Natural Diversity of CRISPR Spacers

Sofia Medvedeva

Skolkovo Institute of Science and Technology

Doctoral Program in Life Sciences

Sorbonne Université

Ecole Doctorale Complexité du vivant

Molecular Biology of Gene in Extremophiles Unit, Institut Pasteur

Natural diversity of CRISPR spacers

By **Sofia Medvedeva**

Doctoral thesis

Supervised by Konstantin Severinov and Mart Krupovic

Date of the defense : 03/06/19

Jury composition:

Prof. Guennadi Sezonov	Co-chair
Prof. Mikhail Gelfand	Co-chair
Prof. Dmitri Pervouchine	Jury member
Dr. Tamara Basta-Le Berre	Jury member
Dr. David Bikard	Jury member
Prof. Olga Soutourina	Jury member
Prof. Konstantin Severinov	Supervisor
Dr. Mart Krupovic	Supervisor

Skolkovo Institute of Science and Technology

Doctoral Program in Life Sciences

Sorbonne Université

Ecole Doctorale Complexité du vivant

Unité Biologie Moléculaire du Gène chez les Extrêmophiles, Institut Pasteur

Natural diversity of CRISPR spacers

Par **Sofia Medvedeva**

Thèse de doctorat de Microbiologie

Dirigée par Konstantin Severinov et Mart Krupovic

Présentée et soutenue publiquement le 03/06/2019 devant un jury composé de :

Prof. Guennadi Sezonov	Président du jury
Prof. Mikhail Gelfand	Président du jury
Prof. Dmitri Pervouchine	Rapporteur de thèse
Dr. Tamara Basta-Le Berre	Rapporteur de thèse
Dr. David Bikard	Examinateur
Prof. Olga Soutourina	Examinatrice
Prof. Konstantin Severinov	Directeur de thèse
Dr. Mart Krupovic	Directeur de thèse

Natural Diversity of CRISPR Spacers

ABSTRACT

CRISPR-Cas is a prokaryotic immunity system against mobile genetic elements, such as viruses and plasmids. The system consists of two components: the clustered regularly interspaced short palindromic repeats (CRISPR) and CRISPR-associated proteins (Cas). In the CRISPR array, short fragments of foreign DNA, called spacers, are interleaved with palindromic repeats. During the adaptation stage of the CRISPR-Cas immunity, new spacers are inserted into the CRISPR array, whereas during the expression and interference stages, spacers are transcribed, processed and complexed with Cas proteins to target the complementary foreign DNA or RNA molecules for degradation. CRISPR array is a fast-evolving part of the genome, with acquisition, duplication, and loss of spacers occurring concurrently to point mutations in the CRISPR repeat and spacer sequences. Thus, sequences of CRISPR arrays can be used to differentiate closely related bacterial lineages. Moreover, analysis of CRISPR spacers is a valuable source of information about virus-host interactions, particularly powerful when applied to metagenomic data.

In this work, we explored the diversity of CRISPR spacers in different natural prokaryotic communities, including extinct Escherichia coli community from a mammoth intestine, Flavobacterium communities from Antarctic surface snow, Thermus communities from four distant hot springs in Italy and Chile, and Sulfolobales community from a Japanese thermal field. The comparison of obtained environmental spacer sets with each other and with spacers from public databases as well as with sequences of viruses allowed us to reach several non-trivial conclusions and to gain insights into virus-host and virus-virus interactions in natural microbial communities.

PUBLICATIONS

1. Savitskaya E, Lopatina A, **Medvedeva S**, Kapustin M, Shmakov S, Tikhonov A, Artamonova I, Logacheva M, Severinov K. Dynamics of Escherichia coli type I-E CRISPR spacers over 42,000 years. Mol Ecol. 2016 Dec 20; 26(7):2019-2026.
2. Lopatina A, **Medvedeva S**, Shmakov S, Logacheva MD, Krylenkov V and Severinov K. Metagenomic Analysis of Bacterial Communities of Antarctic Surface Snow. Front. Microbiol. 2016 Mar 31;7:398
3. Lopatina A#, **Medvedeva S#**, Artamonova D, Sitnik V, Ispolatov J and Severinov K. Natural Diversity of CRISPR Spacers of Thermus: Evidence of Local Adaptation and Global Spacer Exchange. Philos Trans R Soc Lond B Biol Sci. 2018 Mar 25.
4. **Medvedeva S**, Liu Y, Koonin EV, Severinov K, Prangishvili D, Krupovic M. Virus-borne mini-CRISPR arrays promote interviral conflicts and virus speciation. Submitted.
5. Krupovic M, Makarova KS, Wolf YI, **Medvedeva S**, Prangishvili D, Forterre P, Koonin EV. Integrated Mobile Genetic Elements in Thaumarchaeota. Environmental Microbiology. 2019 Feb 17, doi: 10.1111/1462-2920.14564.
6. Musharova O, Vyhovskyi D, **Medvedeva S**, Guzina J, Zhitnyuk Y, Djordjevic M, Severinov K, Savitskaya E. Avoidance of Trinucleotide Corresponding to Consensus Protospacer Adjacent Motif Controls the Efficiency of Prespacer Selection during Primed Adaptation. mBio Dec 2018, 9 (6) e02169-18

#equal contribution

INTRODUCTION

LIST OF ACRONYMS

ATP – adenosine triphosphate

BLAST – Basic Local Alignment Search Tool

BREX system – Bacteriophage Exclusion system

bp – base pairs

Cas – CRISPR-associated

Cascade – CRISPR-associated complex for anti-viral defense

CRISPR – Clustered Regularly Interspaced Short Palindromic Repeats

crRNA – CRISPR RNA

dsDNA – double-stranded DNA

IHF – integration host factor

HTS – High Throughput Sequencing

MGE – mobile genetic element

mRNA – messenger RNA

NGS – Next Generation Sequencing

nt – nucleotides

PAM – Protospacer Adjacent Motif

PCR – polymerase chain reaction

R-M system – restriction-modification system

RNase – ribonuclease

RT – reverse transcriptase

ssDNA – single-stranded DNA

TA system – toxin-antitoxin system

tracrRNA – trans-activating crRNA

INTRODUCTION

1. Defense systems of prokaryotes

Bacteria and Archaea developed a wide range of immune mechanisms to defend themselves against foreign DNA. The restriction-modification (R-M), CRISPR-Cas, pAgos (prokaryotic Argonaute proteins) (1), and BREX systems are all based on self vs non-self DNA (or RNA) discrimination. By contrast, the abortive infection, and toxin-antitoxin systems induce programmed cell death or cell dormancy upon virus infection. Different defense mechanisms often coexist within one genome. Moreover, they are colocalized in genomic regions called "defense islands" (2).

Restriction-modification systems

The R-M systems consist of two enzymes: the endonuclease and the methyltransferase. Methyltransferase transfers methyl groups from S-Adenosyl-L-methionine to specific DNA motifs in the host genome, whereas endonuclease recognizes the same, but unmethylated motifs and cleaves the foreign DNA. R-M systems are classified into different types, depending on the subunit composition, location and type of cleavage and recognition sites (3). Type II R-M systems, which have been harnessed for molecular biology applications, consist of separate methylase and endonuclease enzymes. They recognize 4-8 bp palindromic sequences and cleave within or near the recognition site. Type I and III systems are ATP-dependent hetero-oligomeric complexes with non-palindromic recognition sites. Unlike in other R-M systems, in type IV systems the endonuclease recognizes and cleaves methylated DNA (4).

BREX system

Discovered as part of the "defense islands" in 10% of prokaryotic genomes, the BREX (bacteriophage exclusion) system consists of 6 genes (5). The key member of the BREX gene cassette, pglX gene, is a DNA methyltransferase, which was shown to methylate DNA inside of a 6 nt motif (6). However, unlike in the R-M systems, no degradation of the virus or host DNA in the absence of methyltransferase was observed. Thus, BREX system prevents the replication of a wide range of phages by a mechanism, different from that of the R-M systems.

Prokaryotic Argonaute proteins

Prokaryotic Argonaute proteins (pAgos) are nucleic acid-guided nucleases, discovered in 32% of archaeal and 9% of bacterial genomes (7). They are homologous to proteins of the extensively

characterized eukaryotic Argonaute proteins, which play a central role in RNA silencing processes, as essential components of the RNA-induced silencing complex (8). Guided by small single-strand nucleic acids (13-25 nt in length), which are generated by a "chopping" mechanism (9), pAgos recognize and cleave the invader nucleic acid. All combinations of guide and target types of nucleic acid were observed: DNA-guided DNA interference in T. thermophilus (10), RNA-guided DNA interference in R. sphaeroides (11), DNA-guided RNA interference in A. aeolicus (12) and RNA-guided RNA interference in M. piezophila (13). Two types of pAgo proteins were characterised: 1) "long" pAgos have the same domain composition as eukaryotic Argonaute proteins (i.e., PIWI, MID, and PAZ domains); 2) "short", much less studied, pAgos lack the oligonucleotide-binding PAZ domain and are associated with diverse nucleases (14).

Abortive infection and toxin-antitoxin systems

Abortive infection is an altruistic defense mechanism, when infected cell induces a programmed cell death process, which stops the virus propagation in the population (15). One of the best studied examples is a Rex system of the Escherichia coli bacteriophage lambda. The RexA protein senses the replication of the phage and activates the ion channel RexB, which depolarizes the membrane, causing cell death (16). A great diversity of 20 abortive infection systems, usually encoded in plasmids, was found in Lactococci (17). However, the exact mechanisms of action remain unknown in most cases.

Toxin-antitoxin (TA) systems are composed of stable toxin proteins which target essential cell processes and unstable antitoxins, which prevent the toxin activity. Currently, TA systems are classified into six types, depending on the toxin-antitoxin interaction type and the nature of the antitoxin (18). In type I TA systems, the antitoxin is a small antisense RNA, which binds to the toxin mRNA and promotes the degradation of RNA duplex or inhibits the translation of toxin protein from mRNA by blocking Shine-Dalgarno sequence (19). In type II TA systems, the antitoxin is a protein, neutralizing the corresponding toxin by a protein-protein interaction (20). The type III antitoxin is a repeat-containing RNA, which sequesters the toxin by an RNA-protein interaction (21). In type IV systems the antitoxin protein prevents the binding of the toxin to its target (22). Type V antitoxins are specific ribonucleases, which cleave the mRNA of the toxin (23). The antiviral defense functions of TA systems were demonstrated for types I, II and III (24-26). The arrested translation of host proteins or altered transcription regulation during virus infections may change the ratio of toxin-antitoxin components, which leads to the activation of the toxin and subsequent cell suicide. Although the infected cell dies, the clonal population prevails.

CRISPR-Cas systems

CRISPR-Cas system is an RNA interference-like prokaryotic immune system directed against mobile genetic elements, such as viruses and plasmids (27). The system consists of one or several CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) arrays and Cas (CRISPR-associated) proteins. All Cas proteins can be functionally assigned to adaptation, expression and interference modules (28) (Figure 1). Cas proteins from the adaptation module incorporate fragments of the viral DNA into the CRISPR array as spacers sandwiched between repeats. Transcription and processing of CRISPR array result in production of protective CRISPR RNAs (crRNAs). Interference module proteins, directed by crRNA, recognize and cleave cognate regions in the DNA or RNA of mobile genetic element. By composition of interference and adaptation modules CRISPR-Cas systems are classified into 2 classes, 6 types and ~30 subtypes (29). CRISPR-Cas systems are the focus of this PhD thesis and will be described in more detail in the following sections.

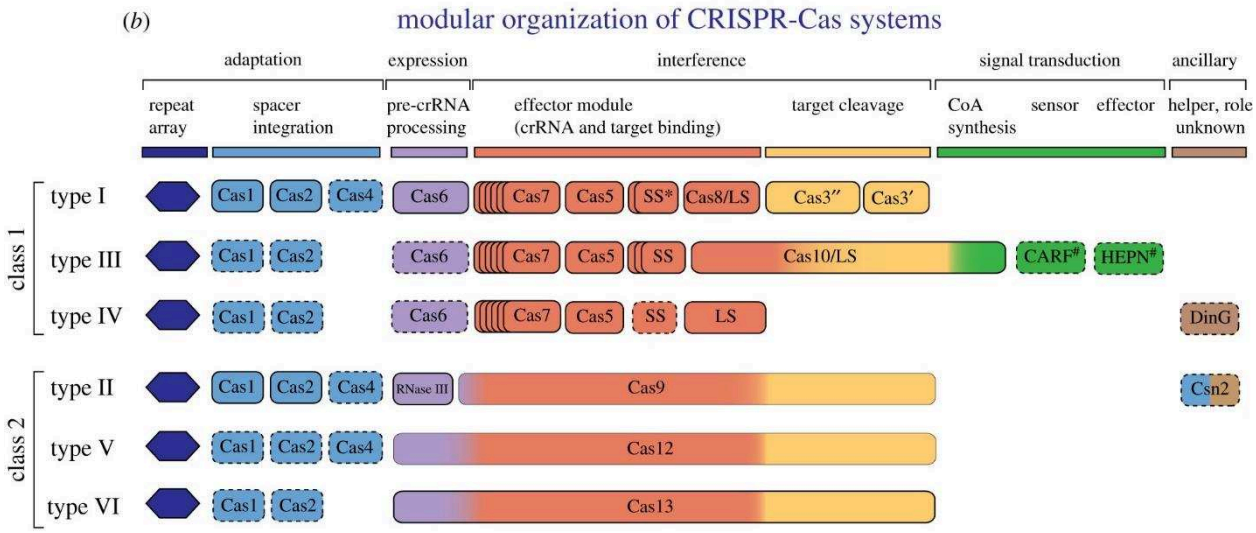


Figure 1. Modular organization of CRISPR-Cas systems. All CRISPR-Cas types have similar modular organization. Cas proteins can be assigned to adaptation (blue), expression (purple) and interference (red+yellow) modules. Functionally dispensable Cas proteins are shown with dashed outlines. Reproduced with permission from (30).

2. CRISPR arrays

CRISPR array is a genomic region containing palindromic repeat sequences interspaced by nonrepetitive sequences, called spacers. Sequences of CRISPR repeats are species- or sometimes even strain-specific and may slightly vary along the length of the CRISPR array. Many, but not all, CRISPR repeat sequences are palindromic and are able to form hairpin structure (31, 32).

One genome can contain several CRISPR arrays with similar or different CRISPR repeats. CRISPR arrays vary in size from ~100 bp with 1 spacer to more than 40000 bp with 587 spacers (type I-U CRISPR array of Haliangium ochraceum) (33). CRISPR arrays and cas gene operons are thought to be subject to horizontal gene transfer, possibly via mobile genetic elements carrying CRISPR loci (34, 35).

A region upstream of the first CRISPR repeat is called the 'leader'. The leader sequence comprises a promoter for transcription of the CRISPR array and sequence elements required for the adaptation process. New spacers are primarily incorporated in the leader-proximal end of array, between the leader and the first CRISPR repeat (36). The replacement of promoter sequence in the leader does not affect spacer acquisition, so transcription of the CRISPR array is not essential for adaptation. Upon deletion or replacement of first 20 or 40 bp of the leader sequence, new spacers were not incorporated in the I-E CRISPR array of E. coli (37), illuminating the critical role of the repeat-proximal region in the adaptation process. Leader sequences are conserved in genomes of the same species, genus or even order. Short conserved nucleotide sequences, probably involved in regulation of adaptation and transcription, were found in similar leader sequences (38).

After transcription, long pre-crRNA is processed by endoribonucleases into small crRNAs. In class 1 CRISPR-Cas systems, Cas6 protein produces individual crRNAs with the spacer sequence and CRISPR repeat-derived 3' and 5' handles by cleaving pre-crRNA inside CRISPR repeat sequences (Figure 2). The processing of pre-crRNA by Cas6 depends on the structure of CRISPR repeat. If CRISPR repeat is palindromic and the canonical stem-loop structure can be formed, Cas6 acts as a single-turnover enzyme: it binds to the stem-loop, cleaves RNA at the base of the stem-loop and later becomes a part of the Cascade effector complex (in type I systems). In the case of a nonpalindromic structure of CRISPR repeat, Cas6 forces the formation of an RNA stem-loop, cleaves pre-crRNA and releases the crRNA (39, 40). In class 2 systems, the processing of pre-crRNA involves binding of tracrRNA to complementary CRISPR repeat sequences in pre-crRNA and their cleavage by RNase III in the presence of Cas9 effector (41) (Figure 2).

New spacers are predominantly incorporated after the leader sequence (42). As a result, the leader-distal end of the CRISPR array contains the oldest spacers and is more conserved than the leader-proximal end (43). Spacers from the leader-distal end of CRISPR array are transcribed less efficiently than the leader-proximal spacers (44). To maintain the optimal number of spacers

in the CRISPR array, old spacers from the trailer end are eliminated, possibly via homologous recombination between CRISPR repeat sequences (45, 46).

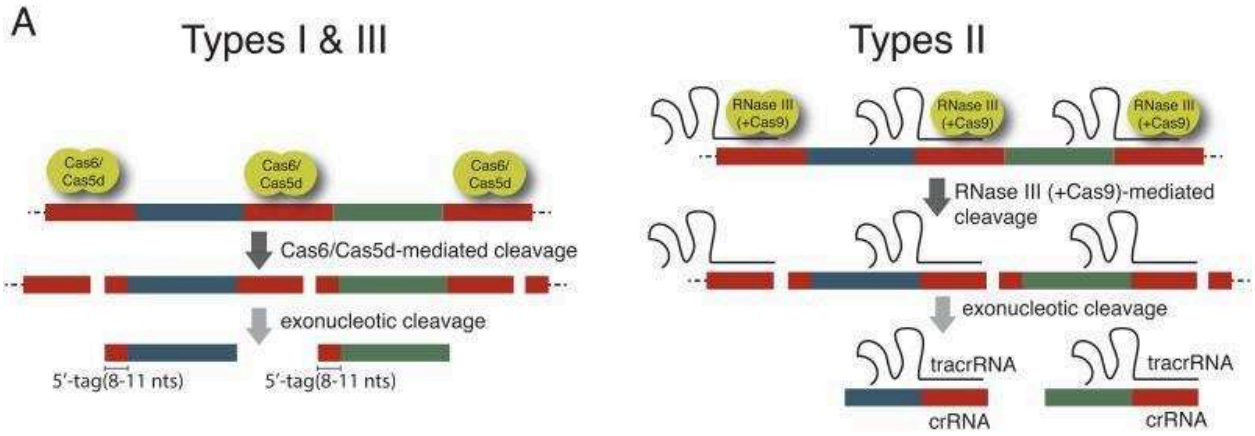


Figure 2. P**rocessing of pre-crRNA in type I, II and III CRISPR-Cas systems.** In type I CRISPR-Cas systems (left panel), Cas6 cleaves pre-crRNA inside the repeat sequence producing crRNA with 8-11 nt repeat-derived 5' tag. In type II CRISPR-Cas systems (right panel), pre-crRNA is processed by RNase III in the presence of Cas9 and tracrRNA. Reproduced with permission from (39).

3. Adaptation module

Adaptation module is conserved in all types of CRISPR-Cas systems and is considered to be a hallmark of the CRISPR-Cas systems. It includes Cas1, Cas2 and, optionally, Cas4 proteins, all of which are nucleases. The mechanism of new spacer integration is similar to site-specific integration of cut-and-paste transposons (Figure 3):

- leader-repeat boundary of CRISPR array is recognized by the Cas1-Cas2 heterohexameric complex carrying prespacer DNA (47);
- two nucleophilic attacks by two 3' OH terminal groups of prespacer are catalyzed by Cas1 nuclease; one 3' end of prespacer is connected to the first nucleotide of repeat sequence on one strand, whereas the other 3' end of the prespacer is connected to the last nucleotide of the repeat on the opposite strand (48);
- the ssDNA gaps formed by the first CRISPR repeat are fill-in repaired and ligated by cell factors.

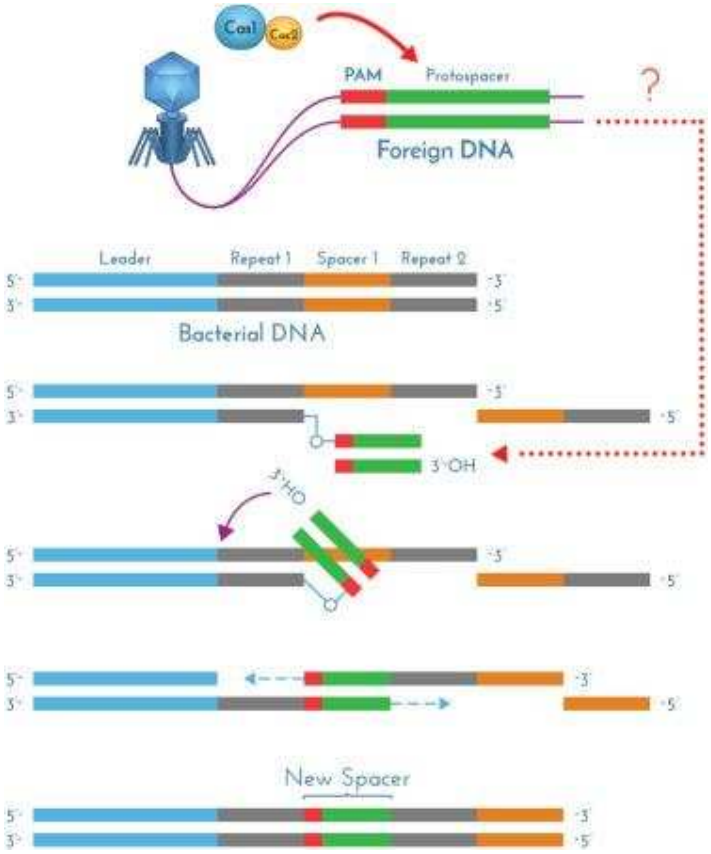


Figure 3. Spacer integration mechanism by Cas1-Cas2. Cas1 catalyzes nucleophilic attacks by two 3' OH terminal groups of prespacer, connecting prespacer with the first nucleotide of repeat sequence on one strand and the last nucleotide of the repeat on the opposite strand. Single-stranded gaps are fill-in repaired by unknown cell factors, resulting in spacer integration and duplication of the Leader-proximal repeat. Reproduced with permission from (42).

The integration of spacers by type I-E system in E. coli requires supercoiled target DNA or DNA bound by IHF (integration host factor), which introduces a stationary bend (49). Off-target spacer acquisition into sequences, which resemble CRISPR repeats and are preceded with IHF binding site, was demonstrated in vivo (50). Recognition of the leader sequence defines specificity of Cas1-Cas2 spacer integration after the first repeat in vitro (51). Although Cas1 protein alone is capable of spacer integration in vitro (48), the addition of Cas2 greatly enhances the efficiency of integration. Enzymatic activity of the Cas2 nuclease is not important for the spacer integration (48). Transcription of cas genes is silenced by H-NS factor in E. coli, and reactivated with LeuO protein (52).

In types I and II CRISPR-Cas systems, for the most of adapted spacers, a protospacer is preceded by protospacer adjacent motif (PAM). PAM is a specific sequence located near 3' end (for type II systems) or 5' end (for type I systems) of the protospacers (53, 54). PAM is recognized by

Cas1 (53) and defines the polarity of spacer integration. During the integration process in type I-E systems, the last nucleotide of PAM in the prespacer becomes the last nucleotide of the CRISPR repeat (55).

In type II CRISPR-Cas systems, adaptation process requires Cas9 with tracrRNA, which forms a complex with Cas1, Cas2 and Csn2 proteins. In this complex, Cas9 is responsible for the selection of protospacers with correct PAM sequence, while nuclease activity of Cas9 can be removed with no influence on the spacer acquisition (56).

In type III CRISPR-Cas systems, Cas1 protein can be fused to a reverse transcriptase (RT) domain. A complex of Cas1-RT and Cas2 was shown to incorporate new spacers originating from either DNA or RNA, with RNA-derived spacers largely matching the highly transcribed genes. The orientation of spacers in the CRISPR array was random with no preference to sense or antisense strand (57).

In type I systems, two variants of the adaptation process – naïve and primed – were demonstrated. Naïve adaptation is mediated by Cas1 and Cas2 proteins. Only ~35% of spacers acquired by naïve adaptation mechanism contain correct AAG PAM sequences, required for the interference in E. coli (37). Spacers selected by Cas1-Cas2 were clustered near double-strand break hotspots, such as replication fork stalling sites (Ter sites), suggesting the involvement of DNA repair machinery in the process of naïve adaptation (58). By contrast, primed adaptation besides the Cas1-Cas integrase requires Cas3 nuclease and the Cascade complex (55). This mechanism is activated under conditions of attenuated interference, when crRNA matches protospacer with one or several mismatches or PAM sequence is not optimal. In this case, the acquisition of spacers located in cis with the targeted protospacer (referred to as a "priming protospacer") is strongly accelerated. The positions of new spacers acquired from the target are characterized by a strand bias and a gradient in the acquisition efficiency with respect to the priming protospacer (55, 59). In contrast to naïve adaptation, most of acquired spacers contain a correct PAM sequence.

Cas4 protein is an exonuclease with a RecB-like domain (60), present in some type I adaptation modules or outside of the context of CRISPR-Cas systems, as standalone protein (61). Cas4 forms a complex with Cas1 and processes the 3' overhangs of prespacer by exonuclease activity (62); increases the number of spacers with valid PAM and reduces the length of acquired spacers (63); and defines the orientation of prespacer integration (64). A similar function has been

attributed to DnaQ exonuclease-like domain fused to Cas2 in type I-E CRISPR-Cas system (PMID: 29891635).

Evolutionary origins of adaptation module components

The adaptation module and CRISPR arrays may have evolved from casposons (65, 66), a recently discovered superfamily of transposon-like elements. Casposons are flanked by terminal inverted repeats and encode a family B DNA polymerase and a Cas1-like protein, termed casposase, which acts as an integrase. The mechanism of casposon integration is highly similar to the integration of spacers in the CRISPR array – the insertion site contains a leader sequence and the target site is duplicated during the casposon insertion (67). Sequential insertion of casposons, one after another, separated by repeats and thus resembling CRISPR arrays was observed in several genomes (68). Besides the universally conserved DNA polymerase and casposase, casposons carry a diverse gene complement, including homologs of Cas4 nucleases. Cas2 proteins have an RNA recognition motif fold similar to one found in VapD proteins from toxin-antitoxin systems, which are known to co-localize with CRISPR-Cas systems (69). The leader sequences and CRISPR repeats have likely originated from the preexisting target site of casposon integration. Notably, CRISPR repeats and PAM sequences were shown to coevolve with Cas1 (Figure 4) (54).

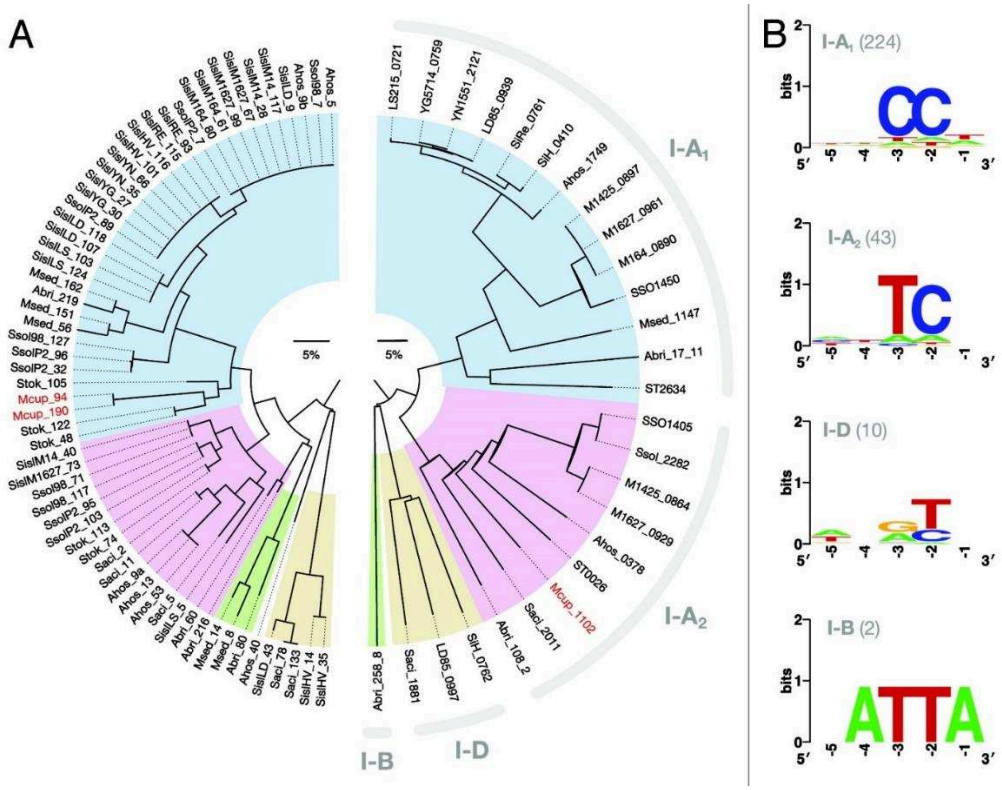


Figure 4. Coevolution of Cas1 proteins (A, left part), CRISPR repeat sequences (A, right part) and PAMs (B) in Sulfolobales genomes. Cas1 proteins and CRISPR repeats of Sulfolobales can be classified into four groups (I-A_1, I-A_2, I-D and I-B). Groups are associated with different PAM sequences (shown as sequence logos). Reproduced with permission from (54).

4. Interference modules

CRISPR-Cas systems are divided into two classes based on the composition of the effector complexes involved in the interference: in class 1 systems (type I, type III and type IV), the interference is conferred by multisubunit complexes, whereas in class 2 systems (type II, type V and type VI), the effector complex consists of a single multidomain protein (29).

Type I interference

Type I interference modules include the Cascade complex (CRISPR-associated complex for antiviral defense) and the Cas3 nuclease. In I-E system of E. coli, Cascade complex consists of the 61 nt crRNA bound to Cas5 and 6 subunits of Cas7, the Cas6 processing nuclease (which holds the 3' end hairpin of the crRNA), large subunit named Cas8, and two small Cse2 subunits (70-72). The mechanism of target recognition and interference for the I-E CRISPR-Cas system is well understood:

1. Large subunit Cas8 recognizes the PAM sequence upstream of the protospacer in the target DNA (73).
2. Binding of Cascade to foreign DNA induces conformational changes in the complex: small Cse2 subunits slide to the 5' end of crRNA and push the C-terminal domain of the Cas8 between the two DNA strands, melting the dsDNA duplex (74).
3. The crRNA hybridizes with the target DNA strand, displacing the nontarget strand and forming an R-loop. Nontarget strand is stabilized by Cse2 and Cas7 subunits (75). Cascade complex can adopt two conformational states depending on the presence of interfering PAM and complementarity between crRNA and protospacer sequences, which determine the size of the R-loop (76).
4. If PAM is recognized and full-size R-loop is formed, the binding site for Cas3 becomes exposed on the Cas8 surface (77). Cas3 nicks the displaced nontarget strand within the protospacer region (78, 79). Through its ATP-dependent helicase activity, Cas3 moves along the nontarget strand in 5' -> 3' direction, unwinding the DNA and generating DNA loops (77, 80).
5. If PAM is not recognized, but protospacer sequence is complementary to crRNA, Cas1-Cas2 complex is recruited before the Cas3 (71).

6. After initial cleavage inside the protospacer, Cas3 nuclease unspecifically degrades the DNA of targeted mobile genetic element (81).

Type II interference

Type II effector complex consists of the endonuclease Cas9 with HNH and RuvC nuclease domains, a crRNA and a trans-activating CRISPR RNA (tracrRNA – a small RNA molecule, partially complementary to the CRISPR repeat). Interference by Cas9 requires a PAM sequence and complementarity between the crRNA and the target protospacer. The type II interference mechanism is relatively simple:

1. Loading of crRNA and tracrRNA onto Cas9 induces conformational changes in the complex, converting Cas9 into an active state (82).
2. Effector complex scans for PAM sequences in target DNA (83). If PAM is found, the formation of R-loop is initiated by bending of the DNA duplex (84).
3. RuvC domain cuts the displaced nontarget strand, while HNH domain cuts the target strand in the RNA-DNA duplex (85). Both cuts are located inside the protospacer, 3 nt upstream of the PAM sequence, which makes Cas9 a blunt-end generating nuclease (86).

Type III interference

Type III interference modules contain the signature Cas10 protein and different sets of accessory proteins: Csm proteins in types IIIA/D and Cmr proteins in types IIIB/C. Effector complexes Csm and Cmr have similar structures: Csm4 or Cmr3 holds 5' tag of the crRNA and is connected to Cas10, while two backbone proteins (Csm3 and Csm2 or Cmr4 and Cmr5) form a filament around the crRNA and are capped by Csm5 or Cmr1. Type III effector complexes recognize a protospacer sequence in RNA, which matches the crRNA, and degrade target RNA and DNA. Type III effector complexes harbor several nuclease activities:

1. The Cas10 HD domain is an ssDNA nuclease. Cas10 acts as a nonspecific nuclease, cleaving the nontemplate strand of target DNA, from which the target RNA recognized by crRNA is transcribed. Cas10 is only activated when the target is transcribed and is repressed, when the 3' region of a protospacer is similar to the CRISPR repeat sequence (87, 88). Cas10 is temporally regulated – binding of the crRNA to a protospacer induces a conformational change in the HD domain and activates the single-stranded DNase activity. When target RNA is destroyed by other nucleases, HD domain reverses back to its inactive state (89, 90).
2. Csm3 and Cmr4 backbone proteins are ssRNA nucleases. Csm3 is present in multiple copies in the effector complex, covering crRNA. The number of copies depends on the length of

crRNA. Csm3 cleaves RNA complementary to the crRNA in multiple sites separated by 6 nt increments (91, 92).

3. Csm6 HEPN domain is a ribonuclease. Csm6 is not part of the effector complex and requires secondary messenger molecule for the activation. Following the target recognition by the effector complex and specific DNA cleavage by the Cas10 HD domain, the Palm domain of Cas10 synthesizes cyclic oligoadenylates from ATP (93, 94). Oligoadenylates bind to the CARF (CRISPR Associated Rossman Fold) domain of unspecific ribonuclease Csm6 and allosterically activate RNA degradation activity of Csm6 HEPN domain. CARF domains were also found in Cas proteins containing HTH (helix-turn-helix) and other nuclease domains, including PIN, RelE and PD-(D/E)xK. Thus, oligoadenylates generated by Cas10 may activate other nucleases and transcription factors (87).

To sum up, the interference in type III systems includes the following steps: specific DNA degradation by Cas10 and RNA degradation by Csm3 upon crRNA recognition and unspecific RNA degradation by Csm6 (93, 94) (Figure 5).

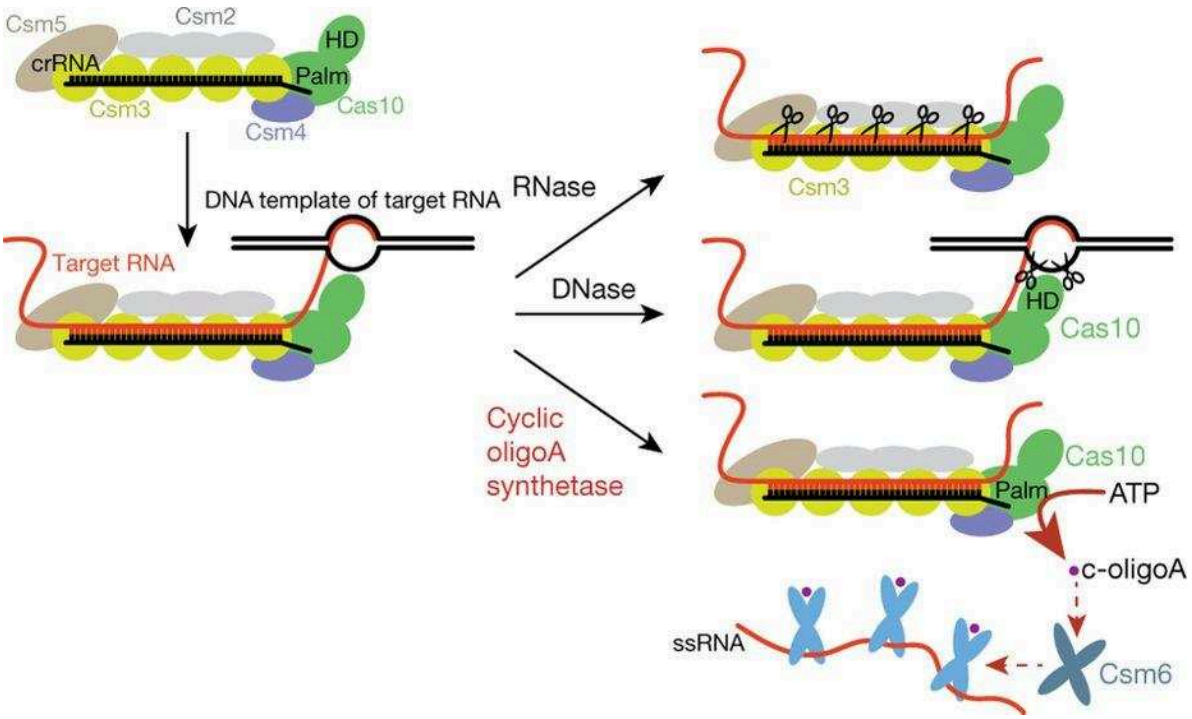


Figure 5. Enzymatic activities of the type III effector complex. Domain composition of type III effector complexes is shown in the top left. Upon target recognition, RNase activity of Csm3 proteins and DNase activity of

HD domain of Cas10 are initiated. In addition to in-built nuclease activities, an independent ssRNase – Csm6 – is activated by c-oligoA synthesized by the Palm domain of Cas10. Reproduced with permission from (94).

Type IV interference

A proposed type IV effector complex consists of the large subunit Csf1 and homologs of Cas5 and Cas7 backbone proteins. In many cases, type IV interference modules are not associated with adaptation modules and probably use crRNAs of other CRISPR types (95). Recently, the structure of type IV effector complex was resolved, confirming protein functions, predicted by homology. Csf5 protein was shown to process the pre-crRNA generating unusual 7 nt 5' end repeat tag of crRNA (96). Type IV-B interference modules are generally encoded on mobile genetic elements (plasmids or viruses) and are associated with several accessory proteins (97).

Type V interference

Type V interference complex includes the Cas12 protein with RuvC and Nuc nuclease domains and a crRNA. The system is characterized by a T-rich PAM sequence, present at the 5' end of protospacer required for the interference (86, 98, 99). RuvC domain performs cleavage of both target and nontarget DNA strands generating staggered ends with 7 nt overhangs (100). Nuc endoribonuclease domain is involved in the processing of pre-crRNAs (101). Different subtypes of type V CRISPR-Cas systems demonstrated great diversity in structure of effector complexes, dependence on tracrRNAs, PAM requirements and nuclease activities (102).

Type VI interference

Type VI effector complex is an RNA-guided RNA nuclease, which consists of Cas13 protein with two HEPN ribonuclease domains and crRNA. Type VI complex is capable of sequence-specific degradation of ssRNA. Similarly to Cas12, Cas13 is responsible for crRNA maturation and does not require a tracrRNA (103). After recognition and cleavage of the main target, Cas13 becomes a nonspecific RNase (104). Cas13 activity is regulated by small accessory proteins.

5. Distribution of CRISPR-Cas systems

CRISPR-Cas systems are found in 90% of Archaea, but only in 50% of Bacteria (105). Furthermore, the distribution of different CRISPR-Cas types across the two domains is not even: type II and type IV CRISPR-Cas systems are present exclusively in Bacteria, while type V is specific to Archaea (95, 106) (Figure 6). In Bacteria, the distribution of CRISPR-Cas systems is mosaic – closely related strains can differ in the presence/absence of CRISPR-Cas systems (107, 108). Furthermore, thermophilic organisms are especially enriched in CRISPR-Cas systems (and

other defense systems) when compared to mesophilic and psychrophilic prokaryotes (106). According to theoretical predictions (109), the CRISPR+ hosts benefit over CRISPR- hosts in conditions of low virus diversity, which is the case in hot geothermal environments. Multiple negative and positive correlations between distribution of CRISPR-Cas systems of particular type in prokaryotic genomes and distribution of components of double-strand break repair systems were reported, suggesting interaction between these two systems (110).

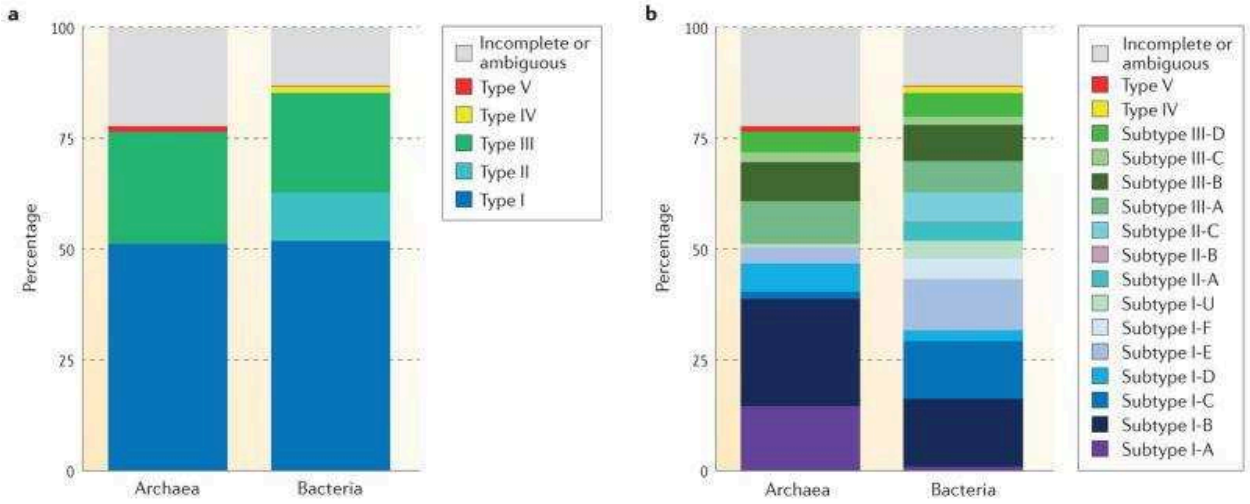


Figure 6. Distribution of CRISPR–Cas system types (a) and subtypes (b) in bacteria and archaea. Incomplete or ambiguous loci are shown with grey color. The dataset was analyzed in 2015, before the identification of type VI CRISPR-Cas systems. Reproduced with permission from (95).

6. CRISPR-Cas immunity of Sulfolobales

Hyperthermophilic archaea from the order Sulfolobales have some of the most complex CRISPR-Cas systems known. For example, the genome of Sulfolobus solfataricus contains six CRISPR arrays, two adaptation modules, five complete and several incomplete interference modules (Figure 7). CRISPR arrays A-E of S. solfataricus are constitutively transcribed from promoters located in long leader sequences (111), whereas CRISPR array F lacks the leader sequence (112). The resulting transcripts are processed by Cas6 into crRNA with 8 nt 5' handle (113). CRISPR arrays A and B are associated with the first adaptation module Cas1AB-Cas2AB, while arrays from C to F are served by the second adaptation module, named Cas1CD-Cas2CD. The CRISPR repeats from A-B and C-D arrays have the same last 8 nucleotides, so the 5' handles of crRNAs generated from these arrays are indistinguishable. As a result, the interference modules of types I-A, III-D, and III-B are not specific to CRISPR array type and can utilize crRNA from all active CRISPR arrays (113).

The Csm (III-D) complex was copurified with crRNAs from all active CRISPR arrays, but mostly with crRNA from A-B CRISPR arrays (114). By contrast, Cmr (III-B) complex showed

preferences to crRNAs from C-D CRISPR arrays (115). This bias can be explained by specificity of Cas6 paralogs. Cas6-1 protein is a multiple turnover enzyme which binds preferably to C-D CRISPR repeats in pre-crRNAs, while Cas6-3, a single-turnover enzyme, associates with the Csm complex and has no preferences to CRISPR repeats (116).

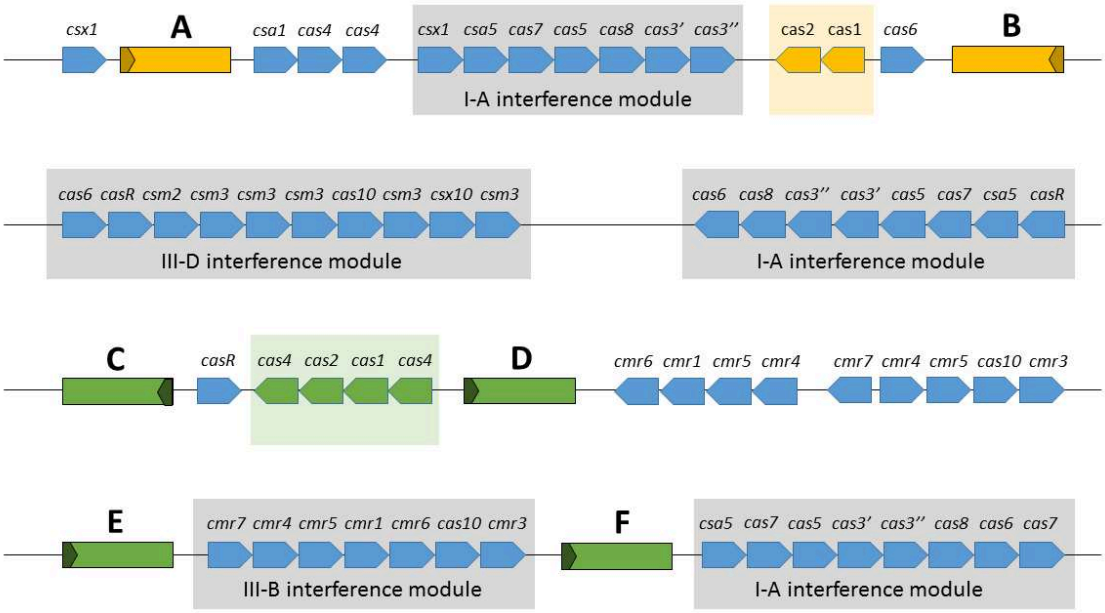


Figure 7. Schematic representation of S. solfataricus CRISPR-Cas modules. Six CRISPR arrays (A-F) are shown as boxes, with an arrow indicating the leader sequence. Two adaptation modules (green and yellow) are associated with the CRISPR arrays of the same colors. Complete interference modules are shown against the grey background.

Adaptation

Two adaptation modules Cas1AB-Cas2AB and Cas1CD-Cas2CD are capable of integration of new spacers in vitro. Acquisition of new spacers into arrays A-B and C-D-E requires "TCN" and "CCN" PAMs, respectively. While in vitro, only Cas1 and Cas2 were required for integration of spacers, in vivo all components of adaptation module were essential: Cas1, Cas2, Cas4, and Csa1 (117). Multiple spacers were incorporated into C, D, and E CRISPR arrays in vivo during infection of S. solfataricus with a mixture of viruses (118). Similarly to I-E CRISPR-Cas system of E. coli, the intact leader sequence and the beginning of the first CRISPR repeat are necessary for spacer integration. Unspecific integration into random positions of plasmid carrying a CRISPR array was observed in in vitro experiments. The addition of known archaeal chromatin proteins did not change this. However, the specificity of spacer integration was restored by unknown host factor(s) from the cell lysate. Cas4 protein slightly increased the specificity of spacer integration and was shown to be involved in trimming of prespacer 3' ends (119).

Interference

Several types of interference modules of Sulfolobus provide immunity against DNA in the presence of PAM, transcribed DNA, or RNA of mobile genetic elements.

Type I-A (Cascade)

Type I-A effector complex consists of Cas7, Cas5, Csa5, Cas8, Cas3 and Cas3'' proteins. In contrast to I-E type, two domains of Cas3 – helicase domain Cas3' and nuclease domain Cas3'' are part of the effector complex. The possible explanation is that recruitment of the trans-acting Cas3 to the pre-formed Cascade complex is inefficient in high temperature environments (120).

When S. solfataricus cells were challenged with MGEs (mobile genetic elements) carrying perfectly matching protospacer or protospacers with mutations, a transcription-independent DNA interference by type I-A interference module occurred. It required an intact PAM sequence and tolerated up to 3 mismatches between the protospacer sequence and crRNA spacer (121). Positions 3-7 and 21-25 of the protospacer were the most important for recognition by the I-A interference complex (122).

Type III-B (Cmr)

The DNA and RNA interference activities of III-B module were demonstrated for S. islandicus. Only plasmids with antisense transcription of protospacer were restricted. PAM sequence was not required for III-B DNA interference, but the presence of 5` sequence similar to the last 6 nucleotides of the CRISPR repeat resulted in loss of targeting (123). The RNA interference occurs when spacer is complementary to the targeted gene transcript, with no PAM sequence required, and multiple mismatches between crRNA spacer and target RNA are tolerated. Two distinct III-B interference modules of S. islandicus have different patterns of RNA cleavage: Cmr-a complex implements a 5' ruler mechanism, cleaving at specific positions, located at 6 nt distance from each other, while Cmr-b complex cuts RNA between UA or UU dinucleotides (124).

Regulation

Transcriptional regulator Csa3a, encoded by a gene located adjacent to the Cas1CD-Cas2CD encoding operon activates the acquisition of new spacers in S. islandicus. Overexpression of the Csa3a protein from a plasmid activates the transcription of the adaptation module genes and leads to the acquisition of hundreds of new spacers with conserved CCN PAM from the host

genome and from the plasmid carrying the csa3a gene. Binding sites for Csa3a were identified in the promoter in front of the adaptation module genes and in the leader sequence of the CRISPR array (125).

Another transcriptional factor, Csa3b, is encoded near type I-A interference modules in S. islandicus and S. solfataricus. Csa3b binds to a palindromic site in the promoter of the interference complex (Cascade-encoding) operon and represses its transcription. The interference complex itself is also shown to interact with a promoter and participate in autorepression of the transcription of its own genes, forming a negative-feedback loop. During virus infection, if protospacer matches the crRNA and a correct PAM is found, Cascade and Csa3b are released from the promoter and transcription of the Cascade genes is reactivated (Figure 8) (126).

Dynamics of cas genes' expression was studied during the infection of S. islandicus with the SIRV2 virus. All cas genes were expressed in uninfected cells, but transcription level for different interference modules varied with I-A and III-B Cmr-B being the most expressed. After 1h of infection the expression of interference modules and CRISPR arrays greatly increased (2-10 fold) and remained at this level during the entire length of the SIRV2 infection cycle (127).

Another regulator of CRISPR-Cas systems in Sulfolobus is Cbp1 protein. Unlike in the case of Csa3a and Csa3B, the gene coding for the Cbp1 is not located near CRISPR-Cas loci. Cbp1 specifically binds to CRISPR repeats (preferring C and D arrays in S. solfataricus), thereby modulating the transcription of CRISPR arrays. Smaller amounts of long pre-crRNAs were found in the cbp1 deletion mutant, whereas overexpression of Cbp1 led to an increased level of pre-crRNAs (128).

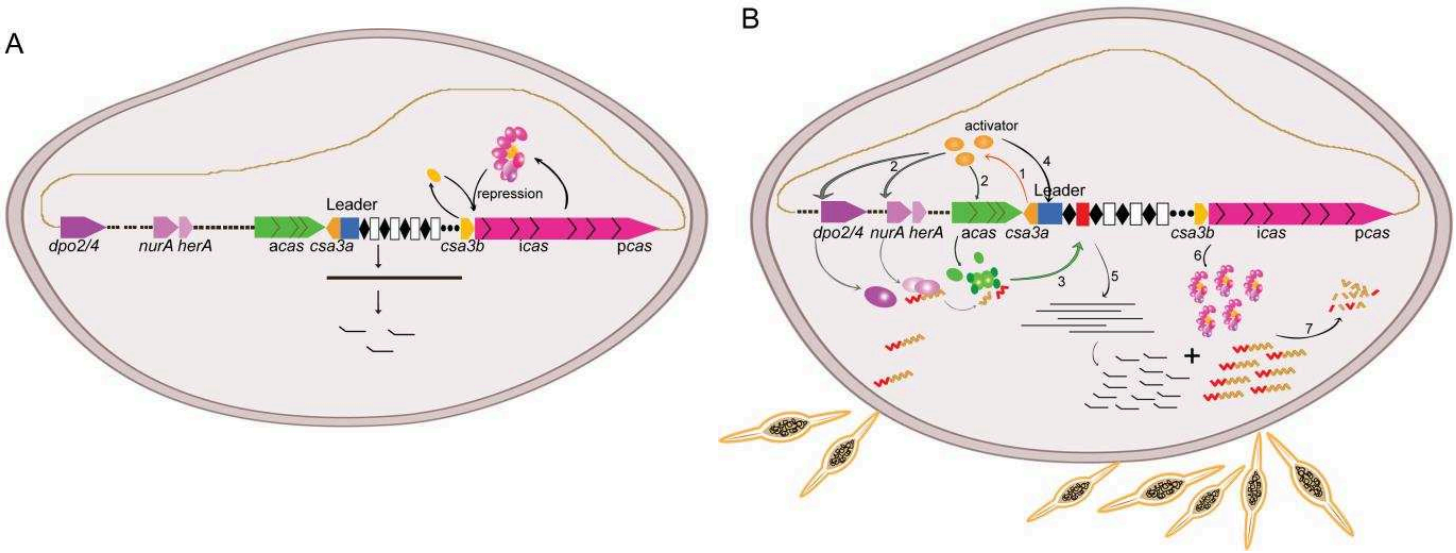


Figure 8. Proposed regulation of type I-A CRISPR-Cas system of Sulfolobus by Csa3b and Csa3a. A. In the absence of virus infection Csa3b and Cascade repress the transcription of interference genes. **B.** Upon virus infection Csa3a activates transcription of adaptation module and pre-crRNA. Cascade complex is released from promoter of interference genes and transcription of interference module is reactivated. Reproduced with permission from (117).

7. Strain subtyping

Even before CRISPR-Cas system was found to be a prokaryotic immune system, analysis of CRISPR arrays was used for subtyping of pathogenic strains (129). CRISPR array is a fast-evolving part of the genome: acquisition, duplication, and loss of spacers occurs together with point mutations in CRISPR repeat and spacer sequences. Thus, sequences of the CRISPR arrays could be used to differentiate closely related bacterial lineages (130). Several CRISPR-based subtyping methods were designed:

1. Spoligotyping: CRISPR spacers are amplified with primers complementary to CRISPR repeat sequences. Labeled PCR products are hybridized with probes containing known spacer sequences (131, 132).
2. Amplification and sequencing of CRISPR arrays: spacer composition of CRISPR arrays and analysis of point mutations in spacer sequences allowed the reconstruction of phylogenetic relationships between Yersinia pestis strains. The similarity of spacer sets correlates with the distance between sites of strain isolation (133). A similar sequence-based method was designed for Salmonella (134).
3. Strain detection with real-type PCR using strain-specific spacer sequences (135).
4. Subtyping based on CRISPR array lengths (136).

CRISPR-subtyping methods demonstrate the best performance in not very active CRISPR-Cas systems, such as I-E system of E.coli. In addition to high rate of spacer acquisition, deletions of spacers and horizontal transfer of CRISPR arrays might hinder the CRISPR-based reconstruction of phylogeny (137, 138).

8. CRISPR in metagenomics

CRISPR spacers represent a catalog of past viral infections and, as such, are a valuable source of information about virus-host interactions. Analysis of spacers can be particularly powerful when applied to metagenomic data. Several bioinformatic tools were implemented for the extraction of CRISPR spacers and reconstruction of CRISPR arrays from metagenomics reads (e.g., CRASS,

CRISPRFinder, PILER-CR, MetaCrast (139-142)). Examples of metagenome-derived spacers matching sequences of phages from the same sampling site were reported, with some spacers targeting low-abundance viruses in the virome (143). In this way, CRISPR spacers can be used to identify viral sequences in metagenomes and monitor changes in viral populations (144, 145). Besides extraction from metagenomic data or CRISPR loci (144, 146), CRISPR spacers can be directly amplified and analyzed either from individual bacterial isolates or whole communities (147, 148). Matching of CRISPR spacers to unannotated metagenomic reads allows identification of plasmid and viral metagenomic sequences (149, 150). Diversity of CRISPR spacers was studied in metatranscriptome data of the human gut (151). Several new variants of CRISPR repeats were identified and long CRISPR arrays were assembled. Most of reconstructed CRISPR arrays were transcribed in one direction, however, several examples of bidirectional transcription were found. Despite relative abundance of type III spacers in matched metagenomic data, only few RNAs were found from type III CRISPR arrays (151).

9. Anti-CRISPR proteins

Mutations in targeted protospacers or associated PAM sequences allow viruses to evade the CRISPR-Cas immunity. In response to such escape mutations, prokaryotes update the collection of spacers via naive or primed adaptation. In addition to mutation-based anti-CRISPR mechanism, some viruses encode small anti-CRISPR proteins (Acrs), which block the action of CRISPR interference complexes (152). More than 20 diverse families of Acrs acting against I-D, I-E, I-F, II-A, II-C, and V-A CRISPR-Cas types have been characterised and Acrs specific for I-A and III-B CRISPR types have been predicted (153, 154).

Several approaches were used for identification of new families of anti-CRISPR proteins:

1. Isolation and analysis of CRISPR-Cas resistant viruses (155).
2. Search for viral genes co-localized with transcriptional regulator aca (anti-CRISPR associated) genes (156).
3. Testing of anti-CRISPR activities of gene products from MGEs integrated in genomes harboring autoimmune spacers (157).

For a few of the discovered Acrs the mechanism of CRISPR-Cas blockage has been studied in detail (Figure 9):

1. I-F anti-CRISPR proteins AcrF1 and AcrF2 interact with different subunits of the Cascade complex and prevent dsDNA binding. AcrF3 freezes Cas3 in an inactive state, preventing the recruitment of Cas3 by the Cascade (156, 158, 159).

2. II-A anti-CRISPR protein AcrIIA4 is a dsDNA mimic, which binds to the PAM recognition site of Cas9 and blocks its activity (160).
3. II-C anti-CRISPR proteins AcrIIC1 and AcrIIC3 have distinct mechanisms of action: AcrIIC1 blocks the DNA cleavage by binding to the Cas9 HNH nuclease domain, whereas AcrIIC3 hinders the binding of Cas9 to dsDNA (160, 161).
4. I-D anti-CRISPR protein AcrID1 contains a lot of negatively charged residues on its surface and probably acts like a dsDNA mimic, similar to ACrIIA4 (162).

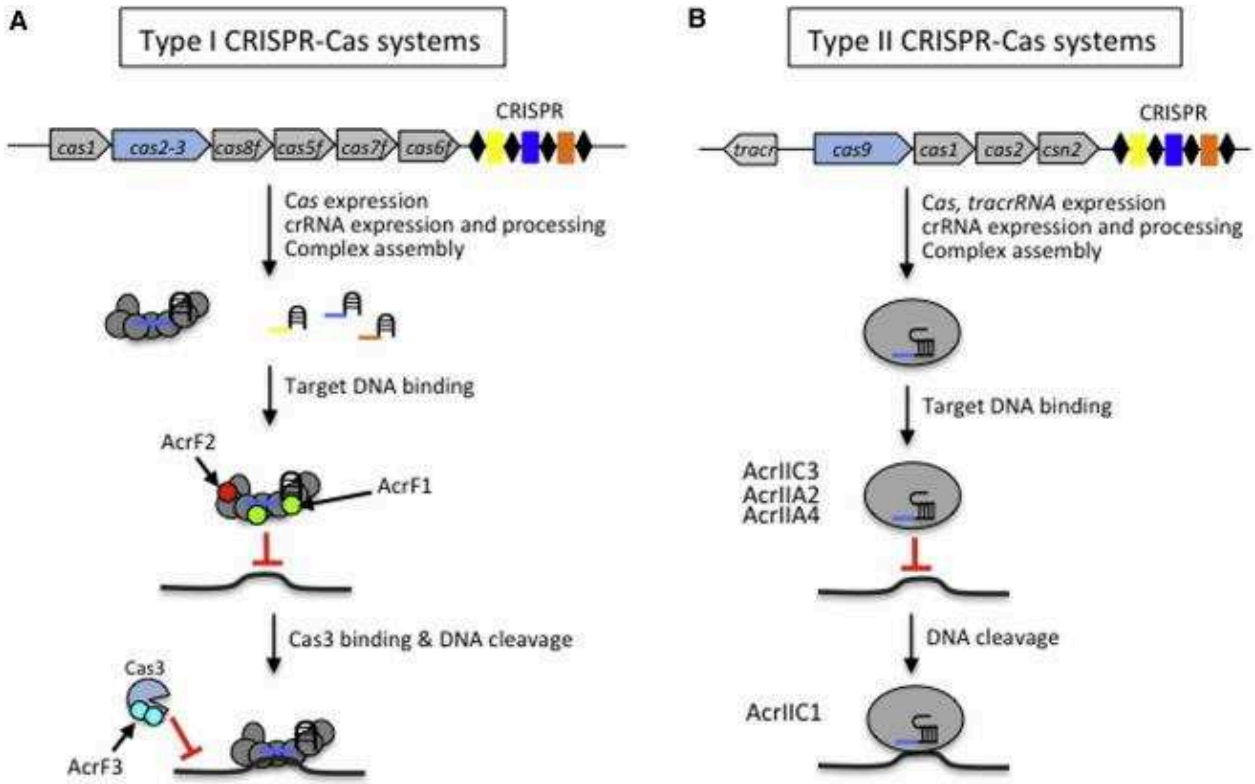


Figure 9. Different strategies of anti-CRISPR proteins for blocking CRISP-Cas activity. In type I systems (**A**) AcrF1 and AcrF2 prevent binding of Cascade complex to the target sequence and AcrF3 blocks the Cas3 nuclease activity. In type II systems (**B**) AcrIIC3, AcrIIA2, AcrIIA4 block Cas9 target recognition and AcrIIC1 inhibits Cas9 nuclease activity. Reproduced with permission from (163).

10. Alternative functions of CRISPR-Cas systems

The role of CRISPR-Cas system in adaptive immunity has been thoroughly studied for different CRISPR-Cas system types in multiple species using variety of conditions. In some cases, for example, I-E CRISPR system of E. coli, CRISPR-Cas machinery does not fulfill its primary purpose as a defense system (164), but rather has alternative functions (165). Below are listed some of the most notable noncanonical CRISPR-Cas functions.

The transcription of cas genes in E. coli is repressed by H-NS factor (52), and no turnover of spacers in CRISPR arrays was observed for a long period of time (108, 166), but potentially harmful immune system is still maintained in the E. coli genome (167). Cas1 protein in E. coli was shown to cooperate with a DNA repair system RecBCD, with the deletion of Cas1 resulting in increased sensitivity to DNA damage stress in mutant cells (168).

Function	CRISPR–Cas type	Mechanism	*Cas* genes involved	CRISPR involved	Species	Experimental evidence
Gene regulation	III-B	Cleavage of complementary mRNA	Yes	Yes	*Pyrococcus furiosus*	No
Gene regulation of group behaviour	I-F	Based on partial complementarity	Yes	Yes	*Pseudomonas aeruginosa*	Yes
	I-C	Unknown	Yes	Unknown	*Myxococcus xanthus*	Yes
Virulence gene regulation	II-C	Cas9-dependent cell surface modification	Yes	No	*Campylobacter jejuni*	Yes
	II-B	Cas9-mediated downregulation of BLP production	Yes	No	*Francisella novicida*	Yes
	II-B	Unknown	Yes	No	*Legionella pneumophila*	Yes
	Orphan CRISPR locus	Regulation of *feoAB* operon by partial complementarity	No	Yes	*Listeria monocytogenes*	Yes
Genome remodelling	I-F	Removal of genomic regions by self-targeting	Yes	Yes	*Pectobacterium atrosepticum*	Yes
DNA repair	I-E	DNA repair by Cas1	Yes	No*	*Escherichia coli*	Yes
Competition between MGEs	I-F	Sequence-specific targeting of competing MGE	Yes	Yes	*Vibrio cholerae* phage ICP1	Yes
Cell dormancy	Not specified	Cas1 and Cas2 function analogously to a TA system to trigger dormancy (and eventual cell death) following phage infection	Yes	No	Not specified	No

Figure 10. Noncanonical CRISPR-Cas functions. Alternative functions of CRISPR-Cas systems of different types, origin of CRISPR-Cas system, Cas proteins or CRISPR array participation in the function are indicated in the table. Reproduced with permission from (169).

Deletion, disruption or mutation of different CRISPR-Cas system components may affect other physiological processes of the cell (Figure 10). The formation of fruiting body in Myxococcus xanthus was significantly reduced by disruption of cas8, cas7 or cas5 genes of I-C CRISPR-Cas system present in the genome (170-172). Moreover, the Cas8 protein was shown to activate the expression of FruA regulator, required for the sporulation process in M. xanthus (171). Another type of group behavior altered in CRISPR-Cas mutants is the biofilm formation in Pseudomonas aeruginosa. Deletion of interference-related cas genes in Pseudomonas strain infected with the temperate phage DMS3 restored the ability to form biofilms (173). The mechanism of biofilm formation regulation by CRISPR-Cas system may involves one of the spacers in Pseudomonas CRISPR array, which is partially complementary to the DMS3 genome (174).

A link between virulence and Cas proteins was demonstrated for several pathogens. In Francisella novicida bacterium, Cas9 protein, tracrRNA and scaRNA (small CRISPR-Cas-associated RNA) downregulate the production of surface lipoprotein BLP, which is involved in recognition of Francisella by the host immune system (175). Similarly, the absence of Cas9 in Campylobacter jejuni influenced binding of host antibodies to the cell surface and altered the swarming behavior (176). Finally, CRISPR adaptation related gene cas2 is necessary for the infectivity of Legionella pneumophila, the mechanism of this regulation, however, remains unknown (177).

AIMS OF THE STUDY

The aim of my PhD thesis project was to answer the following questions:

- How well the CRISPR spacer diversity is represented in current databases? (Chapters I, II, III, IV)
- How variable are the spacer contents in natural populations in short and long terms? (Chapters I, IV)
- Do geographically close/distant prokaryotic populations have similar/different spacer collections? (Chapters II, III, IV)
- Is there biogeographical pattern in virus targeting by CRISPR spacers, i.e., do prokaryotic populations have stronger CRISPR immunity against local viruses? (Chapters II, III, IV, V)
- How do different CRISPR-Cas systems interact with each other in terms of spacer content? (Chapters III, IV)
- Can new facets of virus-host and virus-virus interactions be revealed by studying the spacer diversity in natural microbial populations? (Chapters IV, V)
- What is short-term dynamics of CRISPR spacers during cultivation with viruses? (Annex)
- What are properties of spacer sequences? (Chapter VI, Annex)

RESULTS

CHAPTER I

Dynamics of Escherichia coli type I-E CRISPR spacers over 42 000 years

Introduction:

This chapter describes the development and the first application of CRISPRome (metagenome of CRISPR spacers) analysis by our group. A targeted metagenomics approach was used to assess the diversity of E. coli community from the intestinal content of a mammoth. In addition to natural E. coli community, the model experiment with E. coli strains was performed to evaluate the methodology used. Several pipelines for analysis of CRISPRome data were implemented: extraction of spacers from NGS reads, hierarchical clustering of similar spacer sequences, evaluation of our clustering procedure by comparison to other clustering methods. Developed software was later used for data analysis in Chapter II.

Contribution:

This project had started several years before I joined K. Severinov's lab. By that time, NGS data of E. coli community have been obtained and processed by coauthors and spacer extraction and spacer clustering pipelines were already developed. I applied the developed pipelines to the model experiment with laboratory E. coli strains (Figure 1B). I created the local database of spacers from fully sequenced E. coli genomes and compared the diversity of ancient natural community from a mammoth with diversity of spacers in contemporary E. coli genomes by BLASTN (Figure 2A). I searched for protospacer sequences in sequences of E. coli phages and plasmids with BLASTN (Table 2). Finally, I attempted to reconstruct long CRISPR arrays from the CRISPRome data (Figure 3). I modified the spacer extraction pipeline to obtain pairs and triplets of spacers. De novo reconstruction of CRISPR arrays by overlapping pairs and triplets of spacers led to ambiguous result (possibly due to high number of spacer combinations in pairs). With reference-based CRISPR array reconstruction (using sequences of CRISPR arrays from databases), I was able to find several contemporary CRISPR arrays in the ancient CRISPRome data, suggesting inactivity of I-E CRISPR-Cas system of E. coli. I prepared some figures and tables for the manuscript and contributed to the Methods section. The main text, however, was written by the first author.

MOLECULAR ECOLOGY

Molecular Ecology (2017) **26**, 2019–2026 doi: 10.1111/mec.13961

MICROBIAL LOCAL ADAPTATION

Dynamics of *Escherichia coli* type I-E CRISPR spacers over 42 000 years

EKATERINA SAVITSKAYA,*† ANNA LOPATINA,†‡ SOFIA MEDVEDEVA,*‡ MIKHAIL KAPUSTIN,* SERGEY SHMAKOV,* ALEXEY TIKHONOV,§¶ IRENA I. ARTAMONOVA,**†† MARIA LOGACHEVA‡‡ and KONSTANTIN SEVERINOV*† ‡ §§

*Skolkovo Institute of Science and Technology, Skolkovo, Russia, †Institute of Molecular Genetics, Russian Academy of Sciences, Moscow, Russia, ‡Institute of Gene Biology, Russian Academy of Sciences, Moscow, Russia, §Zoological Institute, Russian Academy of Sciences, St. Petersburg, Russia, ¶Institute of Applied Ecology of the North, North-Eastern Federal University, Yakutsk, Russia, **N.I. Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow, Russia, ††A.A. Kharkevich Institute of Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia, ‡‡M.V. Lomonosov Moscow State University, Moscow, Russia, §§Waksman Institute of Microbiology, Rutgers, the State University of New Jersey, Piscataway, NJ, USA

Abstract

CRISPR-Cas are nucleic acid-based prokaryotic immune systems. CRISPR arrays accumulate spacers from foreign DNA and provide resistance to mobile genetic elements containing identical or similar sequences. Thus, the set of spacers present in a given bacterium can be regarded as a record of encounters of its ancestors with genetic invaders. Such records should be specific for different lineages and change with time, as earlier acquired spacers get obsolete and are lost. Here, we studied type I-E CRISPR spacers of *Escherichia coli* from extinct pachyderm. We find that many spacers recovered from intestines of a 42 000-year-old mammoth match spacers of present-day *E. coli*. Present-day CRISPR arrays can be reconstructed from palaeo sequences, indicating that the order of spacers has also been preserved. The results suggest that *E. coli* CRISPR arrays were not subject to intensive change through adaptive acquisition during this time.

Keywords: *Escherichia coli*, CRISPR spacers, CRISPR arrays, palaeo DNA

Received 29 September 2016; revision received 29 November 2016; accepted 5 December 2016

Introduction

Prokaryotic CRISPR (clustered regularly interspaced short palindromic repeat)-Cas (CRISPR-associated proteins) systems comprise noncoding CRISPR DNA arrays containing variable spacers separated by identical or almost identical repeats and *cas* genes (Makarova *et al.* 2015). Upon CRISPR array transcription and processing, individual CRISPR RNAs containing a single spacer and flanking repeat fragments are bound by Cas proteins. Resulting ribonucleoprotein complexes recognize nucleic acids with sequences matching CRISPR RNA spacer and subsequently degrade them (Barrangou *et al.* 2007; Brouns *et al.* 2008; Marraffini & Sontheimer 2008). New spacers are acquired into one end of CRISPR arrays during a Cas protein-catalysed process referred to as 'CRISPR adaptation' (van der Oost *et al.* 2009). Bioinformatics analysis revealed that some CRISPR spacers are derived from viral and plasmid sequences (Bolotin *et al.* 2005; Mojica *et al.* 2005; Pourcel *et al.* 2005) and it is now commonly accepted that CRISPR-Cas systems control the spread of mobile genetic elements such as plasmids and phages by providing prokaryotes with immunity, which is both adaptive and heritable. Mobile genetic elements can escape the CRISPR-Cas defence by altering sequences recognized by CRISPR RNAs through random mutations or recombination, rendering CRISPR defence inefficient (Andersson & Banfield 2008; Deveau *et al.* 2008; Paez-espino *et al.* 2015) and necessitating acquisition of additional spacers. In several cases, studies of temporal dynamics

Correspondence: Konstantin Severinov, Fax: +1 848 445 5735; E-mail: severik@waksman.rutgers.edu

of bacterial–bacteriophage populations in nature indeed revealed a continuous evolutionary arms race between phages and their hosts driven by cycles of new spacer acquisition followed by accumulation of phage mutants (Andersson & Banfield 2008; Sun *et al.* 2016). Similar dynamics was observed during long-term laboratory cultivation experiments with *Streptococcus thermophilus* (Paez-Espino *et al.* 2013).

The type I-E CRISPR-Cas system of model bacterium *Escherichia coli* is repressed at laboratory conditions (Pougach *et al.* 2010; Pul *et al.* 2010). However, when induced by means of genetic engineering, it efficiently prevents transformation with plasmids and/or infection by phages harbouring sequences matching spacers (Brouns *et al.* 2008; Pougach *et al.* 2010) and is also capable of highly efficient spacer acquisition (Datsenko *et al.* 2012; Yosef *et al.* 2012). The spacer content of natural isolates of *E. coli* is highly variable with overall diversity being higher at CRISPR arrays ends where new spacers are acquired (Diez-Villasenor *et al.* 2010; Touchon *et al.* 2011; Sheludchenko *et al.* 2015), suggesting that the CRISPR-Cas system is active in natural *E. coli* populations. However, compared to some other bacteria, very few *E. coli* spacers match known bacteriophages and plasmids, a surprising result considering the number of known *E. coli* mobile genetic elements (Diez-Villasenor *et al.* 2010; Touchon *et al.* 2011).

Analysis of palaeo DNA offers an unprecedented ability to analyse sequences from distant past and compare them to modern sequences (Hofreiter *et al.* 2015). CRISPR spacers are particularly attractive for such comparative analysis for their small size favours their preservation despite the fragmentation and deterioration of ancient DNA (Dabney *et al.* 2013), while the adaptive nature of CRISPR immunity implies significant turnover of spacers over time. Here, we studied spacers associated with type I-E *E. coli* CRISPR repeats from an extinct pachyderm, a baby mammoth Lyuba that died about 42 000 years ago (Fisher *et al.* 2009), and compared them with annotated contemporary CRISPR spacers available in public databases. To our surprise, we found no evidence of *E. coli* CRISPR spacer turnover. Multiple cases of palaeo CRISPR arrays preservation over the course of 42 000 years have been revealed, implying overall stability of the locus.

Materials and Methods

Sampling

An intact mammoth calf named Lyuba was found at Yamal Peninsula (western Siberia, Russia) in 2007 (Fisher *et al.* 2009) and brought to St. Petersburg without thawing. The carcass was processed in a sterilized laboratory room at −20 °C. The abdominal wall was opened from the left side. All internal organs were in a good shape. The stomach and intestines appeared full. Several grams of intestinal or stomach content were recovered and stored in sterilized packages at −20 °C until further analysis.

DNA extraction

All manipulations with ancient samples, including PCR amplification, were performed in a separate building in laboratory rooms where no prior molecular biology research was conducted. All samples were sterile as judged by the absence of colony formation after aliquots of intestinal or stomach content suspensions used for DNA purification were plated on LB agar plates. DNA was extracted by the following procedure: approximately 0.5 g of material was combined with 600 μL of preheated lysis buffer (10 mM Tris-HCl, pH 7.8, 50 mM EDTA, 150 mM NaCl, 2.5% N-lauroyl sarcosine, 500 mM β-mercaptoethanol, 400 μg/mL proteinase K and 2.5 mM N-phenacylthiazolium bromide (Poinar 1998)), and samples were incubated at 65 °C for at least 4 h with vigorous agitation and extracted with an equal volume of phenol–chloroform (1:1) mixture, followed by chloroform–octanol (24:1) mixture extraction. DNA from aqueous phase was precipitated with isopropyl alcohol (0.6 volume) and 0.1 volume of 3 M sodium acetate. Precipitated DNA was dissolved in 50–100 μL of milli-Q water. A mock control was performed by following the procedure described above with 0.5 ml of distilled water instead of palaeo material. DNA from *Escherichia coli* K12 cells was extracted in standard molecular biology laboratory with genomic DNA purification kit (Thermo Scientific) according to the manufacturer's instruction. Genomic DNA prepared from *E. coli* K12 was shared by sonication on Vibra-Cell VCX130 machine (Sonics) at 100% power for 5 min yielding DNA fragments with a mean ~200 bp length to reproduce the state of degradation of ancient DNA extracted from the mammoth sample.

PCR and sequencing

The method used for spacer amplification is similar to those previously applied for other CRISPR-Cas systems (Sun *et al.* 2016; Lopatina *et al.* 2016). To minimize biases due to variations in individual repeat sequences, primers used for amplification were designed based on a repeat Logo determined with WebLogo 3.0 (Crooks *et al.* 2004) from repeats in all known type I-E *E. coli* CRISPR arrays. PCR amplification was performed using a forward primer Rep1-3 (CGCTGGCGCGGGGAAC WC) and reverse primers Rep 2-1 (GCGCCAGCGGG

GATAAACCG) and Rep 2-2 (GCGCCAGCGGGG ATAAACCN). The molar ratio of Rep2-1/Rep2-2 was 3/1; the overall concentration of reverse primers was the same as that of the forward primer. 50 μL PCR reactions contained 67 mM Tris-HCl, pH 8.3, 17 mM $(NH_4)_2SO_4$, 0.001% Tween 20, 2.5 mM $MgCl_2$, 10 ng of DNA template, 25 pmol of forward primer or reverse primer mix, and 1.25 units of Encyclo Taq polymerase (Evrogen). For each DNA sample analysed, five to ten individual PCR reactions were set up. After amplification, individual reactions were pooled and processed jointly.

Amplicons corresponding to *E. coli* K12 and 'mammoth' samples were used to obtain libraries with TruSeq DNA sample preparation kit according to the manufacturer's instructions. Paired-end sequencing was performed on Illumina MiSeq platform with MiSeq reagent kit v.2 (Illumina), in 250-bp cycles. For palaeo samples, 462, 332 and 402 thousands of pair reads were obtained for first, second and third biological replica, correspondingly. A total of 160 thousands reads were obtained for the K12 sample.

Bioinformatics analysis

Raw sequencing data were analysed using SHORTREAD and BIOSTRINGS packages (Morgan *et al.* 2009). Illumina-sequencing reads were filtered for quality scores of ≥. Reads that contained 32-bp sequences between two CRISPR repeats were selected, and the intervening 32-bp sequences were considered as spacers.

The spacer clustering procedure is presented in detail in the Supporting Information section. Briefly, each spacer was represented as a $32 \times 4 = 128$ dimensional numerical vector in which information about each nucleotide is stored in four corresponding dimensions. The distance between two spacers or clusters was defined as a sum over 128 dimensions of the absolute values of the difference between their coordinates. Spacers were clustered into a three-level branching structure with each subsequent level having clusters of progressively higher similarity between its members. At the last level of segregation, clusters had radii approximately equal to 3, which reflects the maximum number of substitutions between spacers. The code was written in F# and is available upon request. To verify robustness, clustering was performed repeatedly starting with different randomly chosen initial spacer sequences. The procedure converged to same cluster sets for major ($N > 10$) clusters. Next, consensus sequences of each cluster were compared to each other using standard pairwise BLASTn algorithms with an *e*-value less than 10^{-9}. When trivial matchings of each cluster to itself were excluded, overlapping of 0.15% or less of the clusters was detected, indicating that underclustering was minimal. As an independent verification of the clustering procedure, a data set of 30 000 spacers acquired from pG8-C1T plasmid (Shmakov *et al.* 2014) was clustered alone or together with one of the spacer sets analysed in this work. The average number of plasmid-derived spacer clusters corresponded to known number of plasmid protospacers (the ratio did not exceed 1.2), while clustering of combined set of plasmid-derived and palaeo spacers was found to proceed independently, as should be expected because no palaeo spacers match the pG8-C1T plasmid sequences.

The spacer diversity saturation was calculated according to Good's formula: $C = 1 - (n1/N)$, where n1 is the number of sequences that occurred only once and N is the sample size (Good 1953). Spacer clusters of three biological replicates were merged based on pairwise comparison with up to three mismatches tolerated using SHORTREAD and BIOSTRINGS R packages (Morgan *et al.* 2009). Spacers from annotated CRISPR arrays of *Salmonella* and *E. coli* downloaded from GenBank were extracted and clustered in the same way. Pairwise comparison with up to three mismatches tolerated was also used to find intersections between spacer clusters from the mammoth sample and annotated arrays. Two benchmark groups of 'recent' and 'ancient' spacers were composed, correspondingly, from three leader-proximal and three leader-distant spacers from each known array. For each spacer, the frequency of its belonging to one of these groups was determined. Then, the sums of 'recent' and 'ancient' frequency values were next calculated.

To search for protospacers matching spacer sequences, cluster consensus sequences were aligned to nt (2016) databases using BLASTn algorithm adjusted for short sequences. Hits with an e-value>0.001 or matching CRISPR arrays were filtered out.

Reads containing two or three spacers were extracted and grouped with up to three mismatches tolerated in each spacer. Comparisons with fragments of *E. coli* CRISPR arrays present in public databases were performed using SHORTREAD and BIOSTRINGS packages (Morgan *et al.* 2009) with up to three mismatches per each spacer allowed.

To reconstruct CRISPR allele fragments, pairs of neighbouring spacers were represented as a directed graph, where vertices were spacers and edges connecting vertices represented spacers present in one read. Each edge had its own weight reflecting the frequency of two spacers' co-occurrence. To reconstruct most common arrays, we considered only edges with weights above 30. After decomposition of resulting subgraphs into connected components, the longest path for each component was determined. Vertices in the longest path

corresponded to spacer of a reconstructed array. Described algorithms were implemented using SHORTREAD and BIOSTRINGS packages (Morgan *et al.* 2009). Scripts are available from the authors upon request.

Results and Discussion

To determine the overall diversity of spacers associated with *Escherichia coli* type I-E CRISPR repeat in an intestinal sample, a PCR-based method amplifying short spacer-containing fragments of CRISPR arrays with partially overlapping primers complementary to CRISPR repeat was applied (Sun *et al.* 2016; Lopatina *et al.* 2016) (Fig. 1a). The procedure should allow amplification of the entire complement of spacers associated with chosen CRISPR repeat and is particularly well suited for analysis of palaeo DNA which is usually degraded to 50–400-bp fragments (Dabney *et al.* 2013). It should be noted that type I-E CRISPR repeat sequences of *E. coli* and *Salmonella* are identical (Touchon & Rocha 2010), so our procedure cannot distinguish spacers originating from these bacteria. To evaluate the procedure, we applied it to a laboratory *E. coli* strain K12, which contains two CRISPR arrays, CRISPR1 and CRISPR2 according to the classification of Sun *et al.* 2016; with twelve and six different spacers, correspondingly (Fig. 1b) (Diez-Villasenor *et al.* 2010). The K12 genomic DNA was disrupted by sonication to give a mean fragment size of ~200 bp to mimic palaeo DNA. Amplified PCR fragments (Fig. 1c) were purified and subjected to high-density Illumina sequencing. Spacers (defined as 32-nt-long sequences bracketed by CRISPR repeats) were extracted from individual reads and mapped to K12 CRISPR arrays. Reads corresponding to every K12 spacer were obtained (Fig. 1b). The frequency of reads corresponding to different spacers within each array and the mean number of spacers amplified from CRISPR1 and CRISPR2 arrays were not equal, indicating that our procedure provides a representative qualitative but not quantitative view of type I-E repeat-associated spacers. Many of the longer reads contained more than one spacer-repeat unit. When neighbouring spacers from longer reads were analysed, their order matched the order of neighbouring spacers in K12 CRISPR arrays.

Spacer content in samples from baby mammoth Lyuba (Fisher *et al.* 2009) was next investigated. Amplification products were obtained in reactions containing DNA purified from samples of mammoth intestinal content but not in control reactions containing mock-purified DNA or DNA purified from a sample of mammoth stomach content where no *E. coli* was expected (Fig. 1d).

Three independent mammoth intestinal content DNA purifications/amplifications were performed followed by high-density Illumina sequencing. Tens of thousands

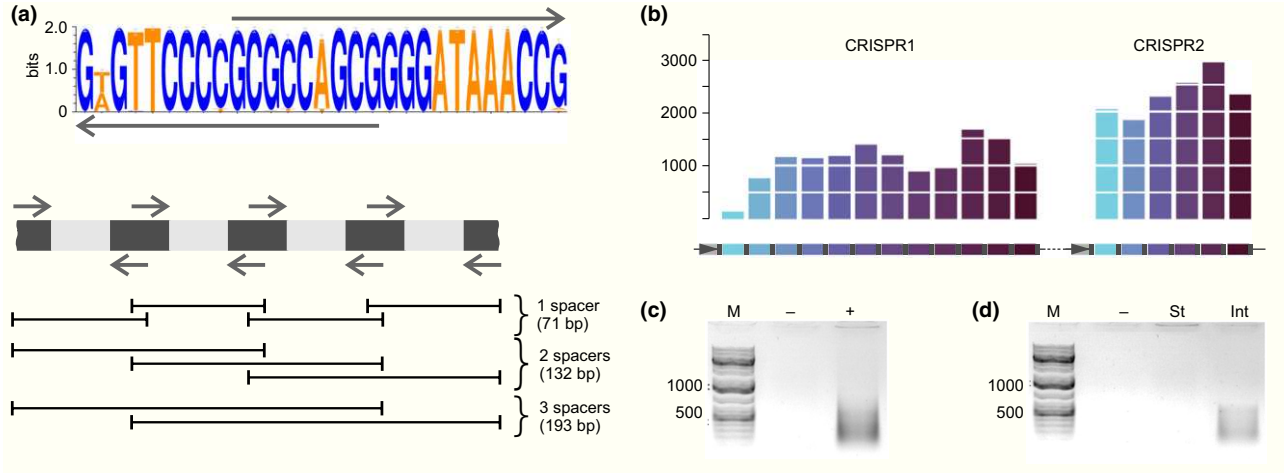


Fig. 1 *Escherichia coli* type I-E CRISPR-Cas system spacer retrieval from K12 strain and a palaeo DNA sample. (a) A Logo of the *E. coli* type I-E CRISPR repeat is shown at the top. The arrows above and below the Logo indicate primers used in PCR amplification. A scheme showing expected products of PCR amplification from an *E. coli* type I-E CRISPR array using repeat-specific primers is presented below. Repeats are dark grey, and spacers are light grey. Expected amplification products are shown below as black lines with their sizes indicated. (b) The procedure outlined in (a) was applied to *E. coli* K12 strain containing two CRISPR arrays (CRISPR1 and CRISPR2, schematically shown at the bottom, with repeats indicated in grey, and spacers are in colour). Rightward horizontal arrows indicate promoters in the leader of each array. Leader-proximal spacers are coloured with lighter shades of blue, while leader-distant spacers are shown in progressively darker colours. The number of Illumina reads corresponding to each spacer is shown on the histograms above. (c, d) Results of *E. coli* type I-E CRISPR spacer amplification from K12 strain (c) and mammoth intestinal ('Int') and stomach ('St') content samples (d). Lanes marked as '-' show results obtained with mock-purified DNA.

of nonredundant spacer sequences were obtained in each replicate (Table 1). Clustering of such a large number of unique sequences based on direct BLAST sequence comparisons of every spacer is a computationally intensive task. Therefore, a faster *k*-means hierarchical clustering-based procedure was utilized (for details of algorithm, threshold values choice and verifications tests, see Materials and Methods and Supporting Information sections). The clustering procedure reduced complexity of spacer sets from each biological replicate to 1.2–1.4 thousands spacer clusters. Sequences that fell into distinct clusters differed from each other in more than three positions. The depth of sequencing allowed us to reach 80–99% coverage of spacer diversity in each replicate as estimated by the Good's criterion (Good 1953) (Materials and Methods and Table 1).

Spacer clusters present in each biological replicate were merged with up to three mismatches tolerated. In this way, a final set of 1883 unique clusters of spacers from the mammoth sample was created (Table 1). To obtain contemporary *E. coli* spacer set for comparison, the clustering procedure was applied to 1728 spacers from *E. coli* type I-E CRISPR arrays present in public databases, producing 1599 spacer clusters. Direct BLAST comparison of the mammoth and contemporary spacer cluster sets revealed 425 common clusters (Fig. 2a).

The set of spacer clusters from public databases for *Salmonella* is much larger than that of *E. coli* (it consists of more than ~3.6 thousands clusters), but the two sets do not overlap. There was a minimal 0.04% overlap between the mammoth and the *Salmonella* sets,

Table 1 Statistics of palaeo-spacer sequencing and clustering

Replicate	CRISPR spacers, total	CRISPR spacers, nonredundant	Clusters	Good's criterion	Cluster combined set
I	824 536	47 429	1411	0.830986	1883
II	448 951	33 226	1220	0.999231	
III	709 795	46 489	1175	0.875101	

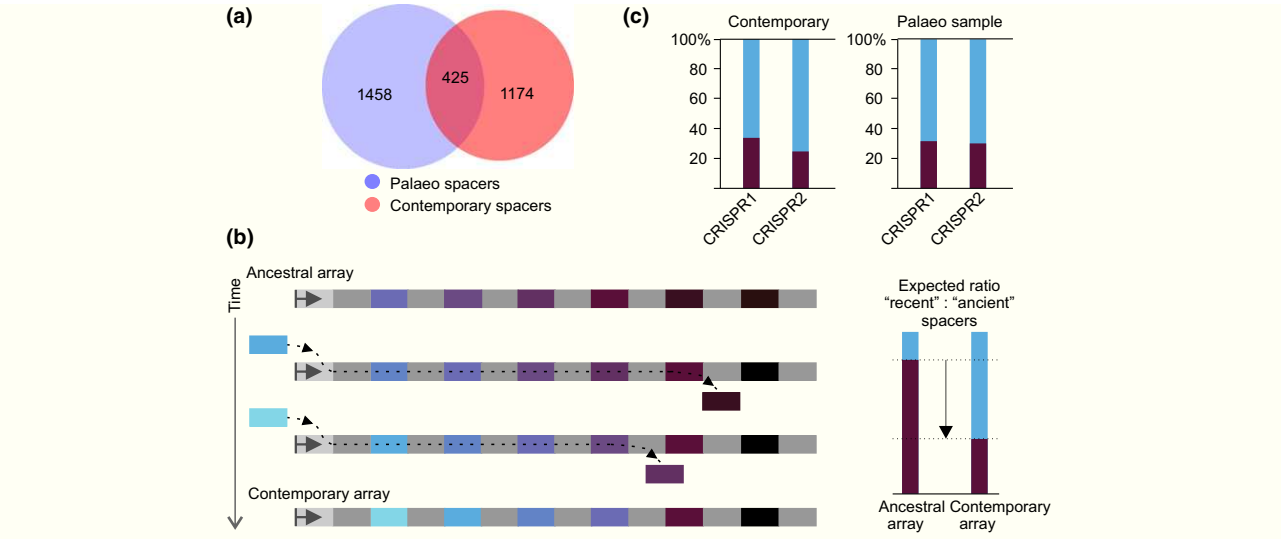


Fig. 2 Comparison of ancient and present-day *Escherichia coli* type I-E CRISPR spacers. (a) Comparison of spacer cluster sets. Numbers within circles correspond to unique and overlapping spacer clusters. Blue circle represents clusters obtained from the mammoth sample; red circle represents known *E. coli* type I-E spacer cluster set. (b) An ancestral CRISPR array is schematically shown at the top. Repeats are light grey, and spacers are coloured. The leader (light grey rectangle with arrow) is shown on the left. With the passage of time, additional spacers (coloured with lighter shades of blue) are acquired at the leader-proximal end, while internal spacers (dark-coloured) are lost. A resulting contemporary array is shown at the bottom. Expected ratios of recently acquired (spacer-proximal) and ancient (spacer-distal) spacers in the ancestral and contemporary arrays are shown at the right. (c) The overall frequency of 'ancient' and 'recent' *E. coli* type I-E CRISPR spacer clusters from known CRISPR arrays present in public databases (DB) and in the mammoth sample is shown. Data for CRISPR1 and CRISPR2 arrays are shown separately.

suggesting that most mammoth sample spacers correspond to *E. coli* type I-E CRISPR arrays spacers.

Spacers are acquired at one end of the array proximal to the leader region, and for every acquired spacer, an additional copy of CRISPR repeat is generated (Barrangou *et al.* 2007; Datsenko *et al.* 2012; Erdmann & Garrett 2012; Lopez-Sanchez *et al.* 2012; Swarts *et al.* 2012). Spacers located close to this end of the array should have been acquired more recently, while distal spacers should correspond to ancient acquisition events. As CRISPR arrays cannot grow indefinitely, the acquisition of new spacers shall be accompanied by the loss of older internal spacers (Deveau *et al.* 2008; Horvath *et al.* 2008; Lopez-Sanchez *et al.* 2012). As a result, a turnover in spacer composition is expected (Fig. 2b). Specifically, recently acquired spacers present in contemporary arrays should have been less frequent or even absent in ancestral arrays (Fig. 2b). For every spacer cluster from contemporary set and for overlapping spacer clusters from the mammoth set, the frequency of spacer occurrence in three leader-proximal ('recent') and leader-distal ('ancient') positions of annotated *E. coli* CRISPR arrays was calculated (see Materials and Methods). The overall frequency of 'recent' and 'ancient' spacer

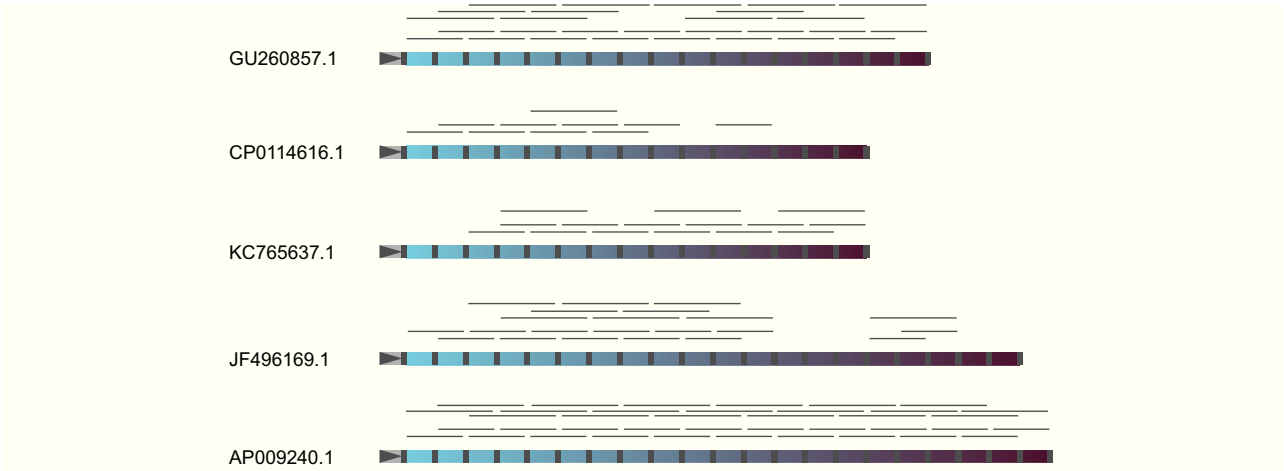


Fig. 3 Reconstruction of contemporary CRISPR arrays from reads containing two or three spacers from the mammoth sample. Mapping results of neighbouring spacer pairs and triplets on five selected CRISPR arrays from contemporary *Escherichia coli* are shown. Repeats are grey, and spacers are coloured. The leader regions are marked by grey triangles on the left of each array. Leader-proximal spacers are coloured with lighter shades of blue, while leader-distant spacers are dark-coloured. Detected reads containing neighbouring spacer pairs or triplets are shown by thin grey lines above each array.

Table 2 Hits of CRISPR spacer clusters originated from the mammoth sample

Cluster consensus sequence	Hit
GCATCTCTTCCACTTAAATCTCCTTGTTACGA	Enterobacteria phage NJ0
CGGGATAATTCAGCTTTCACATCACGGCAAGA	Enterobacteria phage phiEco32
TGCCGGGTTCGACTGGACGCCATTTGCCATCT	Enterobacteria phage epsilon15
GGTAAAAACACGGTCTGAACCGACATTCATGT*	Enterobacteria phage P7
CATTTTTGCGTGGCGAGCTGCGCCGCGTTCTG*	Escherichia phage JLK-2012
ACGATTGGGCAGCCAGAGTTGCCGCCGGGAAA	*Escherichia coli* strain T23 plasmid pEQ1
CGGCCAGGCTGGATTTAAGCGGCACGGCCGCA	Uncultured bacterium plasmid pMBUI4
GTCGCCTCAATAGCGCGTTTACCTTTGCTGTT	Uncultured bacterium plasmid pMBUI4
GCCAGGGCAAGCGGCCCAAGGGCAAGGTCATA	Plasmid pMCBF1
GGGATCTCATCGTCAAAATCGTGAGCCGGATC	*Escherichia coli* strain BK28960 plasmid
CCAGCCGTTCAGTATTGCCGGTGTCAGCAAAA*	*Enterobacter cloacae* strain 34983 plasmid p34983-328.905 kb
GCCGTCGTGCCGTGTTCACCTTTACGAACCTG*	*Klebsiella pneumoniae* ATCC BAA-2146 plasmid pHg
TAAAATGAGAGCTTTTGTTCGCTTGAGCAATA	*Escherichia coli* genome, fimbrial protein
CAAGAAGTACTGAACCGATATACTCGCCAACC	*Escherichia coli* genome, intergenic between two hypothetical proteins
AGGACAGTAAAAATGACGGAATTGTTTATCAG	*Escherichia coli* genome assembly FHI92, tail sheath protein

*Asterisk mark clusters found in both the mammoth and contemporary data sets.

clusters was then determined by summing the values obtained for individual clusters. The spacer content in CRISPR1 and CRISPR2 arrays is unrelated (Diez-Villasenor *et al.* 2010; Touchon & Rocha 2010; Kupczok *et al.* 2015), suggesting that spacers in each array are acquired independently and there is no recombination between arrays. Therefore, 'recent' and 'ancient' spacers from CRISPR1 and CRISPR2 arrays were treated separately. In contemporary *E. coli* spacer set, 'recent' spacers constituted ~70% of the total in both arrays (Fig. 2c). Higher portion of 'recent' spacers arose due to higher diversity of leader-proximal spacers compared to the more homogeneous leader-distant spacers. Strikingly, the overall proportion of spacer clusters matching either 'age' group remained the same in the mammoth set (Fig. 2c). Thus, our analysis failed to reveal a significant turnover of spacers associated with *E. coli* type I-E CRISPR repeats in the course of 42 000 years that separate *E. coli* from mammoth and the present-day *E. coli*.

We next analysed neighbouring spacer pairs in longer high-density Illumina-sequencing reads from the mammoth sample with the hope of reconstructing CRISPR arrays. A total of 902 unique neighbouring spacer pairs were extracted from the mammoth sample and mapped to annotated *E. coli* CRISPR arrays, yielding 257 neighbouring spacer pairs from the mammoth sample that matched annotated CRISPR arrays. Full or almost full-length contemporary arrays could be reconstructed using these spacer pairs. Selected examples of such reconstructions are shown in Fig. 3. The same analysis was performed for triplets of spacers extracted from some of the longer reads. Of a total of 305 cases, 130 triplets corresponded to contemporary arrays, and in several cases, they could be used to reconstruct arrays identical to those reconstructed with spacer pairs (Fig. 3). Thus, some *E. coli* CRISPR arrays or their fragments remained unchanged for more than 40 thousand years.

Most (645) neighbouring spacer pairs from the mammoth sample had no matches to contemporary *E. coli* arrays. They were used to reconstruct longer chains (see Materials and Methods) yielding twelve 3- to 8-spacer-long array fragments that must correspond to CRISPR arrays/array fragments that are either extinct or that have not been isolated yet in contemporary *E. coli*.

The collection of spacers from the 'mammoth' sample considerably expands the variety of unique *E. coli* type I-E CRISPR spacers. Only a small percentage of *E. coli* type I-E CRISPR spacers from the database match sequences of phages and other mobile genetic elements (Diez-Villasenor *et al.* 2010; Touchon & Rocha 2010). In addition to known phage-matching spacers, several novel hits of palaeo spacers to mobile genetic elements were found. However, the overall percentage of hits to genomes of known phages, plasmids and likely prophages for spacer clusters from the mammoth sample remained low (0.6%, Table 2).

Overall, our findings reveal that *E. coli* population contains a vast variety of spacers that remain stable over long periods of time. The order of spacers also appears to be preserved at least in some arrays. Most spacers have no matches to known mobile genetic elements, and their origin and sequences they target remain to be established.

Acknowledgements

We thank Dr. Jaroslav Ispolatov for advice. This work was supported by grants from Russian Science Foundation [14-14-00988 to KS], NIH RO1 grant GM10407 to KS and Russian Foundation for Basic Research [16-04-00767 to E.S and 17-04-02144 to IIA].

References

Andersson AF, Banfield JF (2008) Virus population dynamics and acquired virus resistance in natural microbial communities. *Science*, **320**, 1047–1050.

Barrangou R, Fremaux C, Deveau H *et al.* (2007) CRISPR provides acquired resistance against viruses in prokaryotes. *Science*, **315**, 1709–1712.

Bolotin A, Quinquis B, Sorokin A, Dusko Ehrlich S (2005) Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology*, **151**, 2551–2561.

Brouns SJJ, Jore MM, Lundgren M *et al.* (2008) Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science*, **321**, 960–964.

Crooks GE, Hon G, Chandonia J-M, Brenner SE (2004) WebLogo: a sequence logo generator. *Genome Research*, **14**, 1188–1190.

Dabney J, Meyer M, Paabo S (2013) Ancient DNA damage. *Cold Spring Harbor Perspectives in Biology*, **5**, a012567.

Datsenko KA, Pougach K, Tikhonov A *et al.* (2012) Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. *Nature Communications*, **3**, 945.

Deveau H, Barrangou R, Garneau JE *et al.* (2008) Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. *Journal of Bacteriology*, **190**, 1390–1400.

Diez-Villasenor C, Almendros C, Garcia-Martinez J, Mojica FJM (2010) Diversity of CRISPR loci in *Escherichia coli*. *Microbiology*, **156**, 1351–1361.

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

Fisher D, Rountrey A, Tikhonov A *et al.* (2009) Life history of remarkably preserved woolly mammoth calf from the Yamal peninsula, northwestern Siberia. *Journal of Vertebrate Paleontology*, **29**, 96A.

Good IL (1953) The population frequencies of species and the estimation of population parameters. *Biometrika*, **40**, 237–264.

Hofreiter M, Paijmans JL, Goodchild H *et al.* (2015) The future of ancient DNA: Technical advances and conceptual shifts. *Bioassay*, **37**, 284–293.

Horvath P, Romero DA, Coute-Monvoisin A-C *et al.* (2008) Diversity, activity, and evolution of CRISPR loci in *Streptococcus thermophilus*. *Journal of Bacteriology*, **190**, 1401–1412.

Kupczok A, Landan G, Dagan T (2015) The contribution of genetic recombination to CRISPR array evolution. *Genome Biology and Evolution*, **7**, 1925–1939.

Lopatina A, Medvedeva S, Shmakov S *et al.* (2016) Metagenomic analysis of bacterial communities of antarctic surface snow. *Frontiers in Microbiology*, **7**, 398.

Lopez-Sanchez MJ, Sauvage E, Da Cunha V *et al.* (2012) The highly dynamic CRISPR1 system of *Streptococcus agalactiae* controls the diversity of its mobilome. *Molecular Microbiology*, **85**, 1057–1071.

Makarova KS, Wolf YI, Alkhnbashi OS *et al.* (2015) An updated evolutionary classification of CRISPR–Cas systems. *Nature Reviews Microbiology*, **13**, 722–736.

Marraffini LA, Sontheimer EJ (2008) CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science (New York, N.Y.)*, **322**, 1843–1845.

Mojica FJM, Díez-Villaseñor C, García-Martínez J, Soria E (2005) Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *Journal of Molecular Evolution*, **60**, 174–182.

Morgan M, Anders S, Lawrence M *et al.* (2009) ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics (Oxford, England)*, **25**, 2607–2608.

van der Oost J, Jore MM, Westra ER, Lundgren M, Brouns SJ (2009) CRISPR-based adaptive and heritable immunity in prokaryotes. *Trends in Biochemical Sciences*, **34**, 401–407.

Paez-Espino D, Morovic W, Sun CL *et al.* (2013) Strong bias in the bacterial CRISPR elements that confer immunity to phage. *Nature Communications*, **4**, 1430.

Paez-espino D, Sharon I, Morovic W *et al.* (2015) CRISPR immunity drives rapid phage genome evolution in *Streptococcus thermophilus*. *mBio*, **6**, 1–9.

Poinar HN (1998) Molecular coproscopy: dung and diet of the extinct ground sloth nothrotheriops shastensis. *Science*, **281**, 402–406.

Pougach K, Semenova E, Bogdanova E *et al.* (2010) Transcription, processing and function of CRISPR cassettes in *Escherichia coli*. *Molecular Microbiology*, **77**, 1367–1379.

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

Pul Ü, Wurm R, Arslan Z *et al.* (2010) Identification and characterization of *E. coli* CRISPR- *cas* promoters and their silencing by H-NS. *Molecular Microbiology*, **75**, 1495–1512.

Sheludchenko MS, Huygens F, Stratton H, Hargreaves M (2015) CRISPR diversity in *E. coli* isolates from australian animals, humans and environmental waters. *PLoS One*, **10**, e0124090.

Shmakov S, Savitskaya E, Semenova E *et al.* (2014) Pervasive generation of oppositely oriented spacers during CRISPR adaptation. *Nucleic Acids Research*, **42**, 5907–5916.

Sun CL, Thomas BC, Barrangou R, Banfield JF (2016) Metagenomic reconstructions of bacterial CRISPR loci constrain population histories. *The ISME Journal*, **10**, 858–870.

Swarts DC, Mosterd C, van Passel MWJ, Brouns SJJ (2012) CRISPR interference directs strand specific spacer acquisition. *PLoS One*, **7**, e35888.

Touchon M, Rocha EPC (2010) The small, slow and specialized CRISPR and anti-CRISPR of *Escherichia* and *Salmonella*. *PLoS One*, **5**, e11126.

Touchon M, Charpentier S, Clermont O *et al.* (2011) CRISPR distribution within the *Escherichia coli* species is not suggestive of immunity-associated diversifying selection. *Journal of Bacteriology*, **193**, 2460–2467.

Yosef I, Goren MG, Qimron U (2012) Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. *Nucleic Acids Research*, **40**, 5569–5576.

E.S. designed research, performed experiments, analysed data and wrote the manuscript, A.L. performed experiments, M.K. designed and implemented clustering procedure and analysed data, S.M. and S.S. analysed data, A.T. collected and provided mammoth samples, I.I.A. helped analyse data, M.L. performed high-throughput sequencing, and K.S. designed research, analysed data and wrote the manuscript.

Data accessibility

The cluster sets of palaeo spacers and annotated spacers from *E. coli* and *Salmonella* arrays used in this work as well as CRISPR array fragments reconstituted from palaeo sample can be downloaded from https://doi.org/10.6084/m9.figshare.1613398, and raw data can be download from https://doi.org/10.6084/m9.figshare.4244501.v1.

Supporting information

Additional supporting information may be found in the online version of this article.

Fig. S1 A plot showing the dependence of the number of secondary clusters N vs. the cutoff distance between clusters D.

SUPPLEMENTARY TEXT

k-mean hierarchical clustering: detailed algorithm and parameter adjustment.

Basic statements

Spacers are defined by their 32-nucleotide sequences. A large number ($\mathbf{up\ to}\ 0.5\text{x}10^7$) of spacers needs to be clustered into an initially unknown number of groups, so that spacers in each group are similar to each other and different from spacers from other groups. Also, identical spacers derived from the same protospacer but differing in their orientation (reverse complementary) (Erdmann & Garrett 2012; Lopez-Sanchez *et al.* 2012; Mick *et al.* 2013; Shmakov *et al.* 2014) and spacers produced by imprecise excision (Savitskaya *et al.* 2013), need to be combined and handled together.

A spacer α with a given nucleotide sequence is denoted by the 32 × 4 = 128-dimensional numerical vector $S\alpha$, in which information about each nucleotide is stored in 4 corresponding dimensions in the following way:

- base A is denoted as (1, 0, 0, 0).
- base G is denoted as (0, 1, 0, 0).
- base C is denoted as (0, 0, 1, 0).
- base T is denoted as (0, 0, 0, 1).

For example, sequence [AGGC, . .] corresponds to (1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, . .).

While a vector describing a single spacer $S = (S_1, \ldots, S_{128})$ contains only 0s and 1s, the position $C=(C_1,\ldots,C_{128})$ of the center of a cluster, defined as the arithmetic mean of vectors S_α of constituent n spacers,

$$C_j = \frac{1}{n}\sum_{\alpha=1}^{n} S_{j,\alpha},\ j = 1,\ldots,128, \tag{1}$$

generally is characterized by real numbers $0 \le C_j \le 1$.

The distance $D_{\alpha\beta}$ between two spacers or clusters α and β is defined as a sum over 128 dimensions of the absolute value of the difference between their coordinates,

$$D_{\alpha\beta} = \sum_{j=1}^{128} |C_{j,\alpha} - C_{j,\beta}|. \tag{2}$$

This distance is twice the Hamming distance between spacers, since each replacement of a nucleotide removes 1 from the position of the old base and adds 1 to position corresponding to a new base. The radius of a cluster is defined as the distance from its center to its most remote member.

Sorting into tree-like hierarchy

To reduce the amount of data and accelerate the search, we cluster the spacers into a 3-level branching structure with each subsequent level having clusters of progressively higher similarity between members. At the last level of segregation, clusters have radii approximately equal to 3, which reflect the maximum number substitutions corresponding to biologically similar spacers and sets the "resolution limit" of the process. Parameters defining branching were varied and after several experiments we converged to values listed below. The procedure of placing a new spacer into the system of clusters consists of the following steps:

- The first spacer forms the root, the first-level branch, and the second-level branch of the first tree.
- Each new spacer is first matched with the closest tree root. If no tree is found within a distance of 27, the new spacer forms the root, first-, and second-level branches of a new tree.
- If a matching tree is found, the new spacer is then matched with the closest first-level branch coming out from the root. If no first-level branch is found within a distance of 9 from the spacer, the spacer forms new first-level and second-level branches.

- If the matching first-level branch is found, the spacer is then compared to the second-level branches emanating from the first-level branch. It joins the closest second-level branch, and if no such branch exists within a distance of 3 from the spacer, it forms a new second-level branch.

Thus, in such fully developed hierarchy, a spacer is defined by its membership in a tree, in a first-level branch, and in a second-level branch or "final" cluster. The hierarchical scheme allowed us to substantially speed up the search of the target cluster for each new spacer.

This clustering procedure is repeated several times from the beginning, taking into account the results of the previous rounds of clustering. A new round starts with clustering of spacers, which belong to the largest final cluster of the largest branch of the largest tree. Next, spacers from the second largest cluster are re-clustered, etc. After the second iteration the cluster tree does not change significantly. Naturally, some of the clusters may have final radii smaller than the threshold value of 3, while others may contain spacers that are further than 3 substitutions away from the center of their cluster. The latter happens when a spacer, initially within the distance of 3 from the center, becomes further separated as the center moves away due to subsequent addition of new members. We surmise that such "swelling" of clusters has little effect on the final result since if such swollen clusters were broken, most probably, they would have merged during the second stage of clustering.

Shifting, flipping, and merging clusters

The first procedure allows us to reduce the amount of data, which is now represented by sizes and coordinates of centers of a few thousand clusters with radii ≈ 3. Next, we compute pairwise distances between all clusters, taking into account possible reversions (Erdmann & Garrett 2012; Lopez-Sanchez *et al.* 2012; Mick *et al.* 2013; Shmakov *et al.* 2014) and shifts of their sequences. When comparing one cluster to another, we first compute the distance between two sequences in their original form, then for one sequence shifted by ± 1 and ± 2 bases, and finally we "flip" one sequence, generating a reverse complement sequence and repeat the procedure, looking for the best match. Flips have no distance penalty, but a shift by a single base in either direction adds a 2 to the distance between clusters. In the end, we compute the adjacency matrix of the complete graph where nodes are clusters and edges are labeled by distances between nodes. For a given cutoff distance D, all edges with distances larger than D are removed, normally breaking the complete graph into several disconnected components. Each component is then declared to be a secondary cluster, characterized by its center and the number of constituent spacers. Naturally, the smaller threshold D yields more such secondary clusters; the plot of the number of secondary clusters N vs. D is shown in Fig. S1.

It follows from Fig. S1 that for $5 \leq D \leq 10$, the dependence of N on D is the weakest, which suggests that the natural inter-cluster separation falls into this range. For final clustering of our data, we chose $D = 7$ which is in the middle of this range.

Concluding remarks

Overall, our clustering method offers two main advantages for large CRISPR spacer sets analysis:

- It is significantly faster.

- Compared to clustering based on pairwise BLAST scores, it naturally and simply shows the sequence composition of each cluster and reveals the variability of each nucleotide within the cluster.

References

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

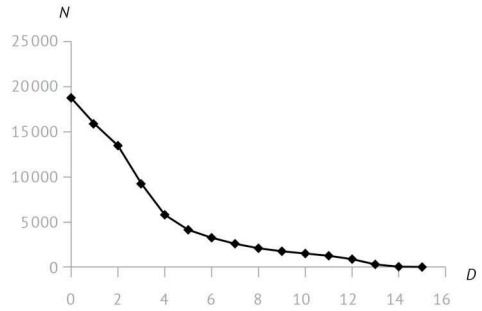
Lopez-Sanchez MJ, Sauvage E, Da Cunha V *et al.* (2012) The highly dynamic CRISPR1 system of Streptococcus agalactiae controls the diversity of its mobilome. *Molecular Microbiology*, **85**, 1057–1071.

Mick E, Stern A, Sorek R (2013) Holding a grudge: persisting anti-phage CRISPR immunity in multiple human gut microbiomes. *RNA Biol*, **10**, 900–906.

Savitskaya E, Semenova E, Dedkov V, Metlitskaya A, Severinov K (2013) High-throughput analysis of type I-E CRISPR/Cas spacer acquisition in E. coli. *RNA biology*, **10**, 716–25.

Shmakov S, Savitskaya E, Semenova E *et al.* (2014) Pervasive generation of oppositely oriented spacers during CRISPR adaptation. *Nucleic Acids Research*, **42**, 5907–5916.

Figure S1. A plot showing the dependence of the number of secondary clusters N vs. the cutoff distance between clusters D.



CHAPTER II

Metagenomic Analysis of Bacterial Communities of Antarctic Surface Snow

Introduction:

In this chapter, the CRISPRome sequencing was performed to complement standard metagenomics approaches, such as 16S amplicon sequencing and metagenome sequencing, to study uncultured bacterial communities of surface snow around four Antarctic stations. Four sampling sites demonstrated different bacterial composition with Flavobacterium genus being one of the most abundant. CRISPR repeats of Flavobacterium, detected in metagenomics reads, were used to construct degenerate primers for CRISPRome amplification. The approach developed in Chapter I was adapted here to study a distinct type of CRISPR-Cas system, the subtype II-C. Analysis of similarities between three sampling sites allowed us to associate the diversity of spacers with geographical distance.

Contribution:

Using the output of automatic metagenome annotation software (MG-RAST), I compared the bacterial composition of different sampling sites (Figure 3b) and generated PCA plots with the STAMP program (Figure 4). My main contribution was in analysis of the CRISPRome data. I applied the spacer extraction and clustering pipelines developed in Chapter I to flavobacterial CRISPRome data. Intersection of spacer diversity between different sites and sequence databases was determined by BLASTN (Figure 5). Protospacers in flavobacterial genomes, viruses and plasmids was found with BLASTN (Figure 6). The first author and the corresponding author wrote the manuscript.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 31 March 2016
doi: 10.3389/fmicb.2016.00398

Metagenomic Analysis of Bacterial Communities of Antarctic Surface Snow

Anna Lopatina[1,2,3], *Sofia Medvedeva*[2,4], *Sergey Shmakov*[4], *Maria D. Logacheva*[5], *Vjacheslav Krylenkov*[6] *and Konstantin Severinov*[1,3,4]*

[1] Department of Molecular Genetics of Cell, Institute of Molecular Genetics, Russian Academy of Sciences, Moscow, Russia, [2] Department of Molecular Genetics of Microorganisms, Institute of Gene Biology, Russian Academy of Sciences, Moscow, Russia, [3] Research Complex of "Nanobiotechnology", Saint-Petersburg State Polytechnical University, Saint-Petersburg, Russia, [4] Center for Data-Intensive Biomedicine and Biotechnology, Skolkovo Institute of Science and Technology, Skolkovo, Russia, [5] Belozersky Institute of Physico-Chemical Biology, Moscow State University, Moscow, Russia, [6] Department of Botany, Saint-Petersburg State University, Saint-Petersburg, Russia

The diversity of bacteria present in surface snow around four Russian stations in Eastern Antarctica was studied by high throughput sequencing of amplified 16S rRNA gene fragments and shotgun metagenomic sequencing. Considerable class- and genus-level variation between the samples was revealed indicating a presence of inter-site diversity of bacteria in Antarctic snow. *Flavobacterium* was a major genus in one sampling site and was also detected in other sites. The diversity of flavobacterial type II-C CRISPR spacers in the samples was investigated by metagenome sequencing. Thousands of unique spacers were revealed with less than 35% overlap between the sampling sites, indicating an enormous natural variety of flavobacterial CRISPR spacers and, by extension, high level of adaptive activity of the corresponding CRISPR-Cas system. None of the spacers matched known spacers of flavobacterial isolates from the Northern hemisphere. Moreover, the percentage of spacers with matches with Antarctic metagenomic sequences obtained in this work was significantly higher than with sequences from much larger publically available environmental metagenomic database. The results indicate that despite the overall very high level of diversity, Antarctic Flavobacteria comprise a separate pool that experiences pressures from mobile genetic elements different from those present in other parts of the world. The results also establish analysis of metagenomic CRISPR spacer content as a powerful tool to study bacterial populations diversity.

Keywords: CRISPR, Antarctica, microbial diversity, genetics, metagenomics

OPEN ACCESS

Edited by:
Mark Alexander Lever, ETH Zürich, Switzerland

Reviewed by:
Casey R. J. Hubert, University of Calgary, Canada
James A. Coker, University of Maryland University College, USA

***Correspondence:**
Konstantin Severinov
severik@waksman.rutgers.edu

Specialty section:
This article was submitted to Extreme Microbiology, a section of the journal Frontiers in Microbiology

Received: *26 November 2015*
Accepted: *14 March 2016*
Published: *31 March 2016*

Citation:
Lopatina A, Medvedeva S, Shmakov S, Logacheva MD, Krylenkov V and Severinov K (2016) Metagenomic Analysis of Bacterial Communities of Antarctic Surface Snow. Front. Microbiol. 7:398. doi: 10.3389/fmicb.2016.00398

INTRODUCTION

Snow covers about 35% of the Earth's surface—permanently or for varying times during the year—and is thus a major climatic and ecological system (Miteva, 2008). It directly affects climate, moisture budget and sea level, and also serves as an interface between different ecosystems (Pomeroy and Brun, 2001; Davis et al., 2005; Zhang, 2005; Hinkler et al., 2008). Snow ecosystems are characterized by harsh conditions such as low temperatures, low atmospheric humidity, low liquid water availability, and high levels of radiation (Cowan and Tow, 2004). The amount of microorganisms on the surface snow varies from 10^2 cells per milliliter of melted snow

on South Pole (Carpenter et al., 2000) to 10^2–10^5 in high mountain and Arctic snow (Segawa et al., 2005; Amato et al., 2007; Liu et al., 2009; Harding et al., 2011). Bacterial diversity from Arctic and alpine snow has been intensively investigated during the last few decades (Blank et al., 2002; Bachy et al., 2011; Varin et al., 2012; Hell et al., 2013; Maccario et al., 2014). Bacteria of several phylogenetic groups have been detected; most were of *Alphaproteobacteria*, *Betaproteobacteria, Gammaproteobacteria, Firmicutes, Bacteroidetes*, and *Actinobacteria* classes (Segawa et al., 2005; Amato et al., 2007; Møller et al., 2013; Maccario et al., 2014; Cameron et al., 2015). Recently, a metagenomic study of Arctic spring snow suggested that snow bacteria can be adapted to photochemical reactions and oxidative stress in addition to cold stress (Maccario et al., 2014), and therefore may form specific communities.

Microorganisms on the surface snow in Antarctica were also analyzed (Carpenter et al., 2000; Brinkmeyer et al., 2003; Christner et al., 2003; Fujii et al., 2010; Lopatina et al., 2013). Representatives of *Proteobacteria*, *Bacteroidetes*, *Cyanobacteria*, and *Verrucomicrobia* have been detected in different sampling sites (Brinkmeyer et al., 2003; Lopatina et al., 2013). Antarctic snow microbial communities have been found to be metabolically active based on the measurements of radioactive thymidine and leucine incorporation (Carpenter et al., 2000; Lopatina et al., 2013). Microbial activity on the surface snow of Dome C was also suggested by the presence of exopolysaccharide-like debris on the DAPI-stained filters and by scanning electron microscopy (Michaud et al., 2014). Also, evidence of active microbial life in the coastal snow of Antarctica was gained during analysis of "red snow" bacterial composition, which was dominated by green alga, producing pigment astaxanthin (Fujii et al., 2010).

Comparative metagenomic analysis of Antarctic show has not been undertaken so far. Availability of such data, particularly from multiple sampling sites, could reveal the presence of particular snow-specific communities or, conversely, point to introduction of snow microorganisms through eolian effects. Here, we performed amplicon library and metagenomic analysis of bacterial sequences from Antarctic snow collected around four Russian stations in Eastern Antarctica. The results reveal very considerable variation between the sites and show clear evidence of deposition of marine bacteria in stations close to open water. We also performed metagenomic analysis of CRISPR spacers in a *Flavobacterium* common in Antarctic snow. The results revealed, surprisingly, a staggering diversity of CRISPR spacers that is distinct from the limited known diversity of flavobacterial spacers from the Northern hemisphere, suggesting that diversity of flavobacterial CRISPR spacers is generated and maintained locally in response to specific genetic parasites.

METHODS

Study Sites

Samples were collected during the austral summer of 2009–2010 year from vicinity of four coastal Russian Antarctic stations—Progress, Druzhnaja, Mirnii, and Leningradskaja as described previously (Lopatina et al., 2013). All stations are located on the coastal part of Eastern Antarctica (**Figure 1**). The distance between stations ranges from ∼150 km between Progress and Druzhnaja to ∼3000 km between Progress and Leningradskaja. The stations vary in indicators of climatic conditions, such as average temperature, humidity and wind speed as shown in **Table 1**.

Total DNA Extraction, Amplification of 16S rRNA Genes, and Sequencing

Samples of total DNA were prepared as described previously (Lopatina et al., 2013). PCR of a bacterial 16S rRNA gene fragment (V3-V4 region) was performed with two universal primers 341F (5′-CCTACGGGNGGCWGCAG-3′) and 805R (5′-GACTACHVGGGTATCTAATCC-3′) under general conditions described by Herlemann et al. (2011). 2 ng of total DNA was used as a template for each PCR reaction. To avoid biases during PCR amplification 10 replicates of each PCR reactions were performed for every sample and mixed prior to further manipulations. Amplicons were visualized on 1% ethidium bromide stained agarose gels and purified using Promega Gel extraction kit according to the manufacturer's instructions. Negative controls (an aliquot of 10 l of Milli Q water subjected to concentration and DNA purification for each sample) resulted in no visible amplification products, confirming that our sample collection and processing techniques were essentially free of contamination. Pair-end sequencing was carried out on Illumina MiSeq platform with MiSeq reagent kit v.2 (Illumina, USA) as described previously (Caporaso et al., 2011).

Sequencing of Metagenomic DNA Libraries

For metagenomic sequencing 100 ng of total DNA from each sample was used to prepare libraries as described previously (Caporaso et al., 2011). Pair-end sequencing was carried out on Illumina MiSeq platform with MiSeq reagent kit v.2 (Illumina, USA).

Analysis of 16S rRNA Gene and Metagemic Libraries

Reads produced by sequencing of 16S rRNA amplicons were subjected to basic trimming (Schloss et al., 2011). First, sequences were demultiplexed, trimmed by quality with Phred score $\geq$ 20 and no admission of ambiguous bases using CLC Genomics 7.0 workbench software (CLC Bio Aarhus, Denmark), and sequences longer than 100 bp were taken for further processing. Homopolymers longer than 8 nt were removed using NGS QC toolkit with HomoPolymerTrimming.pl Perl script (Patel and Jain, 2012) and chimeric sequences were removed using Ribosomal Database Project (RDP) chimera check pipeline (Edgar et al., 2011). Phylotyping and statistical analysis was performed using the RDP classifier via taxonomic supervised method with 80% confidence threshold cut off (Cole et al., 2014), as this approach allows rapid and extensive community comparison (Sul et al., 2011).

Raw reads from shotgun metagenomic sequencing were trimmed by quality with Phred score $\geq$ 20 and no admission of ambiguous bases. Adapters were trimmed using CLC Genomics workbench software (CLC Bio Aarhus, Denmark); reads longer than 50 bp were subjected to further analysis. Trimmed

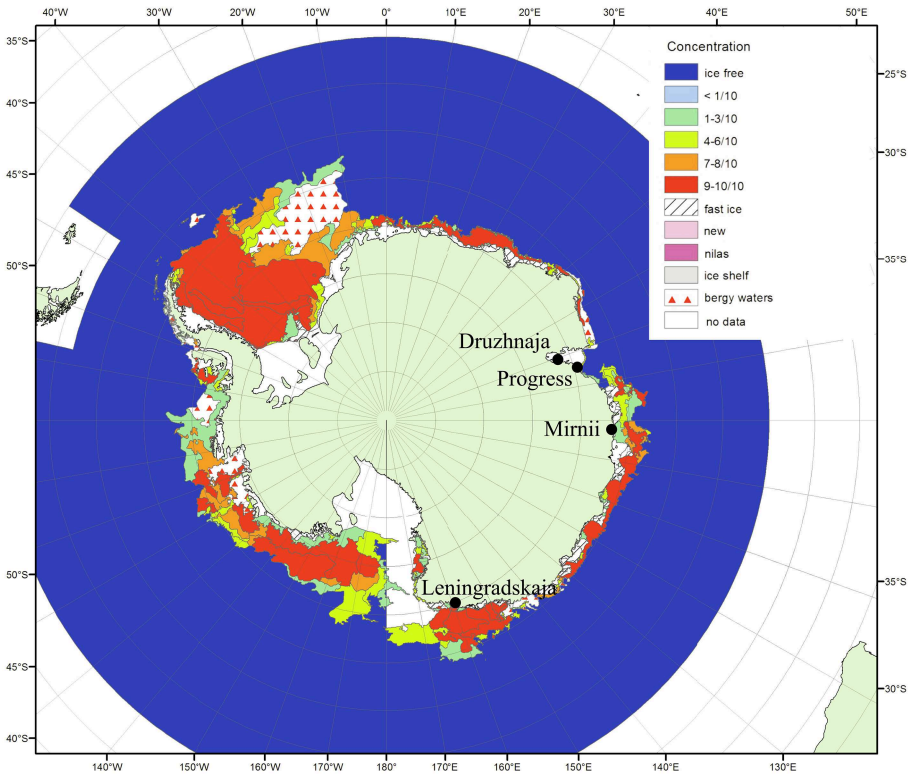


FIGURE 1 | Antarctic surface snow sampling sites. The locations of the four Russian research stations where the snow samples were taken are shown on the map of Antarctica (from the archive of Russian Institute of Arctic and Antarctica http://wdc.aari.ru/datasets/d0040/antarc/png/). The color code indicates ice concentration for January 2010 during the time of sampling. The distances from open water for Mirnii and Progress are 1–5 km, for Druzhnaja—150 km, for Leningradskaja—400 km.

TABLE 1 | Geographical and climatic data for the four sampling sites.

Station	Geographic coordinates	Elevation, m	Mean surface air T, °C	Mean ground T, °C	Mean precipitation, mm	Mean surface wind, m/s
Druzhnaja	69°44′S 72°42′E	No data	No data	No data	No data	No data
Leningradskaja	69°30′S 159°23′E	291	−14.6	−15.4	58.4	8.4
Mirnii	66°33′S 93°01′E	39,9	−11.3	−11.7	43.8	11.3
Progress	69°23′S 76°23′E	14,6	−9.2	−7.4	12.5	5.9

sequences were applied to MG-RAST database (Meyer et al., 2008). Reads were taxonomically and functionally annotated by similarity searching against M5NR database and Subsystems database, respectively, with default parameters (maximum *e*-value cutoff of 10^{-5}, minimum identity cutoff of 60% and minimum alignment length cutoff of 15).

To specifically search for viral sequences in metagenomic libraries, sequences were subjected to Metavir online tool (Roux et al., 2014), where they were blasted against Viral Refseq database (NCBI). Obtained affiliated sequences were filtered from bacterial homologs using supplementary pipeline: firstly, they were blasted against nucleotide (nt) database using blastn standalone application and afterwards viral sequences were extracted using Megan 5.10.1 software (Huson et al., 2011).

Statistical Analysis

Several measurements of alpha diversity were used to estimate the diversity of bacteria in the samples. Species richness estimators S_{chao1} and S_{ace} (Kemp and Aller, 2004b), and community diversity indices Shannon (1948) and Simpson (1949) were calculated using RDP analysis tools. Coverage of 16S rRNA libraries was calculated according to Good's formula: C = 1 – (N/individuals), where N is the number of sequences that occurred only once (Kemp and Aller, 2004a).

Identification and Analysis of CRISPR Arrays

To construct a set of CRISPR arrays for each metagenomic dataset we used CRASS algorithm (Skennerton et al., 2013) with default parameters: repeat lengths 23–47 bp, spacer lengths 26–50

bp, and minimum three spacers in array as default parameters. Spacer and repeat sequences were compared with nucleotide (nt) database using BLAST+ tool installed on Galaxy platform with default parameters for short input sequence (word size 7, gapopen 5, gapextend 2, reward 2, penalty -3, *e*-value 0.01). Repeat sequences from identified CRISPR arrays were classified using CRISPRmap tool (Lange et al., 2013). The *cas* genes search was performed using MG-RAST Subsystems annotation tool (Meyer et al., 2008).

To amplify CRISPR arrays of *Flavobacterium psychrophilum* from total DNA samples primers Flavo_F (CAAAATTGTATTTTAGCTTATAATTACCAAC) and Flavo_R (TACAATTTTGAAAGCAATTCACAAC) were used. Amplification reactions were carried out with Taq DNA polymerase under the following conditions: initial denaturation for 5 min at 95°C, followed by 28 cycles of 30 s at 95°C, 30 s at 55°C, and 40 s at 72°C, and a final extension at 72°C for additional 2 min. Amplicons were visualized on 1% ethidium bromide stained agarose gels and DNA fragments of 200–1000 bp in length were purified from the gel and sequenced on Illumina MiSeq platform as described above. Raw reads were demultiplexed, trimmed by quality with Phred score $\geq$ 20 and no admission of ambiguous bases using CLC Genomics 7.0 workbench software (CLC Bio Aarhus, Denmark).

Spacers from amplified CRISPR arrays were bioinformatically extracted using DNAStringSet function of IRanges package in R. To decrease the amount of spacers and to avoid overrepresented diversity because of mistakes during sequencing, spacers were clustered using a *k*-means algorithm (MacQueen, 1967). The maximum number of substitutions corresponding to biologically similar spacers within one cluster was equal to 5. Coverage and diversity estimates S_{chao} and S_{ace} for total amount of spacers or clusters in each sample were calculated with estimateD function of vegan package in R. Centers of spacer clusters (sequences of mean arithmetic value for each nucleotide position calculated from all spacers within a cluster) were compared against nucleotide collection (nt) and environmental collection (env_nt) databases, as well as against custom-made database containing sequences from Antarctic shotgun metagenomic libraries from the present work, with BLASTn algorithm using default parameters for short input sequences mentioned above and an *e*-value cut off of 0.01. Sequences with <5 mismatches were considered as positive hits. Metagenomic sequences containing protospacers were blasted against nt and nr databases with default parameters for BLASTn algorithm and an *e*-value cut off of 0.001 using BLAST+ tool installed on Galaxy platform. PAM searches were performed with CRISPRTarget online tool (Biswas et al., 2013). Eight nucleotides upstream and downstream of each protospacer were extracted and used for PAM logo search with Weblogo online tool (http://weblogo.berkeley.edu/logo.cgi).

Data Access

The data of 16S rRNA high throughput sequencing were deposited to MG-RAST database under accession numbers 4616914.3 (Druzhnaja), 4616915.3 (Leningradskaja), 4616916.3 (Mirnii), and 4616917.3 (Progress). The data of shotgun metagenomic sequencing were deposited to MG-RAST database under accession numbers 4624083.3 (Druzhnaja), 4624084.3 (Leningradskaja), 4624085.3 (Mirnii), and 4624086.3 (Progress).

RESULTS

Metagenomic Analysis of 16S rRNA Sequences from Antarctic Snow Samples

Earlier, we studied the bacterial diversity of surface snow from two Russian Antarctic stations, Leningradskaja and Druzhnaja, by analyzing individual 16S rRNA gene fragments cloned after PCR amplification of DNA from melted snow samples collected during the 54th (2009) and 55th (2010) Russian Antarctic expeditions (Lopatina et al., 2013). For the present work, we used high-throughput sequence analysis of 16S rRNA amplicons from Leningradskaja and Druzhnaja 55th expedition samples analyzed previously and also included samples collected at the Progress and Mirnii stations during the same time. The microbial diversity at the two latter stations was not analyzed before, however, the biological activity of snow collected at Mirnii was at least 10 times higher than in the Leningradskaja and Druzhnaja samples (Lopatina et al., 2013). For Progress, bioactivity levels were low (4.4 pmol/h∗l of [methyl ^{3}H] thymidine incorporation and 33.1 pmol/h∗l of [^{3}H] L-leucine incorporation) and comparable to those in Leningradskaja and Druzhnaja samples.

DNA concentration was estimated by measuring absorbance by NanoDrop yielding a concentration estimate of 1, 1, 2, and 14 ng/μl for Druzhnaja, Leningradskaja, Progress and Mirnii samples, correspondently. To access bacterial diversity in snow samples, a fragment of bacterial 16S rRNA gene was amplified from total DNA following by Illumina pair-end high throughput sequencing (HTS). The overall sequencing statistics are presented in Table S1. Results of phylogenetic analysis of 16S rRNA sequences from Leningradskaja and Druzhnaja samples generated by HTS and Sanger sequencing of cloned libraries were first compared. Overall, comparisons of class-level distribution revealed by both methods are in very good agreement with each other (**Figure 2B**; Pearson coefficient of correlation for Druzhnaja sample–0.99, for Leningradskaja–0.95). Yet, for both stations, HTS analysis revealed increased relative abundance (or even appearance) of several minor classes, including *Flavobacteriia*, *Alphaproteobacteria*, *Sphingobacteriia*, *Cytophaga*, and *Actinobacteria*.

16S rRNA gene sequences recovered by HTS from the four stations fell into 34 classes based on RDP classification. 3.4, 3.9, 4.5, and 4.3% of 16S rRNA gene reads from, correspondingly, Druzhnaja, Leningradskaja, Mirnii, and Progress samples could not be affiliated to any known bacterial class by the RDP classification tool. Overall, the most abundant classes were: *Alphaproteobacteria*, *Betaproteobacteria*, *Gammaproteobacteria*, *Sphingobacteriia*, *Flavobacteriia*, *Cytophagia*, *Actinobacteria*, *Chloroplast/Cyanobacteria, Bacilli*. While *Betaproteobacteria* were dominant in Leningradskaja, Druzhnaja, and Mirnii samples, *Flavobacteriia* were the major class in the Progress sample, constituting 40% of all sequences (**Figure 2A**). In fact, the latter sample was clearly very different in composition from

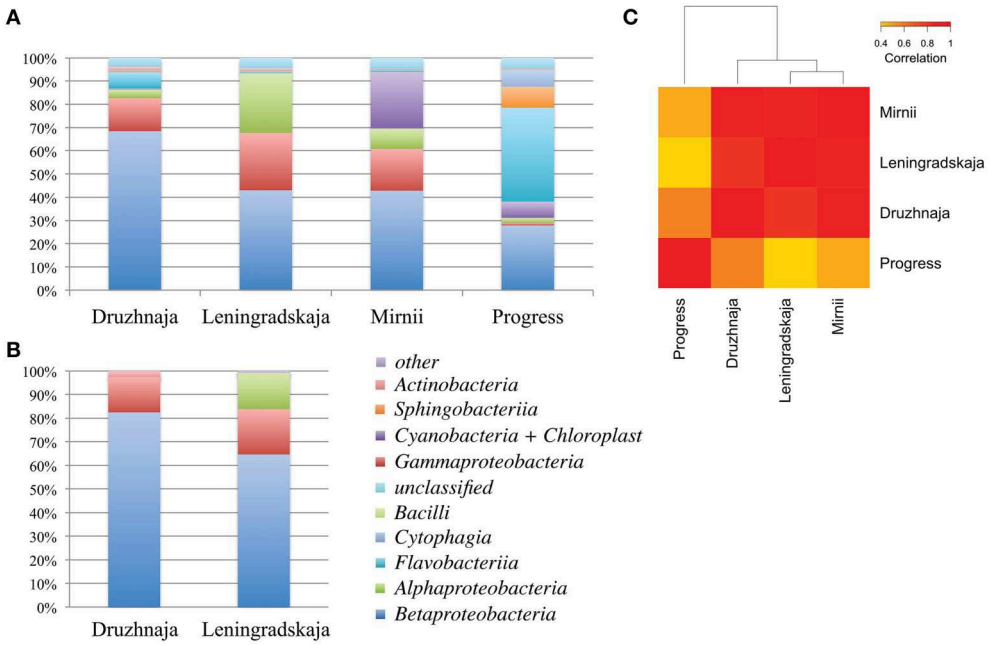


FIGURE 2 | Class-level bacterial diversity in Antarctic snow samples. (A) Relative abundance of class-level bacterial taxonomies based on PCR amplifications and high-throughput sequencing of 16S rRNA gene fragments is shown at the top. "Other" group includes minor classes with <0.6% of total abundance, namely *Verrucomicrobiae*, *Clostridia*, *Gemmatimonadetes*, *Planctomycetia*, *Deltaproteobacteria*, *Spartobacteria*, *Epsilonproteobacteria*, and 14 rare classes found only in one location. **(B)** Similar data using clone library approach for same samples from Druzhnaja and Leningradskaja stations are shown. **(C)** A heatmap comparing class-level bacterial diversity and abundance among the four samples based on high-throughput sequencing results. The colors show the extent of relatedness between the samples as indicated in the legend.

the first three based on Pearson correlation analysis at class level (**Figure 2C**).

Deeper taxonomic affiliation analysis at each site was next performed. 28, 20, 14, and 35% of 16S rRNA gene reads from, correspondingly, Druzhnaja, Progress, Mirnii, and Leningradskaja could not be affiliated to any known genus by the RDP tool. Results of the analysis of remaining reads are shown in **Figure 3A**, where abundances of 20 most prevalent genera are presented. The genus detected in the most abundance in any given sample was *Flavobacterium*, which comprised 39% of the sequences in the Progress library, followed by *Hydrogenophaga* (14%) and *Ralstonia* (7%). In the Druzhnaja sample, 16S rRNA genes from *Janthinobacterium* were dominant (27%), followed by *Ralstonia* (15%), and *Pseudomonas* (11%). In the Leningradskaja sample, 16S rRNA genes from *Caulobacter* (12%), *Acinetobacter* (10%), and *Comamonas* (9%) were most abundant. These genera were also the most abundant during clone library analysis (Lopatina et al., 2013) and in fact the abundance of genera in Druzhnaja and Leningradskaja stations, as revealed by cloning library and HTS approaches, was highly correlated (Pearson correlation coefficient of 0.8 and 0.9, respectively, data not shown). In Mirnii—rRNA gene sequences of *Ralstonia* (31%), *Bacilariophyta* (chloropast-containing diatoms) (24%), and *Rudaea* (8%) were the most dominant. There was no correlation of genera abundance or presence between samples from the four different stations: the Pearson correlation coefficient varied from 0.1 for Progress and Leningradskaja to 0.4 between Mirnii and Druzhnaja (**Figure 3B**).

Shotgun Metagenomic Analysis of Antarctic Snow DNA Samples

DNA samples from the four stations were also subjected to shotgun metagenomic sequencing. The summary of data obtained from four snow samples is shown on Table S2. Sequences that passed the QC criteria were applied to Best hit classification algorithm of the MG-RAST software using M5NR database for phylogenetic affiliation of sequences. The results are summarized in **Table 2**. The percentage of archaeal sequences in shotgun metagenomic libraries was consistently low in all stations (<0.2% of all sequences) and these sequences were not further analyzed; no archaeal sequences were obtained previously in clone 16S rRNA libraries in Druzhnaja and Leningradskaja samples (Lopatina et al., 2013). Viral samples were extracted from metagenomic data through Metavirome tool and were also rare. *Eukaryota* were well-represented in Mirnii library—15% of all sequences. Samples from other stations contained much less eukaryotic sequences (∼1% or less). More than half of eukaryal sequences from Mirnii were from *Bacilariophyta*, suggesting that "cyanobacterial" sequences present in the amplified 16S rRNA gene samples from this station were actually of chloroplast origin. The Mirnii and Progress stations are located within 1–5 km of open water, while Druzhnaja and Leningradskaja,

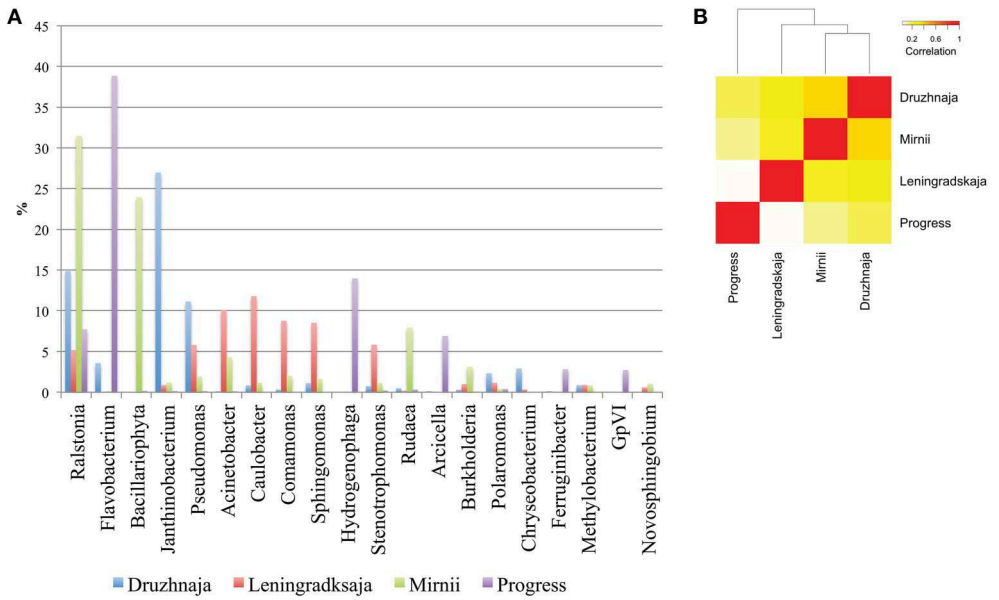


FIGURE 3 | Genus-level bacterial diversity in Antarctic snow samples based on PCR amplicon library. (A) Frequencies of reads corresponding to 20 most abundant genera present in all four samples are shown. For each genus, the height of color-coded bars reflects the percentage of corresponding reads in the entire sample from each station. **(B)** A heatmap comparing genus-level bacterial diversity and abundance for 255 genera detected in Antarctic snow samples. The colors show the extent of relatedness between the samples from each station as indicated in the legend.

are, respectively, about 150 and 400 km away (**Figure 1**). The abundance of *Chloroplasts*/*Cyanobacteria* is thus probably correlated with closeness to open water. Most of metagenomic sequences from all samples corresponded to domain *Bacteria*. Class- and genus-level phylogenetic complexity of bacterial sequences from shotgun and 16S rRNA metagenomic data matched well for all four stations (Pearson coefficient values 0.97–0.99 for class level and 0.68–0.86 for genera level).

Protein-coding sequence reads from snow metagenomes were classified to metabolic functions based on Subsystems database using MG-RAST software. The most abundant functional groups were related to housekeeping functions, such as clustering-based subsystems (functional coupling evidence but unknown function; 14-16%), carbohydrate metabolism (9%), amino acid biosynthesis (8%), and protein metabolism (6.5–8.5%). Stress response related genes constituted 2.3–2.9% of all annotated reads and within this group there was a high proportion of oxidative stress genes (43–44%). Genes of photosynthesis and respiration were clearly more abundant at Mirnii station, where chroloplast/cyanobacterial sequences were common.

Recently, principal component analysis of the relative abundance of annotated reads of functional subsystems from Arctic surface snow metagenomes was presented and a conclusion was made that snow samples grouped together and were well-separated from other ecosystem metagenomes (Maccario et al., 2014). We repeated this analysis including our Antarctic snow metagenomes data. When Antarctic samples were substituted for Arctic samples used in the previous analysis, clear ecosystem clustering similar to the earlier reported result was obtained (**Figure 4A**), seemingly indicating commonalities of microbial communities of Antarctic snow. However, when Arctic snow metagenomic samples were also included, Antarctic samples became indistinguishable from soil and Antarctic microbial mat metagenomes; the free ocean water samples remained tightly clustered and separate, while the Arctic snow samples became very dispersed (**Figure 4B**).

Analysis of CRISPR-Cas Sequences in Antarctic Metagenomes

The CRISPR-Cas systems of adaptive prokaryotic immunity are widespread in bacteria (Marraffini and Sontheimer, 2010; Makarova et al., 2011) and are highly dynamic (Deveau et al., 2008), allowing one, in principle, to monitor the structure of bacterial populations in environment (Bhaya et al., 2011; Sun et al., 2015). We searched for *cas* genes and CRISPR arrays fragments in sequences from our shotgun metagenomic libraries. The *cas* genes of all three CRISPR-Cas system types were found. Specifically, fragments of *cas1, cas2, cas3*, *csn1* (*cas9*) as well as *cas4b* and *cmr1-6* genes were detected. These reads constituted less than 0.03% of all sequences. Fragments of CRISPR arrays were also identified in every library. Some identified repeats matched previously described ones, for example a 46-bp repeat from type II CRISPR-Cas system from *Flavobacterium psychrophilum* (Touchon et al., 2011), found in Progress and Druzhnaja, and a different type II 36-bp repeat matching *Flavobacterium columnare* in Leningradskaja and Progress samples. A type I-F CRISPR-Cas system repeats from *Yersinia pseudotuberculosis* were found in Druzhnaja, Leningradskaja, and Progress (Table S3). CRISPRmap, an automated tool for classification of prokaryotic repeats based on sequence and structure conservation, has been reported to classify as many as 30–40% of repeat sequences from human

TABLE 2 | Overall phylogenetic structure of snow microbial communities.

Station	Prokaryota, %	Eukaryota, %	Viruses, %	Archaea, %	Unclassified, %
Druzhnaja	98.29	1.41	0.06	0.1	0.15
Leningradskaja	99.04	0.77	0.06	0.06	0.08
Mirnii	84.65	14.97	0.14	0.17	0.20
Progress	98.38	1.22	0.04	0.31	0.32

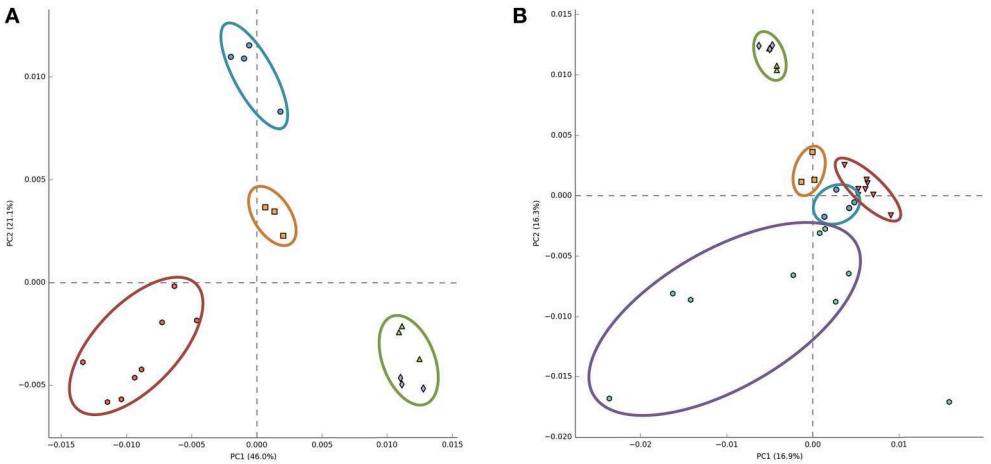


FIGURE 4 | Principal component analysis of 30 metagenomes from five different environments based on frequencies of COG categories. A scatter plot of PCA-scores depicting variance of COG categories detected in different environmental metagenomes. In panel **(A)** COGs from Antarctic snow (cyan) are compared to temperate soil samples (red), sea water (green), and Antarctic microbial mat (orange). In panel **(B)** Arctic snow samples (violet) are included in comparison.

microbiome samples (Lange et al., 2013). In contrast, in the case of Antarctic samples out of a total of 40 distinct repeats identified, only one could be matched with a known family (six could be matched with a known structural motif), indicating that the variety of existing adaptive immunity systems is greatly underexplored.

When spacers extracted from identified Antarctic CRISPR arrays were analyzed, no matches with spacers of previously known CRISPR arrays was detected. Further, when the entire collection of 570 unique spacers recovered from Antarctic snow metagenomic libraries was analyzed against the NCBI nucleotide collection (nt), only a single hit, for a spacer associated with the *F. columnare*-like 36-bp repeat, was found. This spacer matched exactly a fragment of 16S rDNA sequence of another representative of the *Flavobacterium* genus, *Flavobacterium* sp. 136G (NCBI accession number KM021132.1), contrary to the general observation that CRISPR spacers target DNA of mobile genetic elements.

CRISPR interference in type II systems requires a functional protospacer adjacent motif (PAM), located downstream of the protospacer (Chylinski et al., 2014). The PAM sequence of *F. columnare* type II CRISPR-Cas system is not known. Analysis of 43 spacers from CRISPR array of a sequenced *F. columnare* genome (NCBI accession number CP003222.2) revealed four matches with flavobacterial phage FCL-2 protospacers. Sequences adjacent to these protospacers contained a TAA trinucleotide five nucleotides downstream of each protospacer. Both the downstream location of the putative PAM, and its separation from protospacers by a string of non-conserved nucleotides is typical for type II CRISPR-Cas systems (Chylinski et al., 2014). The putative PAM sequence was absent downstream of the *Flavobacterium* sp. 136G 16S rDNA sequence matching the spacer identified from metagenomic data. Thus, the particular 16S rDNA targeting spacer may not be functional (see, however, below). Three spacers—associated with the *F. psychrophilum* 46-bp repeat—were found in both Progress and Druzhnaja samples. The rest of the spacers were unique for each station. Since flavobacterial rRNA was present in samples from all spacers, we were interested in assessing diversity of *F. psychrophilum* spacers in each site. To this end, PCR primers matching 46-bp repeat were designed and used to amplify spacers from each snow community DNA (**Figure 5A**). By design, the procedure allows amplification of spacers associated with the 46-bp repeat, however the information about the order of the spacers in CRISPR arrays is lost. Amplification products were detected in samples from three stations—Progress, Druzhnaja, and Leningradskaja. The amplified material was subjected to Illumina sequencing. A total of ~870,000 spacers with an average length of 30 ± 2 nucleotides was obtained (in published *F. psychrophilum* genomes spacers are 28–31 long). We next clustered spacers in each sample (MacQueen, 1967), combining spacers that differ from each other by <5 nucleotides in the same cluster. After clustering, 2759 unique spacer clusters remained in Leningradskaja, 2584—in Druzhnaja, and 3822—in Progress

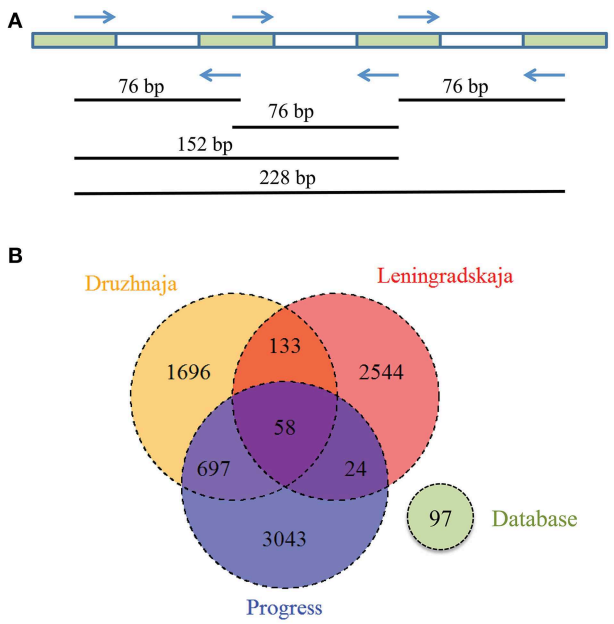


FIGURE 5 | Analysis of *Flavobacterium physchrophilum* CRISPR spacers in Antarctic snow samples. (A) A strategy used to amplify spacers associated with *F. physchrophilum* CRISPR repeats from environmental samples, the length of amplified fragment corresponds to the number of containing spacers. **(B)** A Venn diagram showing the number of shared and unique clusters of spacers associated with *F. physchrophilum* CRISPR repeats in three Antarctic snow samples. Known *F. physchrophilum* spacer clusters from 10 publically available genomes are also shown ("database").

station (**Table 3**, Supplementary Dataset S4). The calculated coverage of the three cluster libraries ranged from 40% for Druzhnaja to 61% for Progress samples (**Table 3**), so true variety in samples was thus 1.5–2.5 times higher than the actual number of clusters obtained. It therefore follows that the diversity of CRISPR spacers associated with the *F. psychrophilum* 46-bp repeat (and, by extension, of *F. psychrophilum*) in Antarctic snow is extremely high. When spacers from each station were compared to each other, only 58 clusters (0.7% of the total) were common for all three stations (**Figure 5B**). The percentage of clusters unique to each station varied from 66% for Druzhnaja to 92% in Leningradskaja. The Druzhnaja spacer set was most similar to Progress (about 30% of common spacers), with much smaller (<7%) overlap with Leningradskaja set. The overlap of Progress and Leningradskaja sets was just 3%. Ninety-five percent of all spacers were located within 14, 29, and 21% of clusters from Progress, Leningradskaja, and Druzhnaja, correspondently, i.e., were highly overrepresented. Bacteria with such spacers must be highly abundant in the samples. Alternatively, overrepresented spacers may be shared between many strains.

A small fraction (1–3%) of self-complementary spacers derived from the same protospacer was observed. Such pair-mated spacers have been reported before for *Streptococcus agalactiae*, *Sulfolobus solfataricus*, and *Escherichia coli* (Erdmann and Garrett, 2012; Lopez-Sanchez et al., 2012; Shmakov et al., 2014). In most cases, when self-complementary spacers were observed, one spacer in the pair belonged to an over-represented group. A high number of such paired spacers were shared between two or more stations (up to 92% self-complementary spacers in the Druzhnaja station sample were also found in other stations).

Many reads corresponded to amplified fragments that contained two spacers and, therefore, harbored a copy of an "internal" repeat, whose sequence, by design, could not be affected by the primers used during amplification step (**Figure 5A**). Analysis of such reads revealed different repeat variants (Table S5). Similar cases of nearly identical repeats sequences were described previously for other organisms, for example, *E. coli* (Touchon and Rocha, 2010) or *H. volcanii* (Maier et al., 2013). The most abundant variant constituted 65.6% of all "internal" repeat sequences and matched the published *F. psychrophilum* repeat consensus used to design oligonucleotide primers for amplification. The second variant had one mismatch from consensus in the 6th position and constituted 34% of all "internal" repeats. Two other repeat variants had, in addition to the 6th position consensus mismatch, changes in the 13th or the 21th positions and were minor (0.2 and 0.1% of all "internal" repeats, correspondingly). The relative proportion of repeat variants was the same in libraries from the three Antarctic sites analyzed. In sequenced *F. psychrophilum* genomes a variant repeat with one mismatch from consensus in the 18th position constitutes 4% of all repeat sequences. This variant is absent from Antarctic samples.

When cluster consensus sequences from each station were compared to the NCBI nucleotide database using BLASTn algorithm a very large number of matches with likely irrelevant (i.e., eukaryotic) sequences was found. We therefore limited comparisons to a custom database containing all known sequences of *Flavobacterium* and their phages. None of Antarctic spacers matched any of the 117 unique spacers associated with 46-bp repeat from fourteen sequenced *F. psychrophilum* strains available in the Genbank (our clustering procedure combined these 117 spacers into 97 clusters). Ten Antarctic spacer clusters matched flavobacterial phages FCL-2, 6H, 11b, or 1/32, while 38 matched *Flavobacterium* chromosomes (Table S6). Interestingly, one cluster consensus sequence (leningradskaja_747) had multiple hits in various flavobacterial genomes (*F. indicum*, *F. psychrophilum*, *F. columnare,* and *F. branchiophilum*). Inspection of genomic sites that matched this spacer revealed that they are composed of non-coding 125 bp-long imperfect palindromic repeats that are spread throughout the *F. indicum* (30 copies) and *F. psychrophilum* (5 copies) genomes and are present in single copies in *F. columnare* and *F. branchiophilum* (**Figure 6A**). Analysis of distribution and genetic neighborhoods of these repeats in *F. indicum* and *F. psychrophilum* (data not shown) genomes revealed that they are clustered in regions containing multiple repeated genes of unknown function, transposes genes, and restriction-modification system genes (**Figure 6B**).

We also analyzed CRISPR cassettes from all *F. psychrophilum* isolates available in the Genbank. Twelve spacers matching flavobacterial phages 6H and 1/32 were identified among the 117

TABLE 3 | Statistics of high-throughput sequencing of PCR amplified Antarctic *Flavobacterium psychrophilum* CRISPR spacers and spacer clustering results.

Station	# of reads	# of spacers	Clusters					
			# of clusters	% of unique clusters	C_{chao1}, %	C_{ace}, %	S_{chao1}	S_{ace}
Druzhnaja	284,286	273,255	2584	65.6	48	40	6382 ± 386	5359 ± 42
Leningradskaja	321,550	313,241	2759	92.2	43	47	5824 ± 303	6477 ± 48
Progress	263,548	255,447	3822	79.6	60	61	6271 ± 170	6332 ± 46

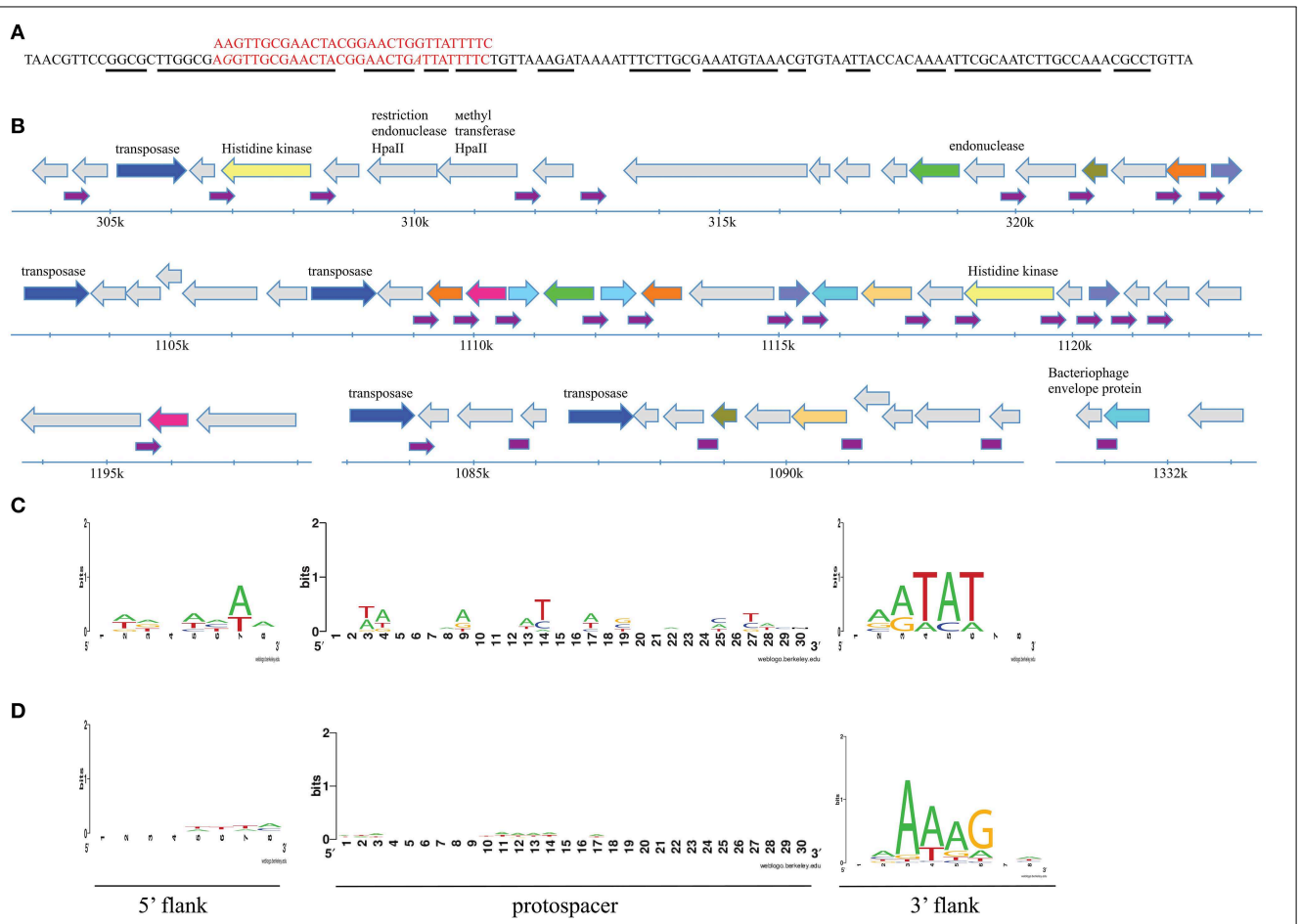


FIGURE 6 | Analysis of sequences matching spacers associated with *F. psychrophilum* CRISPR repeat. (A) The alignment of sequences of cluster consensus and 125 bp-long repeat. The aligned nucleotide positions are shown in red. Nucleotide positions in 125 bp-long repeat that participate in secondary structure formation of the transcript are underlined with a black line. **(B)** At the top, the sequence of a non-coding 125 bp-long imperfect palindromic repeat and matching *F. psychrophilum* spacer are shown. The distribution and genetic neighborhoods of 30 125 bp-long repeats in regions of *Flavobacterium indicum* genome is shown below. The repeats are indicated by small purple arrows. Genes are shown as arrows. Unlabeled gray arrows correspond to unique open reading frames with unknown function. Unique open reading frames with predicted functions are also shown in gray with annotations. Arrows marked with the same (non-gray) color correspond to genes encoding homologous proteins (more than 80% identity) of unknown function. Genes coding for IS110 family transposes are shown in blue. **(C)** A LOGO showing the results of alignment of 12 flavophage protospacers and their 8-bp flanking regions matching CRISPR spacers from *F. psychrophilum* genomes deposited in the GenBank. **(D)** As in **(C)** but showing the results of alignment of 511 sequences from Antarctic metagenomic matching Antarctic *F. psychrophilum* CRISPR spacers.

unique spacers present in *F. psychrophilum* strains sequenced to date. When flanking sequences of these protospacers were compared to each other, a likely PAM, NNATAT, downstream of protospacers was detected (**Figure 6C**). Neither 10 protospacers in the genomes of flavophages nor 38 protospacers in flavobacterial genomes matching Antarctic spacers contain such (or any other) adjacent conserved motive.

We next compared consensus sequences of Antarctic spacer clusters with metagenomic reads obtained in this work as well as with sequences from the metagenomic env_nt

database. A total of 117 hits to env_nt database and 511 hits to Antarctic reads was obtained. When the origin of 511 Antarctic metagenomic reads that contained sequences matching *F. psychrophilum* spacers was investigated, 62% of reads could not be identified by either nt or nr database searches. Of the remaining 38% of reads (corresponding to 194 cluster consensus sequences), 87 originated from flavobacterial chromosomes, 21—from *Flavobacterium* phage 11b or plasmids, 49–from other phages (mostly *Cellulophaga* phage phi10:1), and 37 originated from other eubacterial genomes. 12 and 18 additional hits to *Flavobacterium* chromosomes and flavophages, correspondingly, were obtained when reads with no matches to nt database were analyzed against the nr database. Among matching sequences in the env_nt database, there were four *Flavobacterium* chromosomes and 12 bacteriophages of various hosts. When flanking sequences of protospacers identified in Antarctic metagenomic sequences were compared to each other, an area of strong conservation 3-6 nucleotides downstream of the protospacer—NNAAAG - was detected (**Figure 6D**). This sequence is different from the putative PAM motif detected during searches with spacers from published *F. psychrophilum* genomes (NNATAT, above, **Figure 6C**) but the location of conserved positions is the same. No conservation in flanking sequences was detected for protospacers identified in metagenomic reads from the env_nt database. Neither one of the putative PAM motives is associated with protospacers from 125 bp-long imperfect palindromic repeats (above).

DISCUSSION

In this work, we significantly extended the previous analysis of surface snow microbiota around Russian research stations in Eastern Antarctica by (i) increasing the number of stations analyzed, (ii) using high-throughput sequencing to analyze 16S rRNA genes; (iii) performing metagenomic analysis of snow microbiome, and (iv) analyzing the diversity of CRISPR spacers of flavobacteria common in Antarctic snow. Analysis presented in this work was more extensive than previous limited analysis using cloned 16S rRNA genes fragments (∼50,000 sequences per each sample compared to ∼120 sequences analyzed using clone library approach). Yet, for the two stations where direct comparisons are possible, Druzhnaja and Leningradskaja, a very good correlation between the class- and genus-level composition of microbial sequences in the samples was revealed, indicating that limited sampling of clone libraries did not introduce significant biases in representation of major classes and genera. Moreover, when rRNA gene sequences were extracted from metagenomic reads and class-level phylogenetic complexity was compared with amplified 16S rRNA genes a good match was also observed (Pearson coefficient values between 0.94 and 0.98), indicating that our conditions of PCR amplification of 16S rRNA gene fragments did not introduce significant biases. HTS analysis revealed increased abundance (or even appearance) of several minor classes, including *Flavobacteriia*, *Alphaproteobacteria*, *Sphingobacteriia*, *Cytophaga*, and *Actinobacteria* in both stations These minor classes appeared at the expense of *Betaproteobacteria*, which, nevertheless still remained the major class in both samples. The result is an expected consequence of much deeper coverage obtained with HTS.

Principal component analysis of the relative abundance of annotated reads of functional subsystems from Antarctic surface snow metagenomes revealed some clustering, which, however, was found to be very sensitive to the inclusion of additional environmental samples in the analysis. As expected and recently confirmed by experimental data (Hultman et al., 2015), there is a much greater overlap in shared genes revealed by metagenomic DNA analysis compared to transcriptomic and proteomic analyses of samples from different ecosystems. Such a large overlap may explain the observed instability of results of principal component analysis of functional subsystems in Antarctic metagenomic data. Additional studies will be needed to confirm if there is a characteristic set of gene functions in snow communities.

Spoligotyping, a procedure based on comparisons of spacer sets in different strains of same bacterial species is commonly used for epidemiological tracing of pathogens (Gori et al., 2005). We reasoned that *F. psychrophilum* CRISPR arrays, if present in all four sampled Antarctic sites, may allow us to compare diversity of resident *F. psychrophilum* populations and establish relationships between them. An efficient procedure was elaborated to amplify spacer sets from environmental DNA and k-mean clustering allowed us to parcel the very large number of spacers generated after PCR amplification into a manageable number of spacer clusters. Still, a very high number of spacer clusters was observed in the samples, which is an unexpected result, since a recent report indicated that the *F. psychrophilum* CRISPR-Cas system is inactive and that the spacer content of CRISPR arrays is identical in *F. psychrophilum* isolated in geographically remote locations at different times (Castillo et al., 2015). Spacer sets present in three different Antarctic sites, where successful amplification using *F. psychrophilum* CRISPR repeat-specific primers was achieved differed significantly from each other, with only a very minor portion of spacers being common to all three sites. The larger amount of common spacers between Druzhnaja and Progress agrees with geographical proximity of these stations. Curiously, this similarity, based on common CRISPR spacers was not supported by phylogenetic analysis of bacterial communities based on 16S rRNA genes, according to which Druzhnaja was more similar to Leningradskaja station.

Despite the very large number of *F. psychrophilum* spacers uncovered in our work, no matches with spacers present in *F. psychrophilum* isolates from the Northern hemisphere available in Genbank were observed. Moreover, comparisons with environmental metagenomic data revealed that Antarctic shotgun metagenome from our work, which is orders of magnitude smaller than combined metagenomes stored in the env_nt database contains several times more hits with Antarctic *F. psychrophilum* spacers revealed during HTS analysis of amplified CRISPR spacers. The result suggests that Antarctic *F. psychrophilum* tend to acquire spacers locally. Recent evidence of genetically different pools of viruses in Southern Ocean and Northern hemisphere sampling sites (including Vancouver

Island in British Columbia, Monterey Bay, California, and Scripps Pier in San Diego, California) was recently obtained (Brum et al., 2015). The presence of such separate pools in flavophages could be responsible for observed variations in spacer content (see, however, below). The CRISPR-Cas systems of Antarctic *F. psychrophilum* and strains isolated in the Northern hemisphere may even have evolved different PAM specificities since putative PAMs revealed by comparisons of protospacers matching spacers known for the two sites result in different PAMs. Such a result is not without precedent since varying preferences for PAM selection during spacer acquisition were previously noted for type I-E CRISPR-Cas system variants from different *E. coli* strain (Westra et al., 2012) and for type I-B CRISPR-Cas system of *Haloferax volcanii* (Fischer et al., 2012). The presence of different, non-overlapping sets of CRISPR repeat polymorphisms in our Antarctic samples and in known *F. psychrophilum* CRISPR arrays also supports existence of local variations.

The original theoretical insights about the immune function of CRISPR-Cas systems came after observation of matches between spacer sequences and protospacers in bacteriophage and plasmid sequences specific to a bacterial host (Makarova et al., 2006). Later, self-targeting spacers were also identified and a regulatory function of such spacers was proposed (for detailed review, see Westra et al., 2014). Analysis of *F. psychrophilum* repeat associated spacers suggests, that at least for the Antarctic spacer set, targeting of bacteria related to the host is the most common scenario. Such targeting could help prevent genetic exchange between the species within the genus, although the biological significance of such restriction is unclear.

Previous analysis has revealed the loss of synteny within the *Flavobacterium* spp. genomes likely due to the presence of numerous repeats (e.g., insertion sequences and the *rhs* elements (McBride et al., 2009; Touchon et al., 2011). Our analysis revealed an interesting case of a CRISPR spacer with multiple hits in various flavobacterial genomes. The matching sequence was part of a non-coding 125 bp-long imperfect palindromic repeat that is spread throughout the *F. indicum* and *F. psychrophilum* genomes and is also present in single copies in *F. columnare* and *F. branchiophilum*. The location and the number of these repeats differ in different isolates of *F. psychrophilum*, suggesting that they are subject to horizontal transfer. The 125-bp repeat is distinct from either IS or *rhs* elements, however, it may play a similar role in promoting flavobacterial genome plasticity. Targeting of this element by the CRISPR-Cas system may help control the spread of such elements and is in line with an emerging theme that CRISPR-Cas systems serves as one of the mechanisms of endogenous gene regulation (Westra et al., 2014).

Our analysis of Antarctic spacers has an important caveat in that we determine the identity of spacers associated with a particular repeat and can not exclude that such a repeat (and spacers) are not coming from arrays from other, non-*F. psychrophilum* arrays. We consider this scenario unlikely since at least in Progress station, where rRNA gene sequences from *F. psychrophilum* are most abundant, the spacer variety is also the largest. Besides, the largest number of spacers with matches to metagenomic sequences match *Flavobacterium* chromosomes, which also strengthens the link between spacers identified by our approach and the *Flavobacterium* genus.

AUTHOR CONTRIBUTIONS

AL collected samples, performed experiments, analyzed data, prepared figures; SM performed clustering and PCA analysis, prepared figures; SS performed heatmap analysis; ML performed NGS sequencing; VK collected samples, organized expedition fieldwork; KS designed research, supervised the project, analyzed data, wrote the paper.

FUNDING

This work was supported by foundation of Ministry of education and science of the Russian Federation (No14.B25.31.0004) and by Russian Science Foundation (No14-14-00988). The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmicb.2016.00398

REFERENCES

Amato, P., Hennebelle, R., Magand, O., Sancelme, M., Delort, A.-M., and Barbante, C. (2007). Bacterial characterization of the snow cover at Spitzberg, Svalbard. *FEMS Microbiol. Ecol.* 59, 255–264. doi: 10.1111/j.1574-6941.2006.00198.x

Bachy, C., López-García, P., Vereshchaka, A., and Moreira, D. (2011). Diversity and vertical distribution of microbial eukaryotes in the snow, sea ice and seawater near the North Pole at the end of the polar night. *Front. Microbiol.* 2:106. doi: 10.3389/fmicb.2011.00106

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. doi: 10.1146/annurev-genet-110410-132430

Biswas, A., Gagnon, J. N., Brouns, S. J., Fineran, P. C., and Brown, C. M. (2013). CRISPRTarget: bioinformatic prediction and analysis of crRNA targets. *RNA Biol.* 10, 817–827. doi: 10.4161/rna.24046

Blank, C. E., Cady, S. L., and Pace, N. R. (2002). Microbial composition of near-boiling silica-depositing thermal springs throughout Yellowstone National Park. *Appl. Environ. Microbiol.* 68, 5123–5135. doi: 10.1128/AEM.68.10.5123-5135.2002

Brinkmeyer, R., Knittel, K., Jürgens, J., Weyland, H., Amann, R., and Helmke, E. (2003). Diversity and structure of bacterial communities in Arctic versus Antarctic pack ice. *Appl. Environ. Microbiol.* 69, 6610–6619. doi: 10.1128/AEM.69.11.6610-6619.2003

Brum, J. R., Hurwitz, B. L., Schofield, O., Ducklow, H. W., and Sullivan, M. B. (2015). Seasonal time bombs: dominant temperate viruses affect Southern Ocean microbial dynamics. *ISME J.* 10, 437–449. doi: 10.1038/ismej.2015.125

Cameron, K. A., Hagedorn, B., Dieser, M., Christner, B. C., Choquette, K., Sletten, R., et al. (2015). Diversity and potential sources of microbiota associated with snow on western portions of the Greenland Ice Sheet. *Environ. Microbiol.* 17, 594–609. doi: 10.1111/1462-2920.12446

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., et al. (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl. Acad. Sci. U.S.A.* 108, 4516–4522. doi: 10.1073/pnas.1000080107

Carpenter, E. J., Lin, S., and Capone, D. G. (2000). Bacterial activity in South Pole snow. *Appl. Environ. Microbiol.* 66, 4514–4517. doi: 10.1128/AEM.66.10.4514-4517.2000

Castillo, D., Christiansen, R. H., Dalsgaard, I., Madsen, L., and Middelboe, M. (2015). Bacteriophage resistance mechanisms in the fish pathogen *Flavobacterium psychrophilum*: linking genomic mutations to changes in bacterial virulence factors. *Appl. Environ. Microbiol.* 81, 1157–1167. doi: 10.1128/AEM.03699-14

Christner, B. C., Kvitko, B. H., and Reeve, J. N. (2003). Molecular identification of bacteria and eukarya inhabiting an Antarctic cryoconite hole. *Extremophiles* 7, 177–183. doi: 10.1007/s00792-002-0309-0

Chylinski, K., Makarova, K. S., Charpentier, E., and Koonin, E. V. (2014). Classification and evolution of type II CRISPR-Cas systems. *Nucleic Acids. Res.* 42, 6091–6105. doi: 10.1093/nar/gku241

Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., et al. (2014). Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids. Res.* 42, D633–D642. doi: 10.1093/nar/gkt1244

Cowan, D. A., and Tow, L. A. (2004). Endangered Antarctic environments. *Annu. Rev. Microbiol.* 58, 649–690. doi: 10.1146/annurev.micro.57.030502.090811

Davis, C. H., Li, Y., McConnell, J. R., Frey, M. M., and Hanna, E. (2005). Snowfall-driven growth in East Antarctic ice sheet mitigates recent sea-level rise. *Science* 308, 1898–1901. doi: 10.1126/science.1110662

Deveau, H., Barrangou, R., Garneau, J. E., Labonte, J., Fremaux, C., Boyaval, P., et al. (2008). Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. *J. Bacteriol.* 190, 1390. doi: 10.1128/JB.01412-07

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., and Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 27, 2194–2200. doi: 10.1093/bioinformatics/btr381

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. doi: 10.1111/j.1365-2958.2012.08171.x

Fischer, S., Maier, L. K., Stoll, B., Brendel, J., Fischer, E., Pfeiffer, F., et al. (2012). An archaeal immune system can detect multiple protospacer adjacent motifs (PAMs) to target invader DNA. *J. Biol. Chem.* 287, 33351–33363. doi: 10.1074/jbc.M112.377002

Fujii, M., Takano, Y., Kojima, H., Hoshino, T., Tanaka, R., and Fukui, M. (2010). Microbial community structure, pigment composition, and nitrogen source of red snow in Antarctica. *Microbiol. Ecol.* 59, 466–475. doi: 10.1007/s00248-009-9594-9

Gori, A., Bandera, A., Marchetti, G., Esposti, A. D., Catozzi, L., Nardi, G. P., et al. (2005). Spoligotyping and *Mycobacterium tuberculosis*. *Emerging Infect. Dis.* 11, 1242–1248. doi: 10.3201/eid1108.040982

Harding, T., Jungblut, A. D., Lovejoy, C., and Vincent, W. F. (2011). Microbes in high arctic snow and implications for the cold biosphere. *Appl. Environ. Microbiol.* 77, 3234–3243. doi: 10.1128/AEM.02611-10

Hell, K., Edwards, A., Zarsky, J., Podmirseg, S. M., Girdwood, S., Pachebat, J. A., et al. (2013). The dynamic bacterial communities of a melting High Arctic glacier snowpack. *ISME J.* 7, 1814–1826. doi: 10.1038/ismej.2013.51

Herlemann, D. P., Labrenz, M., Jürgens, K., Bertilsson, S., Waniek, J. J., and Andersson, A. F. (2011). Transitions in bacterial communities along the 2000 km salinity gradient of the Baltic Sea. *ISME J.* 5, 1571–1579. doi: 10.1038/ismej.2011.41

Hinkler, J., Hansen, B. U., Tamstorf, M. P., Sigsgaard, C., and Petersen, D. (2008). Snow and snow-cover in central Northeast Greenland. *Adv. Ecol. Res.* 40, 175–195. doi: 10.1016/S0065-2504(07)00008-6

Hultman, J., Waldrop, M. P., Mackelprang, R., David, M. M., McFarland, J., Blazewicz, S. J., et al. (2015). Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes. *Nat. Lett. Res.* 521, 208–212. doi: 10.1038/nature14238

Huson, D. H., Mitra, S., Ruscheweyh, H. J., Weber, N., and Schuster, S. C. (2011). Integrative analysis of environmental sequences using MEGAN 4. *Genome Res.* 21, 1552–1560. doi: 10.1101/gr.120618.111

Kemp, P. F., and Aller, J. Y. (2004a). Bacterial diversity in aquatic and other environments: what 16S rDNA libraries can tell us. *FEMS Microbiol. Ecol.* 47, 161–177. doi: 10.1016/S0168-6496(03)00257-5

Kemp, P. F., and Aller, J. Y. (2004b). Estimating prokaryotic diversity: when are 16S rDNA libraries large enough? *Limnol. Oceonol.* 2, 114–125. doi: 10.4319/lom.2004.2.114

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. doi: 10.1093/nar/gkt606

Liu, Y., Yao, T., Jiao, N., Kang, S., Xu, B., Zeng, Y., et al. (2009). Bacterial diversity in the snow over Tibetan Plateau Glaciers. *Extremophiles* 13, 411–423. doi: 10.1007/s00792-009-0227-5

Lopatina, A., Krylenkov, V., and Severinov, K. (2013). Activity and bacterial diversity of snow around Russian Antarctic stations. *Res. Microbiol.* 164, e1. doi: 10.1016/j.resmic.2013.08.005

Lopez-Sanchez, M. J., Sauvage, E., Da Cunha, V., Clermont, D., Ratsima Hariniaina, E., Gonzalez-Zorn, B., et al. (2012). The highly dynamic CRISPR1 system of *Streptococcus agalactiae* controls the diversity of its mobilome. *Mol. Microbiol.* 85, 1057–1071. doi: 10.1111/j.1365-2958.2012.08172.x

Maccario, L., Vogel, T. M., and Larose, C. (2014). Potential drivers of microbial community structure and function in Arctic spring snow. *Front. Microbiol.* 5:413. doi: 10.3389/fmicb.2014.00413

MacQueen, J. (1967). "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability in Statistics* (Berkeley, CA: University of California Press), 281–297.

Maier, L.-K., Lange, S. J., Stoll, B., Haas, K. A., Fischer, F., Fischer, E., et al. (2013). Essential requirements for the detection and degradation of invaders by the *Haloferax volcanii* CRISPR/Cas system I-B. *RNA Biol.* 10, 865–874. doi: 10.4161/rna.24282

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

Makarova, K. S., Haft, D. H., Barrangou, R., Brouns, S. J. J., Charpentier, E., Horvath, P., et al. (2011). Evolution and classification of the CRISPR–Cas systems. *Nat. Rev. Microbiol.* 9, 467–477. doi: 10.1038/nrmicro2577

Marraffini, L. A., and Sontheimer, E. J. (2010). CRISPR interference: RNS-directed adaptive immunity in bacteria and archaea. *Nat. Rev. Gen.* 11, 181–190. doi: 10.1038/nrg2749

McBride, M. J., Xie, G., Martens, E. C., Lapidus, A., Henrissat, B., Rhodes, R. G., et al. (2009). Novel features of the polysaccharide-digesting gliding bacterium Flavobacterium johnsoniae as revealed by genome sequence analysis. *Appl. Environ. Microbiol.* 75, 6864–6875. doi: 10.1128/AEM.01495-09

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., et al. (2008). The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *Bioinformatics* 9, 386. doi: 10.1186/1471-2105-9-386

Michaud, L., Giudice, A., Mysara, M., Monsieurs, P., Raffa, C., Leys, N., et al. (2014). Snow surface microbiome on the High Antarctic Plateau (DOME C). *PLoS ONE* 9:e104505. doi: 10.1371/journal.pone.0104505

Miteva, V. (2008). "Bacteria in snow and glacier ice," in *Psychrophiles: from Biodiversity to Biotechnology*, eds R. Margesin, F. Schinner, J.-C. Marx, C. Gerday (Berlin, Heidelberg, Germany: Springer Verlag), 31–50.

Møller, A. K., Søborg, D. A., Al-Soud, W. A., Sørensen, S. J., and Kroer, N. (2013). Bacterial community structure in High-Arctic snow and freshwater as revealed by pyrosequencing of 16S rRNA genes and cultivation. *Polar Res.* 32:17390. doi: 10.3402/polar.v32i0.17390

Patel, R. K., and Jain, M. (2012). NGS QC toolkit: a toolkit for quality control of next generation sequencing data. *PLoS ONE* 7:e30619. doi: 10.1371/journal.pone.0030619

Pomeroy, J. W., and Brun, E. (2001). "Physical properties of snow," in *Snow Ecology: An Interdisciplinary Examination of Snow-Covered Ecosystems,* eds H. Jones, J. W. Pomeroy, D. A. Walker, and R. W. Hoham (New York, NY: Cambridge University Press), 45–126.

Roux, S., Tournayre, J., Mahul, A., Debroas, D., and Enault, F. (2014). Metavir 2: new tools for viral metagenome comparison and assembled

virome analysis. *BMC Bioinformatics* 15:76. doi: 10.1186/1471-2105-15-76

Schloss, P. D., Gevers, D., and Westcott, S. L. (2011). Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS ONE* 6:e27310. doi: 10.1371/journal.pone.0027310

Segawa, T., Miyamoto, K., Ushida, K., Agata, K., Okada, N., and Kohshima, S. (2005). Seasonal change in bacterial flora and biomass in mountain snow from the Tateyama Mountains, Japan, analyzed by 16S rRNA gene sequencing and real-time PCR. *Appl. Environ. Microbiol.* 71, 123–130. doi: 10.1128/AEM.71.1.123-130.2005

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* 27, 379–423. doi: 10.1002/j.1538-7305.1948.tb01338.x

Shmakov, S., Savitskaya, E., Semenova, E., Logacheva, M. D., Datsenko, K. A., and Severinov, K. (2014). Pervasive generation of oppositely oriented spacers during CRISPR adaptation. *Nucleic Acids. Res.* 42, 5907–5916. doi: 10.1093/nar/gku226

Simpson, E. H. (1949). Measurement of diversity. *Nature* 163, 688. doi: 10.1038/163688a0

Skennerton, C. T., Imelfort, M., and Tyson, G. W. (2013). Crass: identification and reconstruction of CRISPR from unassembled metagenomic data. *Nucleic Acids. Res.* 41, e105. doi: 10.1093/nar/gkt183

Sul, W. J., Cole, J. R., Jesus, E. C., Wang, Q., Farris, R. J., Fish, J. A., et al. (2011). Bacterial community comparisons by taxonomy-supervised analysis independent of sequence alignment and clustering. *Proc. Natl. Acad. Sci. U.S.A.* 108, 14637–14642. doi: 10.1073/pnas.1111435108

Sun, C. L., Thomas, B. C., Barrangou, R., and Banfield, J. F. (2015). Metagenomic reconstructions of bacterial CRISPR loci constrain population histories. *ISME J.* 10, 858–870. doi: 10.1038/ismej.2015.162

Touchon, M., Barbier, P., Bernardet, J.-F., Loux, V., Vacherie, B., Barbe, V., et al. (2011). Complete genome sequence of the fish pathogen *Flavobacterium branchiophilum*. *Appl. Environ. Microbiol.* 77, 7656–7662. doi: 10.1128/AEM.05625-11

Touchon, M., and Rocha, E. P. C. (2010). The small, slow and specialized CRISPR and anti-CRISPR of *Escherichia* and *Salmonella*. *PLoS ONE* 5:e11126. doi: 10.1371/journal.pone.0011126

Varin, T., Lovejoy, C., Jungblut, A. D., Vincent, W. F., and Corbeil, J. (2012). Metagenomic analysis of stress genes in microbial mat communities from Antarctica and the High Arctic. *Appl. Environ. Microbiol.* 78, 549–559. doi: 10.1128/AEM.06354-11

Westra, E. R., Buckling, A., and Fineran, P. C. (2014). CRISPR–Cas systems: beyond adaptive immunity. *Nat. Rev. Microbiol.* 12, 317–326. doi: 10.1038/nrmicro3241

Westra, E. R., van Erp, P. B., Künne, T., Wong, S. P., Staals, R. H., Seegers, C. L., et al. (2012). CRISPR immunity relies on the consecutive binding and degradation of negatively supercoiled invader DNA by Cascade and Cas3. *Mol. Cell* 46, 595–605. doi: 10.1016/j.molcel.2012.03.018

Zhang, T. (2005). Influence of the seasonal snow cover on the ground thermal regime: an overview. *Rev. Geophys.* 43:RG4002. doi: 10.1029/2004RG000157

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Table S1. MG-RAST IDs, raw reads statistics, and diversity metrics of 16S Illumina reads

	MG-RAST ID	# of sequences after quality trimming	# of genus observed	Chao1 index	Shannon index	Simpson index	Coverage %
Druzhnaja	4616914.3	39483	179	213	2.66	0.87	93
Leningradskaja	4616915.3	61231	200	231	2.82	0.9	93
Mirnii	4616916.3	50135	181	224	2.57	0.83	30
Progress	4616917.3	46261	194	234	2.51	0.81	76

#of sequences in cloned library
117
126
ND
ND

Table S2. MG-RAST IDs, raw reads statistics of shotgun metagenomic sequences

Stations	MG-RAST ID	# of sequences	Totally, bp	# of rRNA genes	# of predicted proteins with known functions	# of predicted proteins with unknown function	no rRNA genes or predicted proteins
Druzhnaja	4624083.3	101,717	29,907,754	799	71,125	20,320	0
Leningradskaja	4624084.3	315,145	70,173,778	1,713	229,753	64,599	0
Mirnii	4624085.3	273,540	76,130,151	1,862	88,170	154,242	10,553
Progress	4624086.3	104,834	31,256,950	331	64,187	31,495	105

Table S5. Antarctic *Flavobacterium psychrophilum* CRISPR repeat types

Variants of sequences of flavobacterial repeats	# of reads with particular repeat sequence		
	Druzhnaja	Leningradskaja	Progress
GTTGGTAATTATAAGCTAAAATACAATTTTGAAAGCAATTCACAAC	149,900	172,353	152,926
GTTGGGAATTATAAGCTAAAATACAATTTTGAAAGCAATTCACAAC	87,968	88,890	75,732
GTTGGGAATTATAAGCTAAACTACAATTTTGAAAGCAATTCACAAC	666	1,130	484
GTTGGGAATTATGAGCTAAAATACAATTTTGAAAGCAATTCACAAC	357	223	270

CHAPTER III

Natural diversity of CRISPR spacers of Thermus: evidence of local spacer acquisition and global spacer exchange

Introduction:

In this Chapter the diversity of spacers associated with six CRISPR-Cas system types (I-A, I-B, I-C, I-E, I-U, III-A/III-B) of Thermus communities from five geographically distant hot springs was described. In addition, five new Thermus phages were isolated and sequenced from the samples. In comparison to previous chapters, the CRISPRome data analyzed in Chapter III is more complex. Comparisons between six CRISPR-Cas types, five sampling sites, several time points, and different identified local viral populations were performed.

Contribution:

I performed all the bioinformatics analysis (NGS data processing, clustering of spacer sequences, comparison of spacer diversity in different groups, and search of protospacers), prepared figures and tables, and drafted the manuscript.

PHILOSOPHICAL TRANSACTIONS B

royalsocietypublishing.org/journal/rstb

Research

Cite this article: Lopatina A, Medvedeva S, Artamonova D, Kolesnik M, Sitnik V, Ispolatov Y, Severinov K. 2019 Natural diversity of CRISPR spacers of *Thermus*: evidence of local spacer acquisition and global spacer exchange. *Phil. Trans. R. Soc. B* **374**: 20180092.
http://dx.doi.org/10.1098/rstb.2018.0092

Accepted: 20 January 2019

One contribution of 17 to a discussion meeting issue 'The ecology and evolution of prokaryotic CRISPR-Cas adaptive immune systems'.

Subject Areas:
ecology

Keywords:
CRISPR, *Thermus*, diversity of spacers, *Thermus* phages

Author for correspondence:
Konstantin Severinov
e-mail: severik@waksman.rutgers.edu

THE ROYAL SOCIETY PUBLISHING

Natural diversity of CRISPR spacers of *Thermus*: evidence of local spacer acquisition and global spacer exchange

Anna Lopatina[1,2,7,†], Sofia Medvedeva[3,4,†], Daria Artamonova[3], Matvey Kolesnik[3], Vasily Sitnik[3], Yaroslav Ispolatov[5] and Konstantin Severinov[1,3,6,7]

[1]Institute of Molecular Genetics and [2]Institute of Gene Biology, Russian Academy of Sciences, Moscow, Russia
[3]Skolkovo Institute of Science and Technology, Skolkovo, Russia
[4]Pasteur Institute, Paris, France
[5]Department of Physics, University of Santiago de Chile, Santiago, Chile
[6]Waksman Institute, Department of Molecular Biology and Biochemistry, Rutgers, The State University of New Jersey, Piscataway, NJ, USA
[7]Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, Israel

YI, 0000-0002-0201-3396; KS, 0000-0001-9706-450X

We investigated the diversity of CRISPR spacers of *Thermus* communities from two locations in Italy, two in Chile and one location in Russia. Among the five sampling sites, a total of more than 7200 unique spacers belonging to different CRISPR-Cas systems types and subtypes were identified. Most of these spacers are not found in CRISPR arrays of sequenced *Thermus* strains. Comparison of spacer sets revealed that samples within the same area (separated by few to hundreds of metres) have similar spacer sets, which appear to be largely stable at least over the course of several years. While at further distances (hundreds of kilometres and more) the similarity of spacer sets is decreased, there are still multiple common spacers in *Thermus* communities from different continents. The common spacers can be reconstructed in identical or similar CRISPR arrays, excluding their independent appearance and suggesting an extensive migration of thermophilic bacteria over long distances. Several new *Thermus* phages were isolated in the sampling sites. Mapping of spacers to bacteriophage sequences revealed examples of local acquisition of spacers from some phages and distinct patterns of targeting of phage genomes by different CRISPR-Cas systems.

This article is part of a discussion meeting issue 'The ecology and evolution of prokaryotic CRISPR-Cas adaptive immune systems'.

1. Introduction

Bacteriophages are the most abundant and ubiquitous biological entities on the planet [1,2]. Viruses of bacteria have profound influence on population and community structure and microbial evolution [3]. Being constantly under viral predation, bacteria have developed a broad range of mechanisms against phages such as CRISPR-Cas systems, restriction–modification systems, abortive infection systems as well as dozens of others, which are yet poorly investigated [4–6]. CRISPR-Cas systems comprise CRISPR DNA arrays with identical repeats and variable spacers, and CRISPR-associated (*cas*) genes [7]. At one end of the CRISPR array, a leader sequence containing a promoter from which the array is transcribed is located [8]. New spacers can be acquired from the genomes of viruses or plasmids. The spacer is acquired at the leader-proximal end of the array and the acquisition of a spacer also leads to

[†]These authors contributed equally to the study.

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.4400771.

the appearance of an additional copy of the CRISPR repeat. Thus, spacers that are located distal to the leader have been acquired earlier than leader-proximal spacers. The CRISPR array is transcribed and the resulting precursor RNA is processed into individual CRISPR RNAs (crRNAs) each containing a spacer sequence and fragments of flanking repeats [8,9]. Individual crRNAs are bound by Cas effector proteins and can recognize nucleic acids complementary to the crRNA spacer. Upon recognition, foreign nucleic acids are destroyed. In DNA targeting CRISPR-Cas systems, spacers in the CRISPR array are not recognized as, in addition to complementarity with the crRNA spacer, the target must also have a protospacer adjacent motif (PAM) recognized by the effector. Since the part of the CRISPR repeat that is located in the place of PAM is not recognized, discrimination of self from non-self becomes possible. Currently, CRISPR-Cas systems are divided into two classes, six types and 33 subtypes that differ in Cas effector components, details of target recognition, target destruction and self versus non-self discrimination [10].

Analysis of CRISPR spacers is a valuable source of information about virus–host interactions, because short DNA fragments of previously encountered viruses are 'recorded' in CRISPR arrays as spacers, and cells carrying protective spacers are expected to gain an advantage and become more numerous. Such analysis can be particularly powerful when applied to metagenomic data. Besides extraction from metagenomic data or CRISPR loci [11,12], CRISPR spacers can be directly amplified and analysed either from individual bacterial isolates or from whole communities [13–15].

Comparison of CRISPR arrays from isolated populations of the same species revealed great diversity of spacer sequences, which is increased towards the leader-proximal end of arrays [12,16–18]. Analysis of changes of spacer content over time provided examples of new spacers acquisition to the leader-proximal ends of CRISPR arrays, deletion of old spacers from leader-distal ends and recombination of CRISPR arrays between different strains [14,19–21].

CRISPR spacers can be used to identify viral sequences in metagenomes and monitor changes in viral populations [11,22,23]. Examples of spacers that preferably target local phages from the same sampling site were reported [19,24–26]. Theoretical models of coevolution of viruses and hosts demonstrated the efficiency of CRISPR-Cas defence when viral density is small [27]. Host and virus populations were predicted to oscillate short term, with a few dominant strains existing at every given time point [28]. The presence of multiple spacers against a viral genome in host strains makes it more difficult for virus to escape by acquiring mutations in the targeted sites. This may help to maintain spacer diversity over longer time scales [29].

In this work, we investigated the diversity of CRISPR spacers of uncultured communities of *Thermus* strains from distant hot springs and compared them with each other and with a *Thermus* CRISPR database. We also compared *Thermus* bacteriophages and spacers obtained from the same locations. Our analysis reveals, on the one hand, evidence of CRISPR spacer acquisition by *Thermus* communities from local phages and, on the other hand, global distribution of many spacers and arrays suggesting intercontinental migration of at least some *Thermus* strains between their unique ecological niches.

2. Material and methods

(a) Sample collection

The samples were collected from hot gravel of Mount Vesuvius (October 2014, October 2018) or hot springs at Mount Etna (October 2012), the el Tatio region of northern Chile (October 2014), and the Termas del Flaco region of southern Chile (December 2013 and March 2016) and Uzon caldera in Kamchatka, Russia (August 2018). During collection, samples of gravel were taken 5–100 m from each other and water samples were collected from separate hot springs located within a similar distance. In the case of Termas del Flaco, the same hot springs were sampled in 2013 (two samples) and 2016 (three samples). The samples were stored at 4°C and brought to the laboratory within one to two weeks after collection for analysis. Preliminary experiments with laboratory *Thermus thermophilus* strains HB8 and HB27 revealed no loss of viability during conditions and times of storage used. Vesuvius 2018 samples were analysed 2 days after collection.

(b) Enrichment cultures

Five millilitres of TB medium [0.8% (w/v) tryptone, 0.4% (w/v) yeast extract, 0.3% (w/v) NaCl, 0.5 mM $MgCl_2$ and 0.5 mM $CaCl_2$] were inoculated with a 100 µl aliquot of hot spring water sample and incubated overnight at 70°C with vigorous agitation. Enrichment cultures were checked for the presence of *Thermus* by PCR with oligonucleotide primers specific for *Thermus* 16S rRNA gene (electronic supplementary material, table S1). Amplifications were carried out with Taq DNA polymerase under the following conditions: initial denaturation for 5 min at 95°C, followed by 28 cycles of 30 s at 95°C, 30 s at 55°C and 40 s at 72°C, and a final extension at 72°C for an additional 2 min.

(c) Phage isolation

Thermus thermophilus strains HB8 ATCC 27634 and HB27 ATCC BAA-163 were used in enrichment cultures to isolate bacteriophages from environmental samples. Five millilitres of TB medium were inoculated with a 100 µl aliquot of overnight culture of one of the *Thermus* strains and growth proceeded until OD600 reached approximately 0.4. An amount of 0.2–0.5 ml of environmental sample was added and incubation was continued overnight at 70°C with vigorous agitation. To isolate individual phage plaques, 1 ml of enrichment culture was centrifuged for 15 min, and 100 µl aliquots of supernatant were combined with 150 µl of freshly grown *T. thermophilus* HB8 or HB27 cultures (OD600 approx. 0.4). Melted soft (0.75%) TB agar was added, mixtures were poured over 2.5% TB agar plates and incubated overnight at 65°C. Individual plaques were picked with toothpicks and cleaned by several passages on the host *Thermus* strain as described above.

(d) Phage DNA extraction and sequencing

Phage lysates were prepared and DNA was extracted as described previously [30]. Five hundred nanograms of phage DNA were used for library preparation and pair-end sequencing was carried out on the Illumina MiSeq platform with MiSeq reagent kit v. 2 (Illumina, USA) as described previously [31].

(e) Phage genome annotation

Phage genomes were automatically annotated using GeneMark [32] and annotation was further manually checked by the Artemis program [33] and verified by Blastp and HHpred programs. The BlastN tool was used to compare the genomes of newly isolated phages with the database.

(f) Bacterial DNA extraction, amplification and sequencing

DNA was extracted from 2 ml of spring water, mud samples from gravel or enrichment cultures using Blood and Tissue kit (Qiagen) according to the manufacturer's protocol for Gram-negative cells. Different sets of oligonucleotide primers were used to amplify CRISPR spacers (electronic supplementary material, table S1). Amplification was carried out with Taq DNA polymerase under the following conditions: initial denaturation for 5 min at 95°C, followed by 28 cycles of 30 s at 95°C, 30 s at 50–60°C and 40 s at 72°C, and a final extension at 72°C for an additional 2 min. Two nanograms of total DNA were used as a template for each PCR reaction. To avoid biases during PCR amplification, 10 replicates of each PCR reaction were performed for every sample and mixed before further manipulations. Amplicons were visualized on 1% ethidium bromide-stained agarose gel and DNA fragments of 200–1000 bp in length were purified from the gel and sequenced on the Illumina MiSeq platform as described above.

(g) Spacer clustering and analysis

Raw reads were demultiplexed, trimmed by quality with Phred score greater than or equal to 20 and no admission of ambiguous bases using CLC Genomics 8.0 workbench software (CLC Bio Aarhus, Denmark). Spacers were extracted using spget (https://github.com/zzaheridor/spget). To decrease the number of spacers and to avoid overrepresented diversity because of mistakes during PCR and sequencing, spacers were clustered using UCLUST algorithm [34]. The maximum number of substitutions allowed for spacers within one cluster corresponds to 85% identity over the full length of the spacer; end gaps were allowed with zero penalties. Chao index, α and β diversities were calculated with vegan package for R [35]. Good's criterion is defined as $1-(n1/N)$, where N is a total number of spacers in the sample, and $n1$ is a number of singleton spacers.

Centres of spacer clusters (the most highly represented sequence within a cluster) were compared against the NCBI nucleotide collection (nt) and a local database of *Thermus* phages and plasmids with the BLASTn algorithm with parameters for short input sequences (word size 8). Sequences with more than 85% of identity over the entire spacer length and without indels were considered as positive hits.

PAM identification was performed using the CRISPRTarget online tool [36]. Eight nucleotides upstream and downstream of each protospacer were extracted and used for PAM logo search with the Weblogo online tool (http://weblogo.berkeley.edu/logo.cgi). Repeats sequences from identified CRISPR arrays were classified using the CRISPRmap tool [37].

(h) Data access

Phages sequences of phiFa, phiKo, phiLo, phiMa and YS40-Isch were deposited in GenBank under accession numbers MH673671, MH673672, MH673673, MH673674 and MK257744, respectively. Sequences of CRISPR spacers from natural *Thermus* communities are available in the electronic supplementary material.

3. Results

(a) The diversity of CRISPR spacers in complete *Thermus* genomes

Fully sequenced genomes of 26 *Thermus* strains isolated around the world were available in GenBank at the time of writing (electronic supplementary material, table S2). *Thermus* genomes usually contain multiple CRISPR-Cas systems of different types [38] located on the chromosome and/or on megaplasmids present in some isolates. Most *Thermus cas* operons have an adjacent CRISPR array with a specific repeat sequence. Because of a clear connection between the type of *Thermus cas* operon and repeat sequence of adjacent CRISPR array [39], each array (and repeat) can be assigned to a specific CRISPR type or subtype. The III-A and III-B subtype *cas* gene operons have adjacent CRISPR arrays with identical repeat sequences. Moreover, it has been shown that effectors of both subtypes III-A and III-B bind to common crRNAs [40,41]. Therefore, the III-A and III-B subtypes cannot be distinguished from each other and are treated here as a single type, type IIIAB.

For further analysis, we considered six dominant *Thermus* CRISPR-Cas systems: I-A, I-B, I-C, I-E, I-U and IIIAB. Consensus repeat sequences for each system used in our analysis are listed in table 1. We used the *spget* program to extract spacers associated with each consensus repeat sequence from fully sequenced *Thermus* genomes and analysed their diversity. Spacers from different *Thermus* isolates were considered identical if they had fewer than two mismatches in their sequences. In this way, a set of 1567 unique *Thermus* spacers was obtained. Most spacers were found to be strain-specific. For very closely related *T. thermophilus* strains isolated in Japan (labelled as 22, 23 and 24 in figure 1), 19 out of 269 spacers were identical and located one after another in CRISPR arrays of the same type. In *T. scotoductus* strains (labelled as 13–16, figure 1), the oldest, leader-distal spacer in one of the I-E CRISPR arrays was shared [42]. Finally, seven pairs of shared spacers must have been independently acquired from the same locus as they were found in CRISPR arrays belonging to CRISPR-Cas systems of different types and/or were partially overlapping. Similar instances of independent spacer acquisition were reported earlier in other microbes [12].

In total, only 31 *Thermus* spacers (2.0%) were found in more than one genome (see electronic supplementary material, table S3). For comparison, in a well-studied I-E CRISPR-Cas system of *Escherichia coli*, 90.9% of spacers were shared between at least two isolates (data not shown). These observations imply that the diversity of *Thermus* CRISPR spacers in current databases is very undersampled. BlastN analysis of 1567 unique spacers revealed, respectively, 52 (3.3%) and 80 (5.1%) spacers matching *Thermus* phages and prophages, 14 (0.9%) matches to plasmids and 48 (3.1%) matches to *Thermus* chromosomes in locations other than CRISPR arrays. Most spacers that matched DNA fragments from *Thermus* phages were from I-E and I-B arrays (21 and 20, respectively), suggesting that I-E and I-B systems are most active against known *Thermus* phages.

(b) Amplification of CRISPR spacers from natural *Thermus* communities

Given that spacer diversity in known *Thermus* genomes is underestimated, we decided to investigate spacer diversity in natural *Thermus* communities by amplifying spacers associated with specific repeats from samples collected from Mount Vesuvius hot gravel, and hot springs at Mount Etna, the el Tatio region in northern Chile, the Termas del Flaco region in southern Chile and Uzon caldera in Kamchatka, Russian Far East. At each collection site, the temperature was within

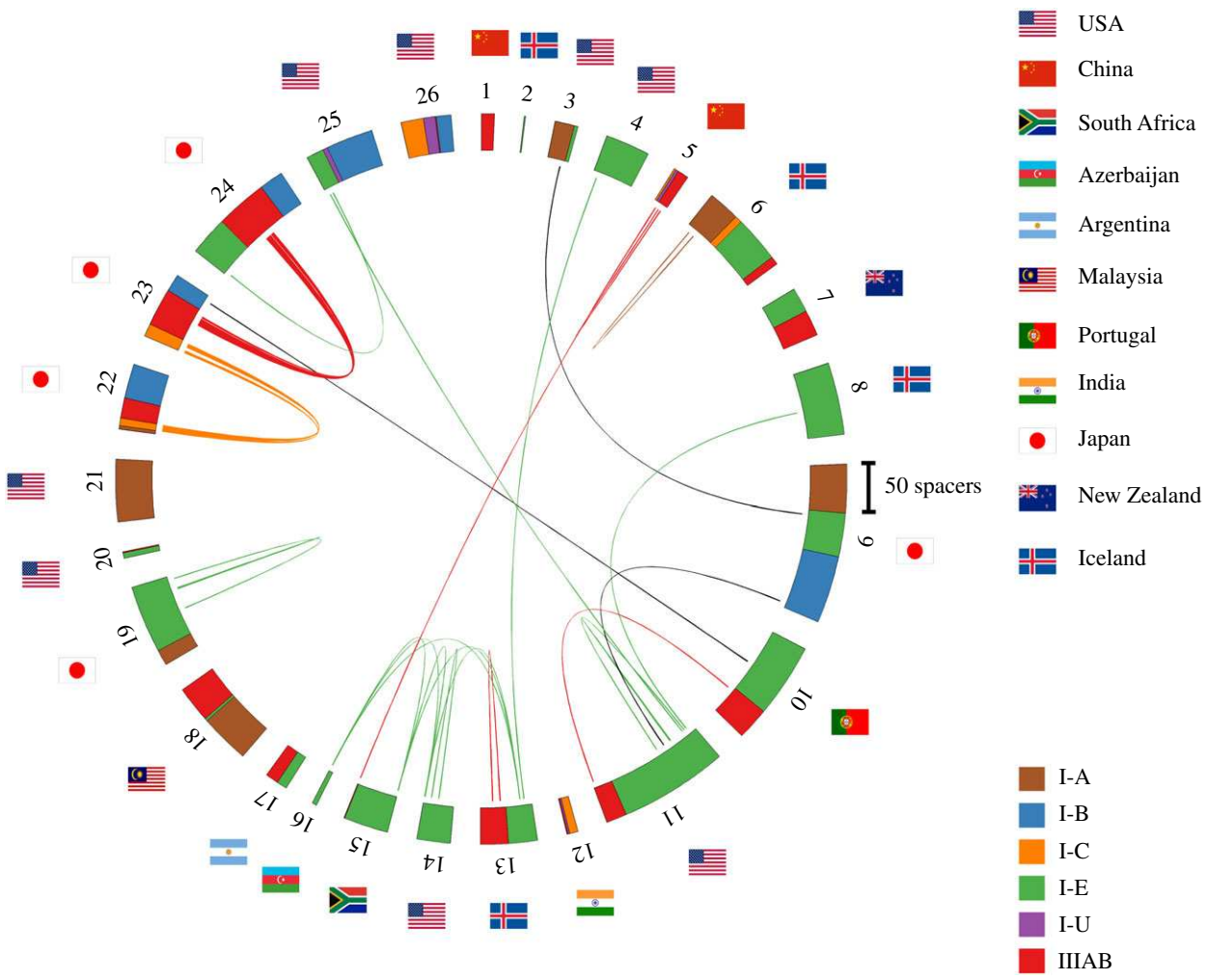


Figure 1. The diversity of CRISPR spacers in fully sequenced *Thermus* genomes. A total of 1567 spacers present in 26 fully sequenced *Thermus* sp. genomes are shown on a circular diagram. *Thermus* isolates used for analysis are numbered outside the spacer diagram (a full list of isolates can be found in electronic supplementary material, table S2). Spacers belonging to arrays of the same CRISPR-Cas systems types/subtype are indicated by identical colours. Spacers that differ from each other by fewer than two nucleotides are connected by lines whose colours correspond to colours indicating CRISPR-Cas systems types/subtypes. Spacers shared by arrays of different types/subtypes are connected by black lines. National flags indicate countries where each strain was isolated.

Table 1. Types of CRISPR repeats present in *Thermus* sp. CRISPR arrays. Consensus sequences built using repeat sequences present in CRISPR arrays of fully sequenced *Thermus* genomes listed in figure 1 are shown.

N	type of CRISPR-Cas	repeat sequence	average length of spacer
1	III	GTTGCAMRRGWKKSWKCCCCGYMAGGGGATKRHYDC	41
2	I-E	GTAGTCCCCACRCRYGTGGGGATGGMCSD	32
3	I-C	GTTGCACCGGCCCGAAAGGGCCGGTGAGGATTGAAAC	38
4	I-B	GTTGCAAACCYCGTYAGCCTCGTAGAGGATTGAAAC	36
5	I-U	GTTGCATCCAAGCTTCACAGCTTGGCTACGTTGCAGG	36
6	I-A	GTTTCAAACCCTYATAGGTACGGTYMRAAG	36

65–70°C and the pH was neutral, so we expected to find *Thermus* there. Degenerate partially self-complementary primers corresponding to each of the six *Thermus* CRISPR repeat consensus sequences (electronic supplementary material, table S1) were used for PCR amplification. The procedure (electronic supplementary material, figure S1A) was previously used to characterize spacer diversity in various environmental samples [31,43]. As the procedure was not previously applied to *Thermus* CRISPR arrays, primer pairs for I-B, I-E and IIIAB repeats were validated using DNA purified from *T. thermophilus* HB8 strain which harbours the corresponding CRISPR-Cas systems. With each primer pair, a characteristic ladder of amplification products was observed (an example of PCR fragments obtained with primers specific for the I-E type repeat is shown in electronic supplementary material, figure S1B, lane 3). We did not observe amplification products when DNA prepared directly from environmental samples was used as a template for PCR (as an example, see electronic supplementary material, figure S1B, lanes 4 and 5), probably because of the low concentration of *Thermus* cells. However,

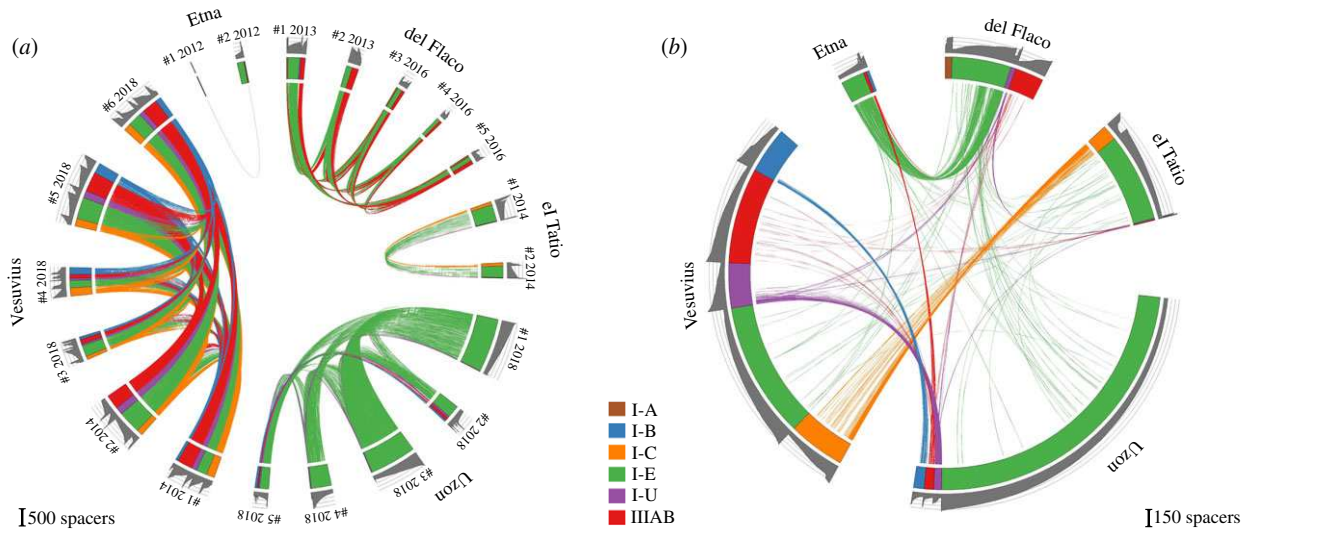


Figure 2. The diversity of CRISPR spacers in environmental *Thermus* samples. (*a*) The diversity of 14 872 spacers (spacer cluster centres) associated with *Thermus* CRISPR repeats from enrichment cultures obtained from samples collected at indicated sites is shown in the circular diagram. Spacers from the same location that differ from each other by fewer than two nucleotides are connected by matching colour lines. For del Flaco, samples #1 and #2 were collected in December 2013 and samples #3–#5 in March 2016. For Vesuvius, samples #1 and #2 were collected in October 2014 and samples #3–#6 in October 2018. (*b*) Spacers from the same location are merged. The resulting diversity of unique 7877 spacers is shown in the circular diagram. Spacers from different locations that differ from each other by fewer than two nucleotides are connected by matching colour lines. The colour labelling scheme is the same as in figure 1. Grey colour histograms on the outside show cluster size in $\log_{10}$ scale.

robust amplification products were seen with DNA prepared from enrichment cultures grown overnight at 70°C in rich medium (see Material and methods). The observed amplification patterns were reproducibly distinct for enrichment cultures seeded with material from different locations (as an example, see electronic supplementary material, figure S1B, lanes 6 and 7).

(c) The diversity of CRISPR spacers in *Thermus* communities

For each site, amplified material corresponding to spacers from different arrays was combined and subjected to Illumina sequencing. Using a spacer extraction pipeline similar to the one described earlier [31,43], a total of approximately 17.8 million spacers (defined as sequences of an expected length located between two repeats sequences of the same type) were extracted. Spacers with identity of more than 85% over their entire length and belonging to the CRISPR arrays with repeats of the same type were clustered, separately for each sample. The most abundant sequence in a cluster was considered as the cluster centre. Overall, implementation of the procedure described above resulted in 109 843 clusters. We measured α-diversity (Shannon entropy) for each sample and calculated the coverage of spacer diversity based on the number of lowly abundant clusters (see electronic supplementary material, table S4). The lowest coverage was observed for Vesuvius samples 3–6 (27–31%) and Uzon samples 3–5 (25–35%). Given undersequencing of spacers with low abundance, further analysis was performed for clusters that contained more than 10 spacers (14 872 clusters). For simplicity, below we will refer to cluster centres as 'spacers'.

When spacers from different sites were compared, 7246 unique spacers were identified. The collection of *Thermus* spacers obtained from environmental samples exceeds the number of spacers from sequenced isolates by more than fourfold (7246 compared with 1567). Yet, only 1.2% of spacers from natural *Thermus* communities are similar to database spacers. This value becomes even smaller if minor, less abundant spacers revealed by our analysis, are considered. The result emphasizes the extent of diversity of CRISPR spacers in *Thermus* and, presumably, reflects the high level of activity of *Thermus* CRISPR-Cas systems in spacer acquisition.

As the overall number of unique spacers (7246) is considerably less than the sum of spacers present in each site (14 872), it follows that some spacers are present in more than one sample. Spacers shared between samples collected from the same locality/reservoir are shown in figure 2*a*. The number of shared spacers ranges from 1.0% in Etna (because of the low number of spacers in Etna 1 sample) to 66% between Vesuvius 1 and Vesuvius 2 samples. Samples from Termas del Flaco, which were taken 27 months apart, illustrate the temporal stability of spacer content in time, with 37–49% of spacers shared between all samples. Even more dramatically, 36–63% of spacers collected 4 years apart at Vesuvius were also common. Interestingly, the frequencies of occurrence of common spacers (as evidenced by the size of the clusters that contain them) were comparable in samples collected at the same site (0.55–0.98 Pearson's coefficient). It can be argued that the enrichment procedure used to prepare cultures suitable for spacer amplification could have introduced a bias in observed spacer content. The stability of spacer sets in samples collected at the same location but separated by extended periods of time makes this possibility unlikely.

By identifying overlapping spacer pairs and triplets in longer Illumina reads, we reconstructed fragments of *Thermus* arrays containing 10–35 spacers (electronic supplementary material, table S5). As an example, three shared I-A CRISPR array fragments from Thermas del Flaco 1 and Thermas del Flaco 5 samples are shown in electronic supplementary material, figure S2. One array remained unchanged over the course of 27 months, another lost three spacers from the leader-distal end and the third was completely renewed except for one pair

of spacers. Overall, these observations are consistent with the existence of stable local *Thermus* communities sharing a conserved set of CRISPR spacers but also show evidence of temporal changes due to spacer acquisition and loss.

Analysis of spacers shared between remote sites is shown in figure 2*b*. For simplicity, all spacers present at the same location were combined to create this figure. Electronic supplementary material, figure S3 shows the results when individual samples from the same locations are treated separately. As can be seen, many spacers are shared between different locations. Four hundred and five spacers were shared between two sites, 78 between three sites and four spacers were shared between four sites. Our analysis revealed, rather strikingly, little overlap between spacer sets present in distant localities at the same continent compared with intercontinental spacer sets. For example, there are less common spacers between the El Tatio and Termas del Flaco sets than between the El Tatio and Vesuvius spacers ($p < 10^{-5}$, Fisher's exact test). The same result was obtained from hierarchical clustering of samples by pairwise β-diversity (electronic supplementary material, figure S4). The number of shared spacers also did not correlate with geographical distance (electronic supplementary material, figure S5). At present, we are unable to explain this observation. It is possible that certain physicochemical properties of water that were not recorded during sample collection are responsible. Careful control of ecological parameters of habitat at the collection sites and extension of analysis presented here to other *Thermus* communities around the world may help resolve this issue.

It could be argued that some spacers were acquired independently in different sites. Several identical partially reconstructed arrays were found in different sites. As chances of independent acquisition of identical spacers in the same order are negligible, the results show that some CRISPR arrays (and, presumably, strains that contain them) are shared between distant locations. Shared arrays contained sample-specific spacers, which were acquired at the leader-proximal end of the array (see electronic supplementary material, figure S6 for several examples). The result appears to mirror the situation with another thermophile, an archaeon *Sulfolobus*. In the full genome sequence of *Sulfolobus solfataricus* 98/2 isolated in Italy, 107 out of 189 CRISPR spacers are identical to spacers from *S. solfataricus* P2 isolated in the Yellowstone National Park [44]. Similarly, we found that *S. acidocaldarius* N8 from thermal fields in Japan and *S. acidocaldarius* GG12-C01-09 from Yellowstone share 95% of CRISPR spacers (data not shown).

(d) The provenance of *Thermus* spacers

We next examined sequences of spacers obtained in this work. Most spacers (94.5%) had no matches to the Genbank nucleotide collection, a situation that is typical for all CRISPR-Cas systems [39]. The remaining spacers matched *Thermus* phages (3.3%), small plasmids (0.4%) or non-CRISPR chromosome/megaplasmid sequences of *Thermus* (1.8%). Alignments of protospacers (sequences matching spacers) and their flanking sequences revealed a putative AAG protospacer-associated motif (PAM) on the 5′-protospacer flank for the I-E system, a GGTN PAM for the I-B system and a TTC PAM for the I-C and I-U systems (electronic supplementary material, figure S3). The AAG PAM has also been reported for the *E. coli* I-E system [43].

More than 100 *Thermus* bacteriophages have been isolated [45–49]. However, only eight complete genomes of *Thermus* phages are available in the Genbank database: IN93, p23-77, YS40, TMA, P23-45, P74-26, phiOH3 and phiOH16. Myoviruses YS40 and TMA, inoviruses phiOH3 and phiOH16, and siphoviruses P23-45 and P74-26 have closely related sequences, respectively. In the course of this work, we have isolated, sequenced and annotated five additional *Thermus* bacteriophages from samples that were used for amplification of spacers. Three phages, phiFa, phiKo and YS40-Isch were isolated from Mount Vesuvius samples, and two (phiLo and phiMa) from el Tatio samples. PhiFa is a siphovirus and most of its genes are homologous to long-tailed phages P23–45 and P74–26 isolated earlier in Kamchatka [50]. PhiKo (11 129 bp, 26 ORFs) belongs to *Tectiviridae* phage family. One PhiKo gene product is homologous to the lysozyme of *Thermus* phage 2119, and three others are homologous to proteins encoded by prophage region of *Thermus* sp. 2.9 isolate. PhiLo (178 531 bp, 165 ORFs) and phiMa (51843 bp, 66 ORFs) are myoviruses. Approximately 10% of phiLo proteins are homologous to proteins encoded by other *Thermus* phages (including YS40, TMA, IN93, P74-26), while 60% of phiMa proteins are most similar to proteins encoded by prophage region of *Thermus* sp. 2.9. YS40-Isch is highly similar to YS40 (87% DNA identity, 85% coverage by BLASTn) and TMA (86% identity, 84% coverage) phages earlier isolated in Japan [46,47].

When the five new *Thermus* phage genomes were taken into account, the percentage of matches of unique spacers with phage sequences increased from 3.3 to 6.3%, indicating that the diversity of *Thermus* phages is greatly undersampled. The results of spacer mapping to known *Thermus* phage genomes are summarized in table 2. The overwhelming majority of spacers that matched phiMa and phiKo genomes came from spacer sets from the same localities ($p < 10^{-15}$, Fisher's exact test). Only 17 spacers targeted YS40 isolated from Japan, while 33 spacers from Vesuvius matched YS40_Isch, a local phage. Spacers targeting IN93 were present in spacer sets from every sample. On the basis of the abundance of IN93 targeting spacers in different locations, it appears that, unlike the apparently 'local' phages such as phiKo and phiMa, the IN93-like phages are globally spread, possibly because of their ability to lysogenize their hosts.

It is apparent that different phages are targeted with widely different frequencies by spacers in our collection. For example, IN93, a small phage with an approximately 20 kb genome, contains 189 protospacers (constituting 38% of the total of phage sequences matching *Thermus* spacers), while some much larger phages, namely YS40-Isch or phiFa, are each targeted by about 30 spacers. It is also apparent that different phages are preferentially targeted by different CRISPR-Cas systems (table 3). Thus, most IN93 targeting spacers belong to the I-E subtype, while phiFa and YS40-Isch are preferentially targeted by IIIAB systems. Interestingly, the I-E system, which contains most unique spacers, has a relatively small percentage of spacers that match phage genomes (4%). This value is significantly higher for spacers of I-C (11%), I-U (9%) and I-B (17%) types.

The locations of protospacers in phage genomes that had most matches with *Thermus* spacers—IN93, phiKo and phiFa—are shown in figure 3 and, for phiMa, a phage with a large genome, in electronic supplementary material, figure S2. While the IN93 phage is globally distributed, for

Table 2. BlastN hits of spacers from different sites. The number of BlastN hits for 'not-unique spacers', i.e. identical spacers found in different sites, is shown. Only hits with greater than 85% identity over entire spacer length are included. Fisher's exact test was used to test for each virus that the number of protospacers depends on the sample site. The resulting *p*-values are given in the last row.

	IN93	P23-77	phiLo	phiMa	phiKo	phiFa	YS40_Isch	ϕOH3		
	Japan	New Zealand	El Tatio	El Tatio	Vesuvius	Vesuvius	Vesuvius	Japan	spacers with hits	total *N* of spacers
Vesuvius	111	2	12	3	107	23	30	0	288 (9.4%)	3123
Etna	11	0	1	0	2	0	0	0	14 (6.2%)	243
El Tatio	26	4	7	28	1	0	1	11	78 (9.0%)	869
Del Flaco	28	0	0	5	0	0	1	0	34 (4.1%)	822
Uzon	13	2	0	8	14	20	1	17	75 (2.7%)	2820
total *N* of hits	189	8	20	44	124	43	33	28	489 (6.3%)	7877
p-value	5.5×10^{-11}	7.7×10^{-2}	3.1×10^{-2}	2.2×10^{-16}	3.7×10^{-15}	9.6×10^{-8}	4.8×10^{-3}	4.6×10^{-14}		

Table 3. BlastN hits of spacers from different CRISPR-Cas systems. The number of hits for 'not-unique spacers', i.e. identical spacers belonging to different CRISPR-Cas system types, is shown. Only hits with greater than 85% identity over entire spacer length are included. The predicted PAMs for each system are presented in the second column (PAM logos are shown in electronic supplementary material, figure S3). Fisher's exact test was used to test for each virus that the number of protospacers depends on the type of CRISPR-Cas system. The resulting *p*-values are given in the last row.

		IN93	P23-77	phiLo	phiMa	phiKo	phiFa	YS40_Isch	ϕOH3		
	PAM	Japan	New Zealand	Chile	Chile	Vesuvius	Vesuvius	Vesuvius	Japan	spacers with hits	total *N* of spacers
I-E	AAG	100	6	10	43	18	0	0	28	205 (4.0%)	5114
I-C	TTC	44	1	5	1	13	0	4	0	68 (11.4%)	596
IIIAB	—	3	0	1	0	18	43	27	0	92 (8.4%)	1134
I-U	TTC	31	1	3	0	7	0	0	0	42 (8.7%)	482
I-B	GGTN	11	0	1	0	68	0	2	0	82 (16.6%)	494
I-A	?	0	0	0	0	0	0	0	0	0	57
total *N* of hits		189	8	20	44	124	43	33	28	489 (6.3%)	7877
p-value		1.0×10^{-30}	2.7×10^{-1}	9.0×10^{-3}	1.2×10^{-8}	3.2×10^{-30}	9.7×10^{-33}	2.5×10^{-18}	2.1×10^{-5}		

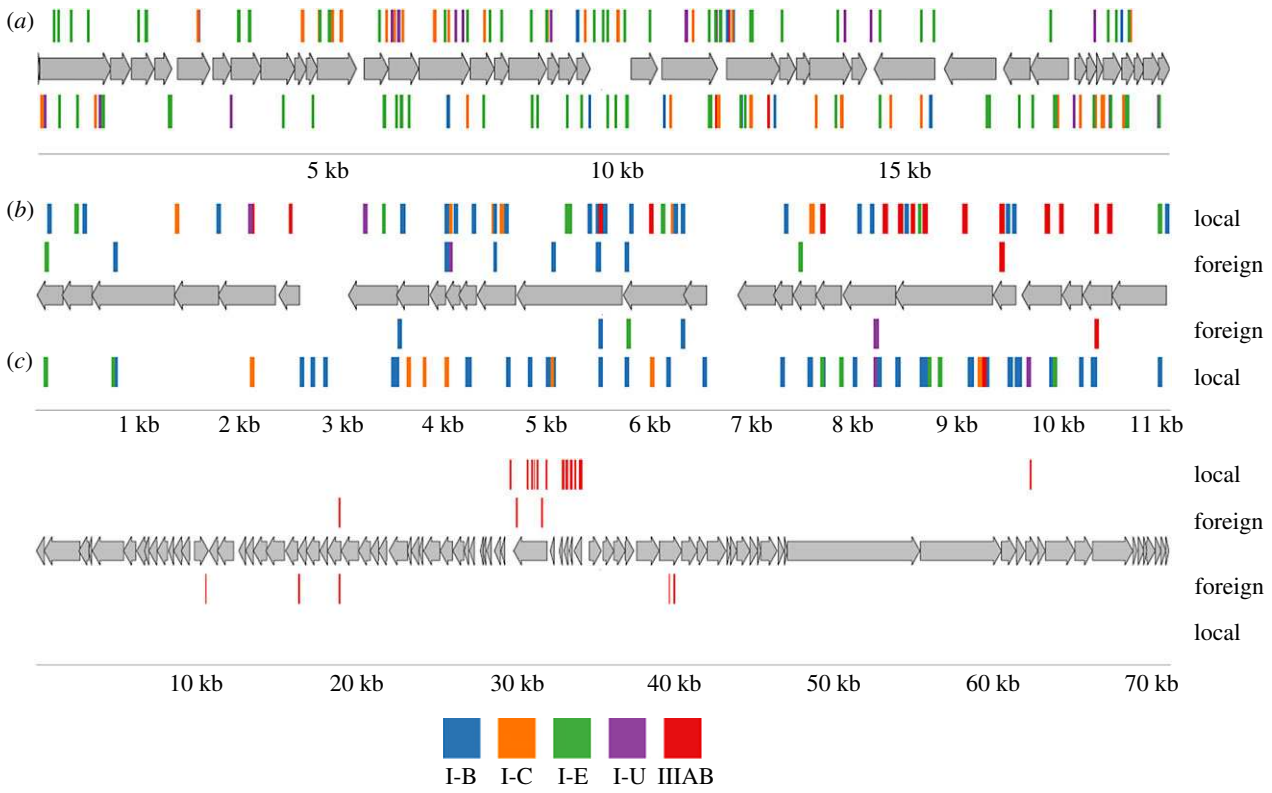


Figure 3. Mapping of protospacers in the genomes of *Thermus* phages. The double-stranded DNA genomes of *Thermus* bacteriophages IN93 (*a*), phiKo (*b*) and phiFa (*c*) are schematically shown. Numbers below indicate genome coordinates, in kilobases. Phage genes are indicated by grey arrows, with arrow directions matching the direction of transcription. Protospacers matching spacers associated with *Thermus* CRISPR repeats are shown as vertical lines above and below phage genomes. The colour of lines representing protospacers indicates the type of CRISPR-Cas systems to which the matching spacers belong (the colour scheme legend is shown at the bottom of the figure). For phiKo and phiFa, mapped spacers are separated into 'local', i.e. found at the site of phage isolation, and 'foreign', i.e. found at distant sites.

phiKo and phiFa phages we performed separate mapping of 'local' spacers recovered at the isolation site and 'foreign' spacers observed elsewhere. As can be seen from figure 3*b*,*c* and table 2, most spacers matching phiKo and phiFa are local.

The I-E and I-B spacers mapped evenly throughout phage genomes to both DNA strands (figure 3). By contrast, most protospacers matching IIIAB spacers were located on the transcribed strand of viral genes, an expected result given that interference by type III systems is transcription-coupled [51]. The observed location of type III protospacers suggests that phages do exert pressure on *Thermus* communities, for in the absence of such pressure non-functional type III spacers targeting the non-transcribed strand of phage DNA could have been expected. The distribution of type IIIAB protospacers along the genome was also highly uneven in the PhiKo (figure 3*b*) and, most prominently, in the PhiFa genomes (figure 3*c*). In the latter case, protospacers were located in a narrow central region of the genome, where, based on homology to *Thermus* P23-45 phage, the early genes are located. In the case of phiKo, type IIIAB protospacers mapped to the part of the genome where transcription of viral genes likely initiates. The result may indicate that spacers acquired from other regions of phage genomes do not provide bacteria that acquire them protection from the virus and are thus not retained in the population [52]. Alternatively, there may be specific aspects of phage development strategy that limit the adaptation machinery of the host to these regions. The availability of new phages described in this work will allow us to address these questions experimentally.

Data accessibility. Sequences of CRISPR spacers from natural *Thermus* communities are available in Supplementary material.

Authors' contributions. A.L. collected samples in Vesuvius and Enta, performed PCR amplification of spacers, and isolated and sequences four phages; S.M. conducted bioinformatics analyses and prepared figures and tables; D.A. performed PCR amplification of spacers, M.K. performed PCR amplification of spacers and isolated phiYS40-Isch phage, V.S. implemented the spget program for spacer extraction, Y.I. collected samples in Chile, K.S. supervised the project and drafted the manuscript.

Competing interests. We declare we have no competing interests.

Funding. This study was supported by the Russian Science Foundation (grant no. 14-14-00988) and National Institute of General Medical Sciences (R01 GM10407) to K.S.

Acknowledgements. To Dr M. Yakimov for help in organizing sample collection in Etna and Vesuvius regions, Svetlana Dubiley and Dmitry Travin for sample collection in Kamchatka and Viktor Fedorchuk for collecting Vesuvius 2018 samples.

References

1. Hendrix RW, Smith MC, Burns RN, Ford ME, Hatfull GF. 1999 Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. *Proc. Natl Acad. Sci. USA* **96**, 2192–2197. (doi:10.1073/pnas.96.5.2192)
2. Suttle CA. 2007 Marine viruses—major players in the global ecosystem. *Nat. Rev. Microbiol.* **5**, 801–812. (doi:10.1038/nrmicro1750)

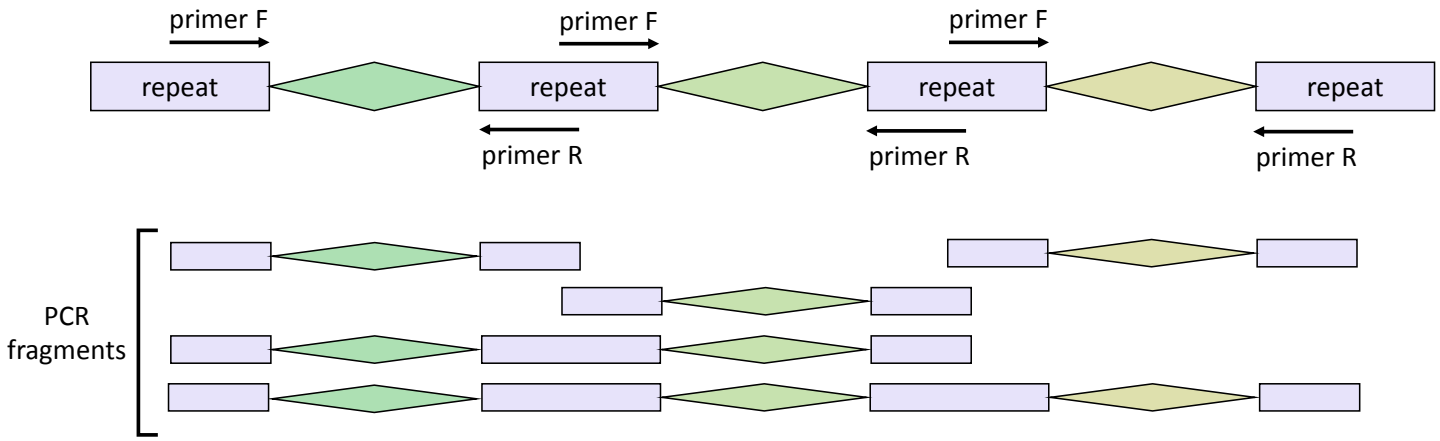
3. Rohwer F, Prangishvili D, Lindell D. 2009 Roles of viruses in the environment. *Environ. Microbiol.* **11**, 2771–2774. (doi:10.1111/j.1462-2920.2009.02101.x)
4. Goldfarb T, Sberro H, Weinstock E, Cohen O, Doron S, Charpak-Amikam Y, Afik S, Ofir G, Sorek R. 2015 BREX is a novel phage resistance system widespread in microbial genomes. *EMBO J.* **34**, 169–183. (doi:10.15252/embj.201489455)
5. Ofir G, Melamed S, Sberro H, Mukamel Z, Silverman S, Yaakov G, Doron S, Sorek R. 2018 DISARM is a widespread bacterial defence system with broad anti-phage activities. *Nat. Microbiol.* **3**, 90–98. (doi:10.1038/s41564-017-0051-0)
6. Doron S, Melamed S, Ofir G, Leavitt A, Lopatina A, Keren M, Amitai G, Sorek R. 2018 Systematic discovery of antiphage defense systems in the microbial pangenome. *Science* **359**, eaar4120. (doi:10.1126/science.aar4120)
7. Makarova KS *et al.* 2015 An updated evolutionary classification of CRISPR-Cas systems. *Nat. Rev. Microbiol.* **13**, 722–736. (doi:10.1038/nrmicro3569)
8. Yosef I, Goren MG, Qimron U. 2012 Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. *Nucleic Acids Res.* **40**, 5569–5576. (doi:10.1093/nar/gks216)
9. Deltcheva E, Chylinski K, Sharma CM, Gonzales K, Chao Y, Pirzada ZA, Eckert MR, Vogel J, Charpentier E. 2011 CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471**, 602–607. (doi:10.1038/nature09886)
10. Koonin EV, Makarova KS, Zhang F. 2017 Diversity, classification and evolution of CRISPR-Cas systems. *Curr. Opin. Microbiol.* **37**, 67–78. (doi:10.1016/j.mib.2017.05.008)
11. Gudbergsdóttir SR, Menzel P, Krogh A, Young M, Peng X. 2016 Novel viral genomes identified from six metagenomes reveal wide distribution of archaeal viruses and high viral diversity in terrestrial hot springs. *Environ. Microbiol.* **18**, 863–874. (doi:10.1111/1462-2920.13079)
12. Held NL, Herrera A, Cadillo-Quiroz H, Whitaker RJ. 2010 CRISPR associated diversity within a population of *Sulfolobus islandicus*. *PLoS ONE* **5**, e12988. (doi:10.1371/journal.pone.0012988)
13. Robles-Sikisaka R, Naidu M, Ly M, Salzman J, Abeles SR, Boehm TK, Pride DT. 2014 Conservation of streptococcal CRISPRs on human skin and saliva. *BMC Microbiol.* **14**, 146. (doi:10.1186/1471-2180-14-146)
14. Pride DT, Sun CL, Salzman J, Rao N, Loomer P, Armitage GC, Banfield JF, Relman DA. 2011 Analysis of streptococcal CRISPRs from human saliva reveals substantial sequence diversity within and between subjects over time. *Genome Res.* **21**, 126–136. (doi:10.1101/gr.111732.110)
15. Kuno S, Yoshida T, Kaneko T, Sako Y. 2012 Intricate interactions between the bloom-forming cyanobacterium *Microcystis aeruginosa* and foreign genetic elements, revealed by diversified clustered regularly interspaced short palindromic repeat (CRISPR) signatures. *Appl. Environ. Microbiol.* **78**, 5353–5360. (doi:10.1128/AEM.00626-12)
16. Andersson AF, Banfield JF. 2008 Virus population dynamics and acquired virus resistance in natural microbial communities. *Science* **320**, 1047–1050. (doi:10.1126/science.1157358)
17. Kunin V *et al*. 2008 A bacterial metapopulation adapts locally to phage predation despite global dispersal. *Genome Res.* **18**, 293–297. (doi:10.1101/gr.6835308)
18. Tyson GW, Banfield JF. 2008 Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. *Environ. Microbiol.* **10**, 200–207.
19. Held NL, Whitaker RJ. 2009 Viral biogeography revealed by signatures in *Sulfolobus islandicus* genomes. *Environ. Microbiol.* **11**, 457–466. (doi:10.1111/j.1462-2920.2008.01784.x)
20. Held NL, Herrera A, Whitaker RJ. 2013 Reassortment of CRISPR repeat-spacer loci in *Sulfolobus islandicus*. *Environ. Microbiol.* **15**, 3065–3076.
21. Kimura S, Uehara M, Morimoto D, Yamanaka M, Sako Y, Yoshida T. 2018 Incomplete selective sweeps of microcystis population detected by the leader-end CRISPR fragment analysis in a natural pond. *Front. Microbiol.* **9**, 425. (doi:10.3389/fmicb.2018.00425)
22. Stern A, Mick E, Tirosh I, Sagy O, Sorek R. 2012 CRISPR targeting reveals a reservoir of common phages associated with the human gut microbiome. *Genome Res.* **22**, 1985–1994. (doi:10.1101/gr.138297.112)
23. Snyder JC, Bateson MM, Lavin M, Young MJ. 2010 Use of cellular CRISPR (clusters of regularly interspaced short palindromic repeats) spacer-based microarrays for detection of viruses in environmental samples. *Appl. Environ. Microbiol.* **76**, 7251–7258. (doi:10.1128/AEM.01109-10)
24. Emerson JB, Andrade K, Thomas BC, Norman A, Allen EE, Heidelberg KB, Banfield JF. 2013 Virus-host and CRISPR dynamics in archaea-dominated hypersaline Lake Tyrrell, Victoria, Australia. *Archaea* **2013**, 1–12. (doi:10.1155/2013/370871)
25. Sorokin VA, Gelfand MS, Artamonova II. 2010 Evolutionary dynamics of clustered irregularly interspaced short palindromic repeat systems in the ocean metagenome. *Appl. Environ. Microbiol.* **76**, 2136–2144. (doi:10.1128/AEM.01985-09)
26. Pride DT, Salzman J, Relman DA. 2012 Comparisons of clustered regularly interspaced short palindromic repeats and viromes in human saliva reveal bacterial adaptations to salivary viruses. *Environ. Microbiol.* **14**, 2564–2576. (doi:10.1111/j.1462-2920.2012.02775.x)
27. Iranzo J, Lobkovsky AE, Wolf YI, Koonin EV. 2013 Evolutionary dynamics of the prokaryotic adaptive immunity system CRISPR-Cas in an explicit ecological context. *J. Bacteriol.* **195**, 3834–3844. (doi:10.1128/JB.00412-13)
28. Childs LM, Held NL, Young MJ, Whitaker RJ, Weitz JS. 2012 Multiscale model of CRISPR-induced coevolutionary dynamics: diversification at the interface of Lamarck and Darwin. *Evolution* **66**, 2015–2029. (doi:10.1111/j.1558-5646.2012.01595.x)
29. Childs LM, England WE, Young MJ, Weitz JS, Whitaker RJ. 2014 CRISPR-induced distributed immunity in microbial populations. *PLoS ONE* **9**, e101710. (doi:10.1371/journal.pone.0101710)
30. Strotksaya A, Semenova E, Savitskaya E, Severinov K. 2015 Rapid multiplex creation of *Escherichia coli* strains capable of interfering with phage infection through CRISPR. In *Methods in Molecular Biology* (eds M Lundgren, E Charpentier, P Fineran), pp. 147–159. New York, NY: Humana Press.
31. Lopatina A, Medvedeva S, Shmakov S, Logacheva MD, Krylenkov V, Severinov K. 2016 Metagenomic analysis of bacterial communities of Antarctic surface snow. *Front. Microbiol.* **7**, 398. (doi:10.3389/fmicb.2016.00398)
32. Besemer J. 2001 GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.* **29**, 2607–2618. (doi:10.1093/nar/29.12.2607)
33. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B. 2000 Artemis: sequence visualization and annotation. *Bioinformatics* **16**, 944–945. (doi:10.1093/bioinformatics/16.10.944)
34. Edgar RC. 2010 Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461. (doi:10.1093/bioinformatics/btq461)
35. Dixon P. 2003 VEGAN, a package of R functions for community ecology. *J. Veg. Sci.* **14**, 927. (doi:10.1111/j.1654-1103.2003.tb02228.x)
36. Biswas A, Gagnon JN, Brouns SJJ, Fineran PC, Brown CM. 2013 CRISPRTarget: bioinformatic prediction and analysis of crRNA targets. *RNA Biol.* **10**, 817–827. (doi:10.4161/rna.24046)
37. Lange SJ, Alkhnbashi OS, Rose D, Will S, Backofen R. 2013 CRISPRmap: an automated classification of repeat conservation in prokaryotic adaptive immune systems. *Nucleic Acids Res.* **41**, 8034–8044. (doi:10.1093/nar/gkt606)
38. Grissa I, Vergnaud G, Pourcel C. 2007 The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinf.* **8**, 172. (doi:10.1186/1471-2105-8-172)
39. Shmakov SA, Sitnik V, Makarova KS, Wolf YI, Severinov KV, Koonin EV. 2017 The CRISPR spacer space is dominated by sequences from species-specific mobilomes. *MBio* **8**, e01397-17. (doi:10.1128/mBio.01397-17)
40. Staals RHJ *et al*. 2014 RNA targeting by the type III-A CRISPR-Cas Csm complex of *Thermus thermophilus*. *Mol. Cell* **56**, 518–530. (doi:10.1016/j.molcel.2014.10.005)
41. Staals RHJ *et al*. 2013 Structure and activity of the RNA-targeting Type III-B CRISPR-Cas complex of *Thermus thermophilus*. *Mol. Cell* **52**, 135–145. (doi:10.1016/j.molcel.2013.09.013)
42. Heidelberg JF, Nelson WC, Schoenfeld T, Bhaya D. 2009 Germ warfare in a microbial mat community: CRISPRs provide insights into the co-evolution of host and viral genomes. *PLoS ONE* **4**, e4169. (doi:10.1371/journal.pone.0004169)
43. Savitskaya E, Semenova E, Dedkov V, Metlitskaya A, Severinov K. 2013 High-throughput analysis of type I-E CRISPR/Cas spacer acquisition in *E. coli*. *RNA Biol.* **10**, 716–725. (doi:10.4161/rna.24325)
44. Shah SA, Garrett RA. 2011 CRISPR/Cas and Cmr modules, mobility and evolution of adaptive

immune systems. *Res. Microbiol.* **162**, 27–38. (doi:10.1016/j.resmic.2010.09.001)

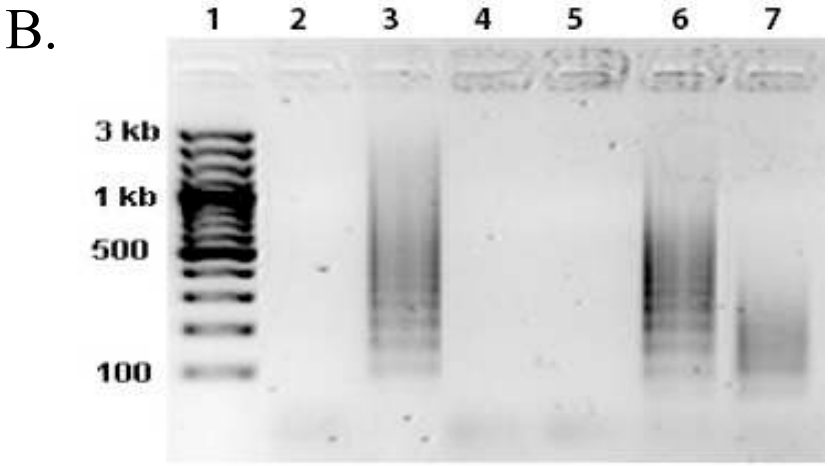
45. Yu MX, Slater MR, Ackermann H-W. 2006 Isolation and characterization of *Thermus* bacteriophages. *Arch. Virol.* **151**, 663–679. (doi:10.1007/s00705-005-0667-x)
46. Tamakoshi M *et al*. 2011 Genomic and proteomic characterization of the large Myoviridae bacteriophage ϕTMA of the extreme thermophile *Thermus thermophilus*. *Bacteriophage* **1**, 152–164. (doi:10.4161/bact.1.3.16712)
47. Naryshkina T *et al*. 2006 *Thermus thermophilus* bacteriophage phiYS40 genome and proteomic characterization of virions. *J. Mol. Biol.* **364**, 667–677. (doi:10.1016/j.jmb.2006.08.087)
48. Nagayoshi Y *et al*. 2016 Physiological properties and genome structure of the hyperthermophilic filamentous phage φOH3 which infects *Thermus thermophilus* HB8. *Front. Microbiol.* **7**, 50. (doi:10.3389/fmicb.2016.00050)
49. Matsushita I, Yanase H. 2009 The genomic structure of *Thermus* bacteriophage IN93. *J. Biochem.* **146**, 775–785. (doi:10.1093/jb/mvp125)
50. Severinov K, Minakhin L, Sekine S-I, Lopatina A, Yokoyama S. 2014 Molecular basis of RNA polymerase promoter specificity switch revealed through studies of *Thermus* bacteriophage transcription regulator. *Bacteriophage* **4**, e29399. (doi:10.4161/bact.29399)
51. Hale CR, Zhao P, Olson S, Duff MO, Graveley BR, Wells L, Terns RM, Terns MP. 2009 RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. *Cell* **139**, 945–956. (doi:10.1016/j.cell.2009.07.040)
52. Strotskaya A, Savitskaya E, Metlitskaya A, Morozova N, Datsenko KA, Semenova E, Severinov K. 2017 The action of *Escherichia coli* CRISPR-Cas system on lytic bacteriophages with different lifestyles and development strategies. *Nucleic Acids Res.* **45**, 1946–1957.

Supplementary figure S1. Amplification of Thermus CRISPR spacers from enrichment cultures.

A.

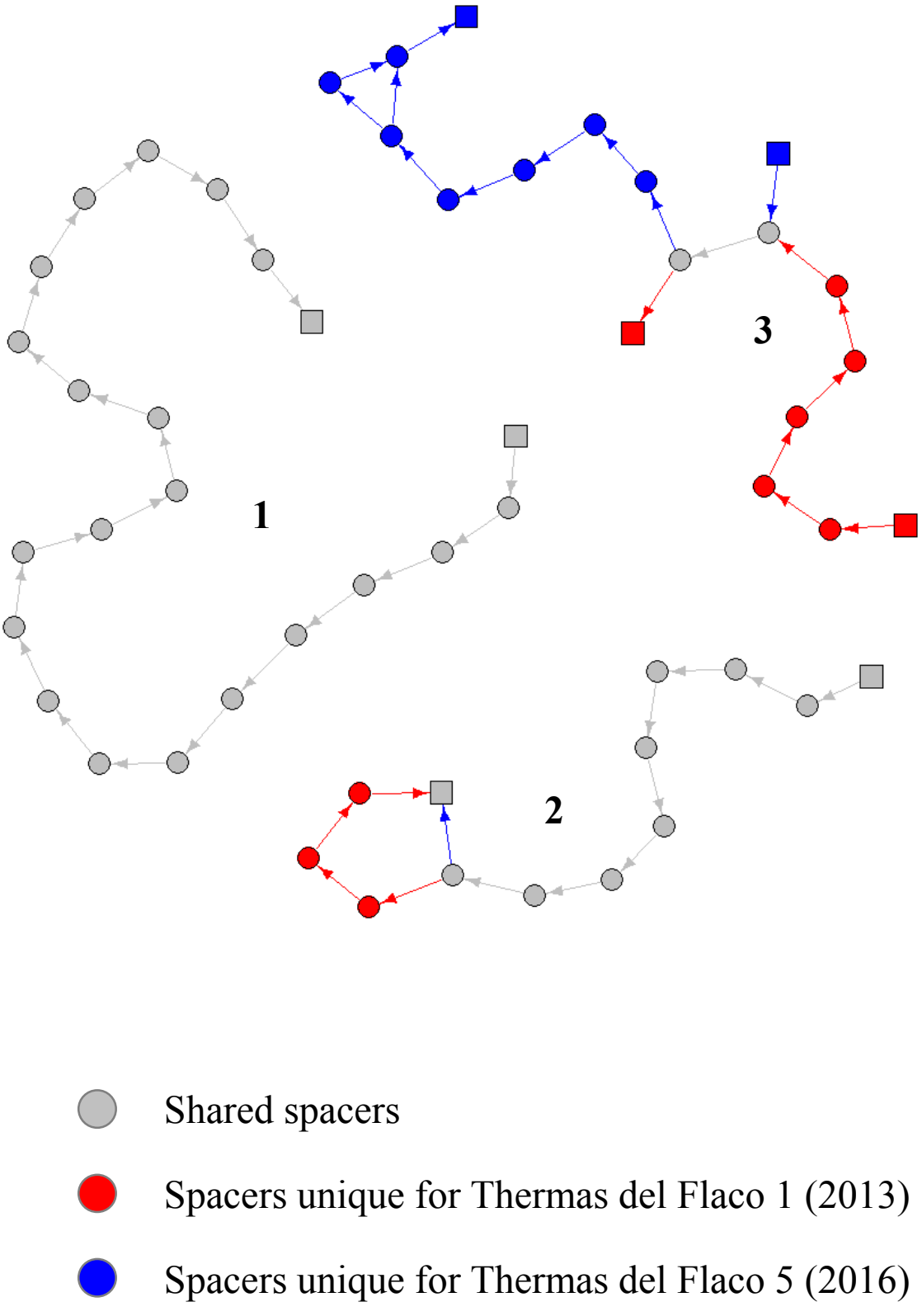


B.

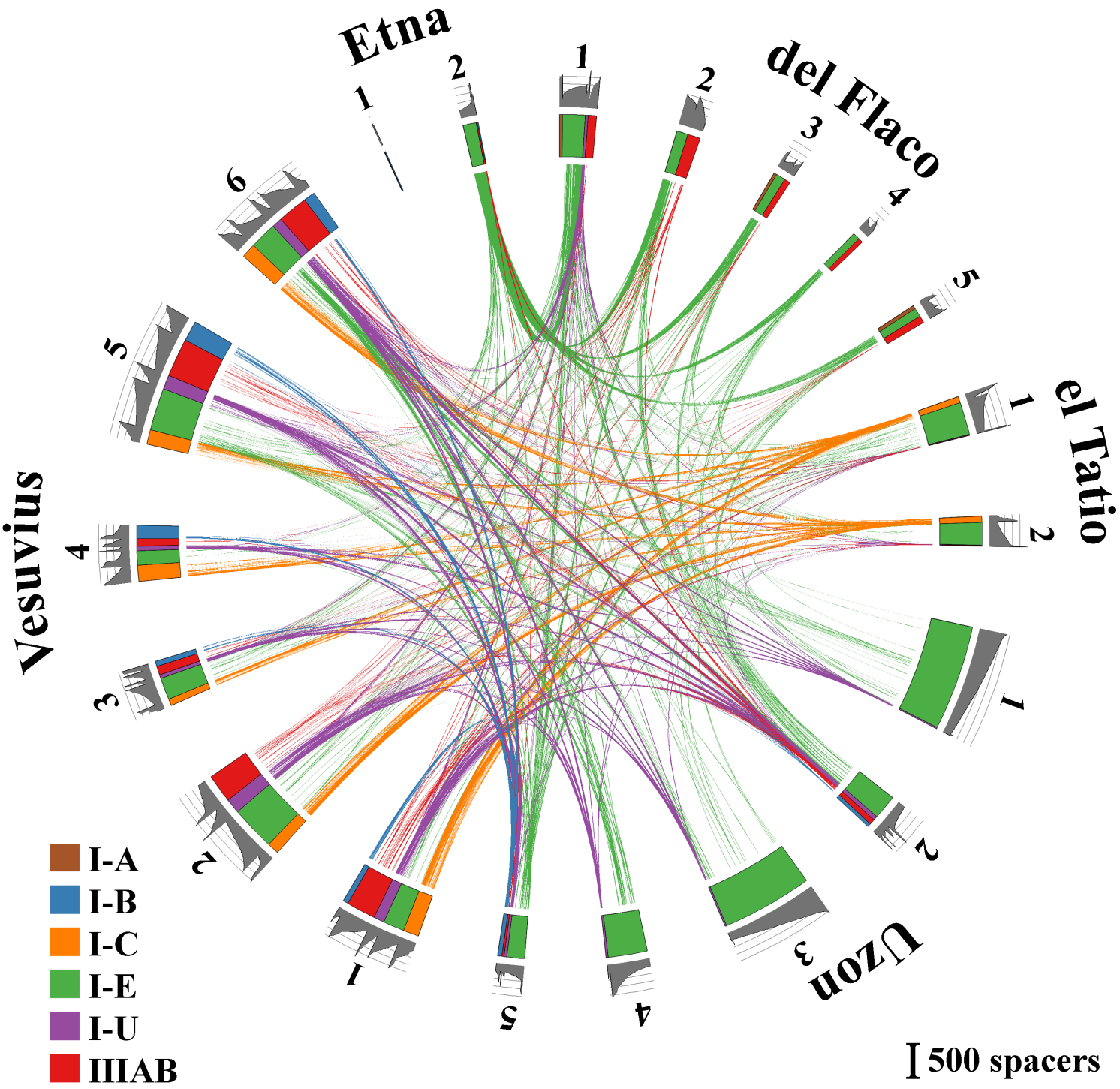


A. A strategy used to amplify spacers associated with Thermus CRISPR repeats from environmental samples.
B. An agarose gel showing the results of separation of products of PCR amplification with primers specific to Thermus I-E type CRISPR repeat using the following DNA as amplification template.
lane 2 - negative control, no input DNA;
lane 3 - T. thermophilus HB8 genomic DNA;
lane 4 - Thermas del Flaco sample 1, no enrichment;
lane 5 - Thermas del Flaco sample 2, no enrichment;
lane 6 - Thermas del Flaco sample 1, enrichment culture;
lane 7 - Thermas del Flaco sample 2, enrichment culture.
lane 1 is a DNA molecular weight marker.

Supplementary Figure S2. Intersection of reconstructed I-A CRISPR arrays for Thermas del Flaco 1 and Thermas del Flaco 5. Arrows show direction from leader sequence.



Supplementry Figure S3. The diversity of CRISPR spacers in environmental *Thermus* samples.
The diversity of 14872 spacers (spacer cluster centers) associated with *Thermus* CRISPR repeats from enrichment cultures from samples collected at indicated sites is shown in circular diagram. Spacers from different locations that differ from each other by less than 2 nucleotides are connected by matching color lines.

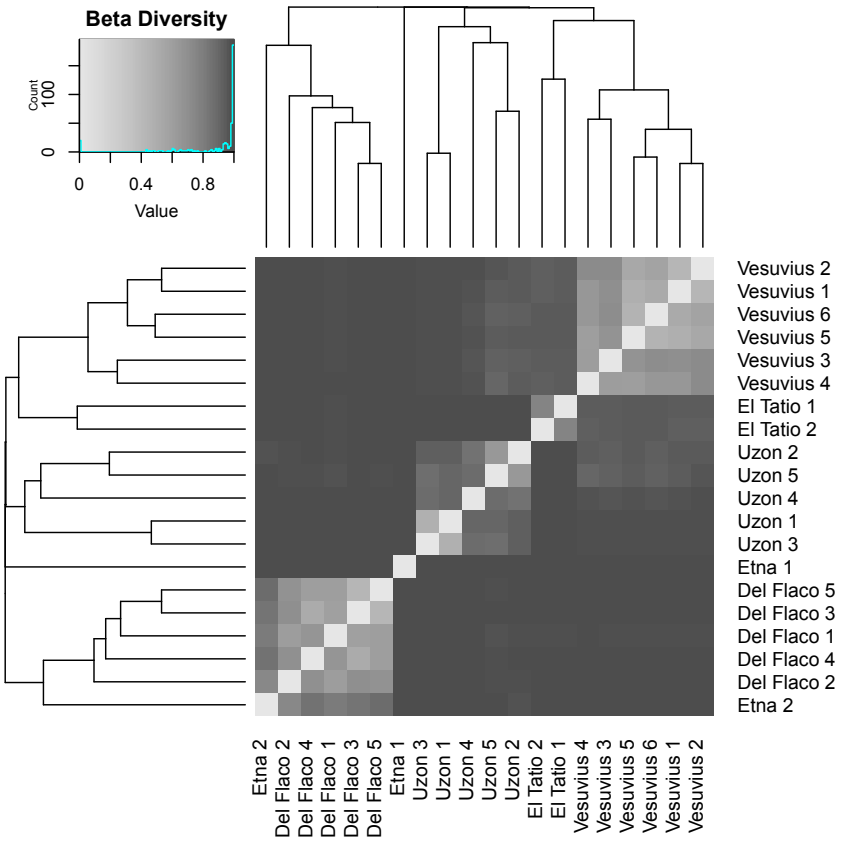


Supplementary Figure S4. Clustering of samples by beta-diversity

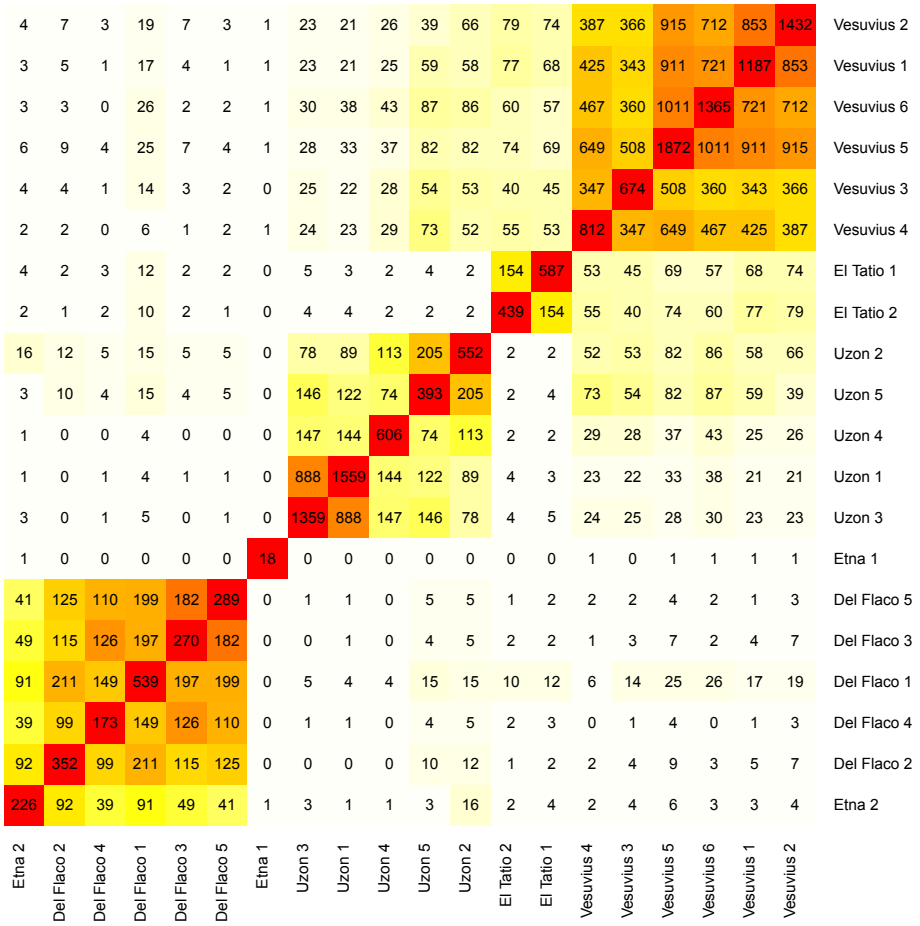
A. Grey-scale heatmap with dendrogam of sample clustering

B. Red-scale heatmap with number of shared spacers indicated in each cell

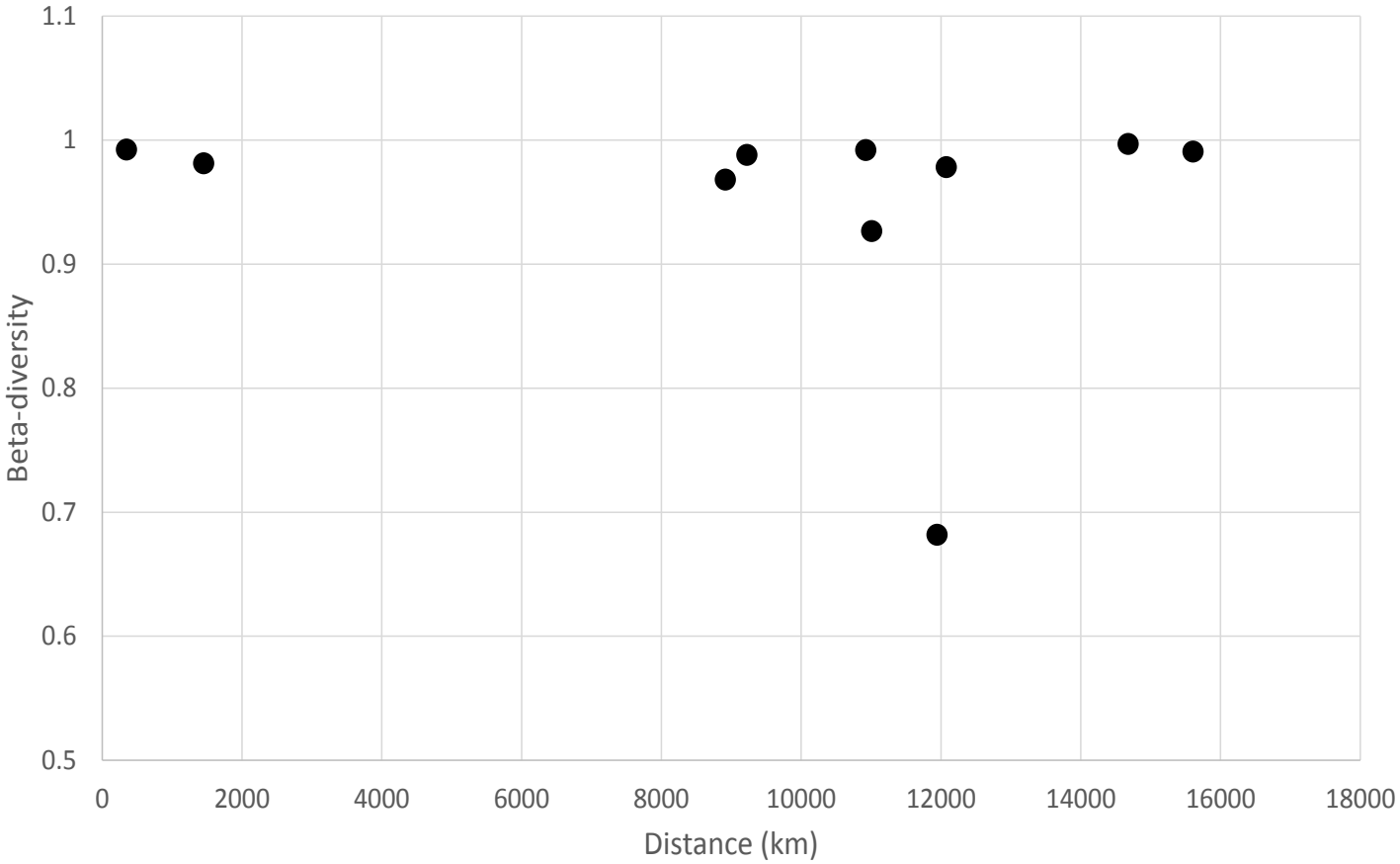
A



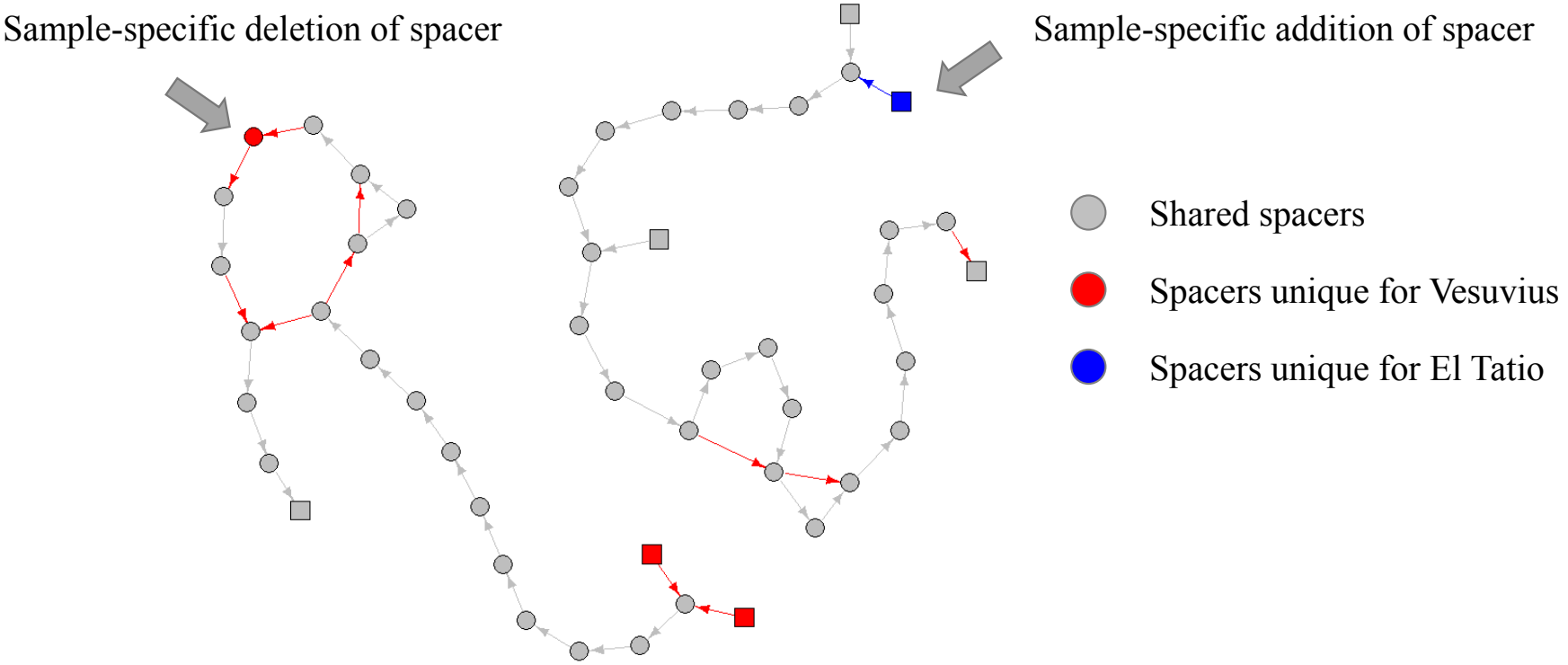
B



Supplementary Figure S5.

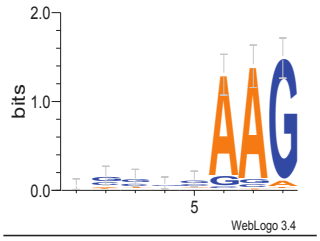


Supplementary Figure S6. Intersection of reconstructed I-C CRISPR arrays between Vesuvius and El Tatio samples. Arrows show direction from leader sequence.

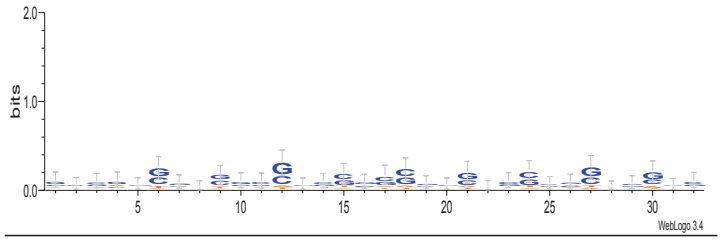


Supplementary figure S7 PAM motifs identified by matches to *Thermus* phages

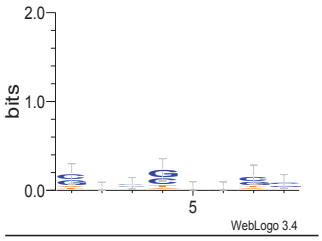
I-E system



5’ flank

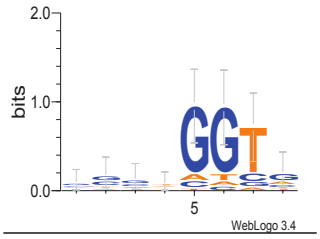


protospacer

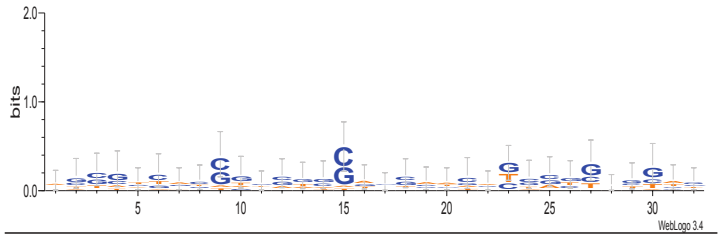


3’ flank

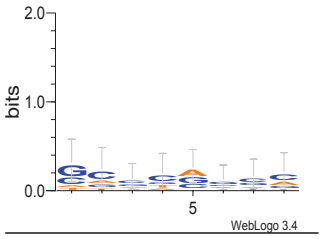
I-B system



5’ flank

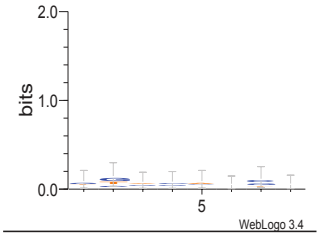


protospacer

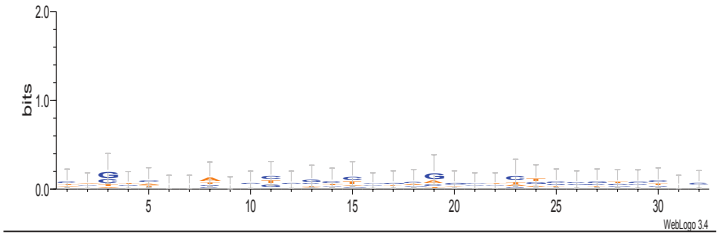


3’ flank

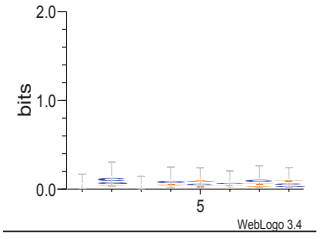
IIIAB system



5’ flank

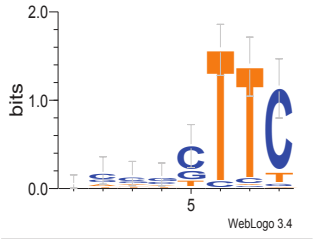


protospacer

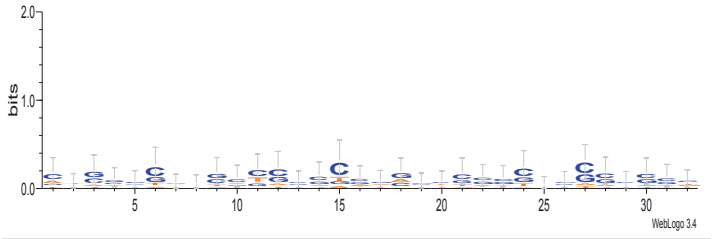


3’ flank

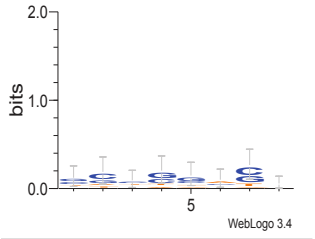
I-C system



5’ flank

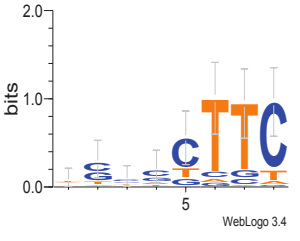


protospacer

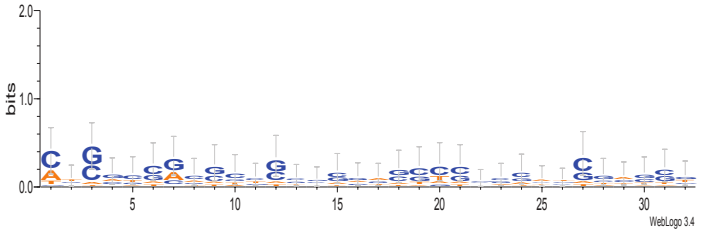


3’ flank

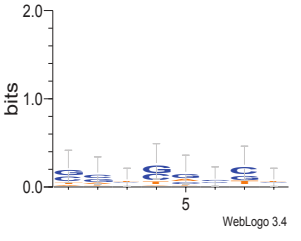
I-U system



5’ flank



protospacer



3’ flank

Supplemetary table S1.

System type	primer name	primer sequence	% of primer in the primer mix (total 100%)
IIIAB system	IIIAB_rep_F	GGGCTCAATCCCTTGCAAC	50%
	IIIAB_rep_R1	CCCGTAAGGGGATTGCGAC	45%
	IIIAB_rep_R2	CCCCGTAAGGGGATKRHYDC	5%
I-E system	I-E_rep_F	CCACRYGYGTGGGGACTAC	50%
	I-E_rep_R1	RCRYGTGGGGATGGMCCG	45%
	I-E_rep_R2	RCRYGTGGGGATGGMCSD	5%
I-C system	I-C_R	GGGCCGGTGAGGATTGAAAC	50%
	I-C_F	CTTTCGGGCCGGTGCAAC	50%
I-B system	I-B_rep_R	AGCCTCGTAGAGGATTGAAAC	54%
	I-B_rep_F	GGCTAACGAGGTTTGCAAC	53%
	I-B_rep_rest_F	GCCTCGTAGAGGATTGAAAC	6%
	I-B_rep_rest_R	GCTRACGRGGTTTGCAAC	6%
	I-B_uniq_F	CTCGTAGAGGATTGGCCA	1%
I-A system	I-A_rep_R	CGTACCTATAAGGGTTTGAAA	35%
	I-A_rep_F	CCTTATAGGTACGGTTCAAAG	35%
	I-A_rep_F1	ACCTATGAGGGTTTGAAAC	15%
	I-A_rep_R2	TCATAGGTACGGTCAGAAC	15%
I-U system	I-U_R	CAGCTTGGCTACGTTGCAGG	50%
	I-U_F	AGCTGTGAAGCTTGGATGCAA	50%

Thermus 16S	Thermus_F3	GTCTCCTGGGGGCCGAAGCTA	50%
	Thermus_R1	ACCCCAGGCTTTCACCCGGGT	50%

Supplementary table S2. A full list of 26 Thermus isolates analyzed in this work.

№	Strain	number of spacers	isolation cite
1	Thermus amyloliquefaciens YIM 77409	53	China
2	Thermus antranikianii DSM 12462	1	Iceland
3	Thermus aquaticus YT-1	48	USA
4	Thermus aquaticus Y51MC23	25	USA
5	Thermus arciformis CGMCC 1_6992	17	China
6	Thermus brockianus GE-1	101	Iceland
7	Thermus filiformis ATCC 43280	55	New Zealand
8	Thermus igniterrae ATCC 700962	78	Iceland
9	Thermus kawarayensis JCM 12314	172	Japan
10	Thermus oshimai DSM 12092	115	Portugal
11	Thermus oshimai JL-2	140	USA
12	Thermus parvatiensis RL	12	India
13	Thermus scotoductus DSM 8553	61	Iceland
14	Thermus scotoductus KI2	36	Hawaii
15	Thermus scotoductus SA-01	49	South Africa
16	Thermus scotoductus K1	5	Azerbaijan
17	Thermus sp 2_9	25	Argentina
18	Thermus sp CCB_US3_UF1	96	Malaysia
19	Thermus sp JCM 17653	90	Japan
20	Thermus sp NMX2_A1	7	USA
21	Thermus tengchongensis YIM 77401	67	USA
22	Thermus thermophilus ATCC 33923	71	Japan
23	Thermus thermophilus HB27	74	Japan
24	Thermus thermophilus HB8	124	Japan
25	Thermus thermophilus JL-18	75	USA
26	Thermus thermophilus SG0_5JP17-16	54	USA

Supplementary table S4. Clustering statistics

Sample	Year	CRISPR spacers, total	Clusters	Good's criterion	Alpha diversity (Shannon)	Schao	Clusters (n > 10)
Vesuvius 1	2014	1 730 954	2 610	0.73	5.47	3512 ± 91	1 212
Vesuvius 2	2014	1 777 462	3 235	0.75	4.84	4208 ± 91	1 468
Vesuvius 3	2018	573 464	7 991	0.29	5.54	26446 ± 805	680
Vesuvius 4	2018	657 890	10 078	0.28	5.70	35343 ± 991	823
Vesuvius 5	2018	546 902	10 675	0.31	6.73	49915 ± 1763	1 916
Vesuvius 6	2018	759 772	12 658	0.27	6.20	52197 ± 1470	1 381
Etna 1	2012	1 027 256	286	0.40	2.84	677 ± 88	18
Etna 2	2012	926 693	765	0.66	3.14	1201 ± 75	226
El Tatio 1	2014	471 175	1 201	0.81	4.21	1468 ± 46	588
El Tatio 2	2014	456 459	1 122	0.70	3.09	1573 ± 67	439
Del Flaco 1	2013	2 103 097	965	0.67	4.88	1745 ± 132	542
Del Flaco 2	2013	3 653 540	1 041	0.55	5.35	1814 ± 100	355
Del Flaco 3	2016	15 328	595	0.87	5.40	642 ± 14	271
Del Flaco 4	2016	11 000	448	0.85	4.97	484 ± 11	173
Del Flaco 5	2016	14 875	695	0.87	5.63	746 ± 14	289
Uzon 1	2018	520 819	5 923	0.40	6.52	23070 ± 1070	1 572
Uzon 2	2018	512 485	11 477	0.40	4.51	23682 ± 420	555
Uzon 3	2018	1 011 741	10 508	0.25	5.98	53793 ± 1905	1 363
Uzon 4	2018	774 875	18 506	0.27	4.98	54063 ± 954	607
Uzon 5	2018	293 179	9 064	0.35	5.05	21357 ± 471	394

CHAPTER IV

Virus-borne mini-CRISPR arrays promote interviral conflicts and virus speciation

Introduction:

In this Chapter, the CRISPRome analysis was performed for an archaeal hyperthermophilic community – members of the order Sulfolobales from Beppu hot spring in Japan. Instead of comparison of geographically distant communities (like in Chapter III), Chapter IV focuses on short-term dynamics of spacer diversity. CRISPR spacers from the original environmental sample and 10-days and 20-days enrichment cultures were analyzed. We serendipitously discovered CRISPR mini-arrays in the genomes of SPV1 and SPV2 viruses, which became the main focus of the project.

Contribution:

I performed all the bioinformatics analysis, prepared Figures and drafted the text of the manuscript.

Virus-borne mini-CRISPR arrays promote interviral conflicts and virus speciation

Sofia Medvedeva[1,2,3], Ying Liu[1], Eugene V. Koonin[4], Konstantin Severinov[2], David Prangishvili[1], Mart Krupovic[1]*

1 – Institut Pasteur, Unité Biologie Moléculaire du Gène chez les Extrêmophiles, 75015, Paris, France.
2 – Center of Life Sciences, Skolkovo Institute of Science and Technology, Skolkovo, Russia.
3 – Sorbonne Université, Collège doctoral, 75005, Paris, France.
4 – National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD 20894, USA.

* - correspondence to
mart.krupovic@pasteur.fr

The incessant arms race between viruses and cells drives the evolution of both conflicting parties, structuring their populations across time and space[1-3], spawning major evolutionary innovations[4-6], and affecting the major biogeochemical cycles[7]. The CRISPR-Cas adaptive immunity systems are at the forefront of antivirus defense in bacteria and archaea[8] and can specifically target viral genomes and other mobile elements that carry protospacers – sequences matching spacers stored in the CRISPR arrays. We performed deep sequencing of the CRISPRome — all spacers contained in a microbiome — of hyperthermophilic archaea recovered directly from environmental samples and from the laboratory enrichment cultures. The 25 million CRISPR spacers sequenced from a single sampling site dwarf the diversity of spacers from all available Sulfolobales isolates and display complex temporal dynamics. The majority of the spacers with identifiable protospacers target viruses from the same sampling site, indicative of local adaptation. Comparison of closely related virus strains shows that CRISPR targeting drives virus genome microevolution. We discover that some of the most abundant spacers in the CRISPRome come from mini-arrays carried by archaeal viruses themselves. These mini-arrays contain only 1-2 spacers, are preceded by leader sequences but are not associated with cas genes. Spacers from these mini-arrays target closely related viruses present in the same population. Targeting by virus-borne spacers might represent a distinct mechanism of superinfection exclusion and appears to promote archaeal virus speciation.

Viruses and other mobile genetic elements (MGEs) have likely coevolved with their cellular hosts for billions of years, ever since the dawn of life, and established a range of complex interaction regimes[5,6]. At the interface of these interactions, various mechanisms of defense and counter-defense have emerged[9-13]. These vary from physical barriers, which abrogate the delivery of foreign genetic material into the host interior, to specific recognition and degradation of the invading nucleic acids inside the cell, to suicide of infected cells that can save the clonal population[14,15]. Concurrently, MGEs evolved elaborate ways to overcome the host defenses. The prime example of such systems in many bacteria and most archaea is the CRISPR-Cas adaptive immunity and the MGE-encoded anti-CRISPR proteins[11]. The defense systems evolve by widely different mechanisms which often involve recruitment of MGEs or their components. Once in existence, the defense and counter-defense systems can change their 'owner' according to the 'guns-for-hire' concept[16]. Indeed, CRISPR-Cas systems are not exclusive to cellular organisms and have been captured and exploited by various MGEs, including bacteriophages, plasmids and transposons[17,18].

Hyperthermophilic archaea of the order Sulfolobales harbor some of the most complex among the studied CRISPR-Cas systems: most of the genomes contain several CRISPR arrays with different CRISPR repeats, several adaptation modules for acquisition of new spacers into CRISPR arrays and several type I and type III interference modules that degrade the DNA and/or RNA molecules of encountered MGEs[19]. Concurrently, members of the Sulfolobales harbor an extremely diverse virome[20]. As a countermeasure to sophisticated defense systems of the host, at least some viruses of Sulfolobales encode anti-CRISPR proteins[21,22]. CRISPR-Cas immunity of Sulfolobus has been extensively explored in vitro, providing insights into the mechanisms of adaptation, expression and interference[23-25]. In parallel, in vivo experiments have demonstrated that new spacers can be inserted into the CRISPR arrays upon infection with a single or multiple viruses[26,27]. Interference with the targeted MGE at the level of DNA and/or RNA has been described for different CRISPR interference modules[28,29].

The sequence of each CRISPR spacer and its position in the array, respectively, provide information about the encountered MGEs and the order of their interaction with the host. Analysis of the CRISPR spacer content in Sulfolobus populations demonstrated high spacer diversity[30,31], reassortments of CRISPR arrays between strains in the course of 10 years[32], as well as specificity of CRISPR spacers to local viruses[33,34]. To gain insights into the diversity and dynamics of CRISPRome — all spacers contained in a microbiome — we studied the natural population of Sulfolobales in the previously

characterized environmental samples from a thermal field in Beppu, Japan[35] (see Methods). To this end, we amplified by PCR[36] the CRISPR spacers from the four principal CRISPR repeat sequences present in Sulfolobales[37], followed by high-throughput sequencing (HTS) of the amplicons (see Methods). Notably, in Sulfolobus, the interference modules of types I and III can utilize crRNA from CRISPR arrays with different repeat sequences[38], precluding unambiguous assignment of CRISPR arrays to interference modules. Thus, hereinafter, we refer to the four consensus CRISPR repeat sequences as A, B, C, and D (Figure 1). All four CRISPR repeat sequences are exclusive to the Sulfolobales, including the genera Sulfolobus, Acidianus and Metallosphaera (Supplementary table 1). The temporal dynamics of the CRISPRome was analyzed in two parallel series of enrichment cultures established from environmental samples J14 and J15 (Ref [35]), in media that favor the growth of Sulfolobus and Acidianus species prevalent in terrestrial hot springs and grow well under laboratory conditions (Supplementary Figure 1).

More than 25 million spacers were sequenced from all the samples (Supplementary table 2), which after clustering of sequences with 85% identity resulted in 40,705 unique spacer clusters (Supplementary Data 1). The clustered spacer collection obtained here from a single sampling site dwarfs (6-fold increase) the size of the Sulfolobales spacer database from strains (n=6699 unique spacers) that have been previously isolated from geographically diverse locations (Figure 1B). The largest intersection (48 spacers) was found between the Beppu spacer set and spacers from sequenced Sulfolobales strains isolated in Japan (Figure 1A), indicative of the presence of a biogeographical pattern in the Sulfolobales virome, consistent with previous observations from other geographical locations[33,34]. The original environmental sample comprised 86% of the 40,705 spacer clusters, with 64% of spacers found exclusively in this sample. In contrast, the 10-days and 20-days enrichment samples, respectively, contained only ~20% and ~15% of the total collection of Beppu spacers (Supplementary Figure 2a). The massive loss of spacer diversity must result from extinction of certain Sulfolobales strains during cultivation under laboratory conditions. Indeed, we found that the initially less abundant spacers (with coverage < 30) were the most strongly affected by the cultivation procedure, with 85% disappearing in the enrichment cultures, whereas only 7% of initially dominant spacers, sequenced more than 500 times, were lost after 20 days (Supplementary Figure 2b). This result indicates that, as one would expect, the bottleneck primarily affects the strains with a small population size.

To assess the provenance of the spacers, we matched the Beppu spacer set against the available Sulfolobales genomes, viruses and plasmids (Figure 1c). Using the threshold of >85% identity over the full length of the alignment, protospacers were identified for ~6% of spacers, a value that is close to the ~7% mean observed in previous analyses of the global dataset of spacers from all available sequenced genomes[39]. Notably, protospacers associated with the C-type CRISPR array were overrepresented in plasmids (P-value < 10^{-5}) and underrepresented in viruses (P-value < 10^{-36}), suggesting specialization among the CRISPR types to combat different types of MGE. The CRISPRome of the Sulfolobales community from Beppu included spacers against 53 viral genomes isolated from all over the world, but the most frequently targeted ones were those sequenced from the same sampling site[35], further indicating local adaptation of the Sulfolobales viruses. Notably, fusellovirus SSV1 isolated from the same Beppu site 35 years ago[40] is the fourth most targeted virus, suggesting that SSV1 and its derivatives are persistent components of the Beppu virome (Figure 1d). Spacers associated with different CRISPR repeat types showed specificity to different viruses (Figure 1d, Supplementary Figure 3), possibly reflecting distinct host ranges of the corresponding viruses. For example, related viruses SBFV1 and SBFV3 are primarily targeted by spacers from types A and D, respectively. Rudivirus SBRV1 is targeted by as many as 841 unique spacers belonging to different types, signifying that SBRV1-like viruses are or were associated with broadly diverse hosts. Such dense coverage of protospacers would allow reconstruction of 53% of the SBRV1 genome. Moreover, tiling the sequences of overlapping spacers allowed assembly of several additional viral contigs (see Supplementary text). Furthermore, mapping the spacers against the

Sulfolobales chromosomes proved to be an efficient approach to identify integrated MGEs (see Supplementary text).

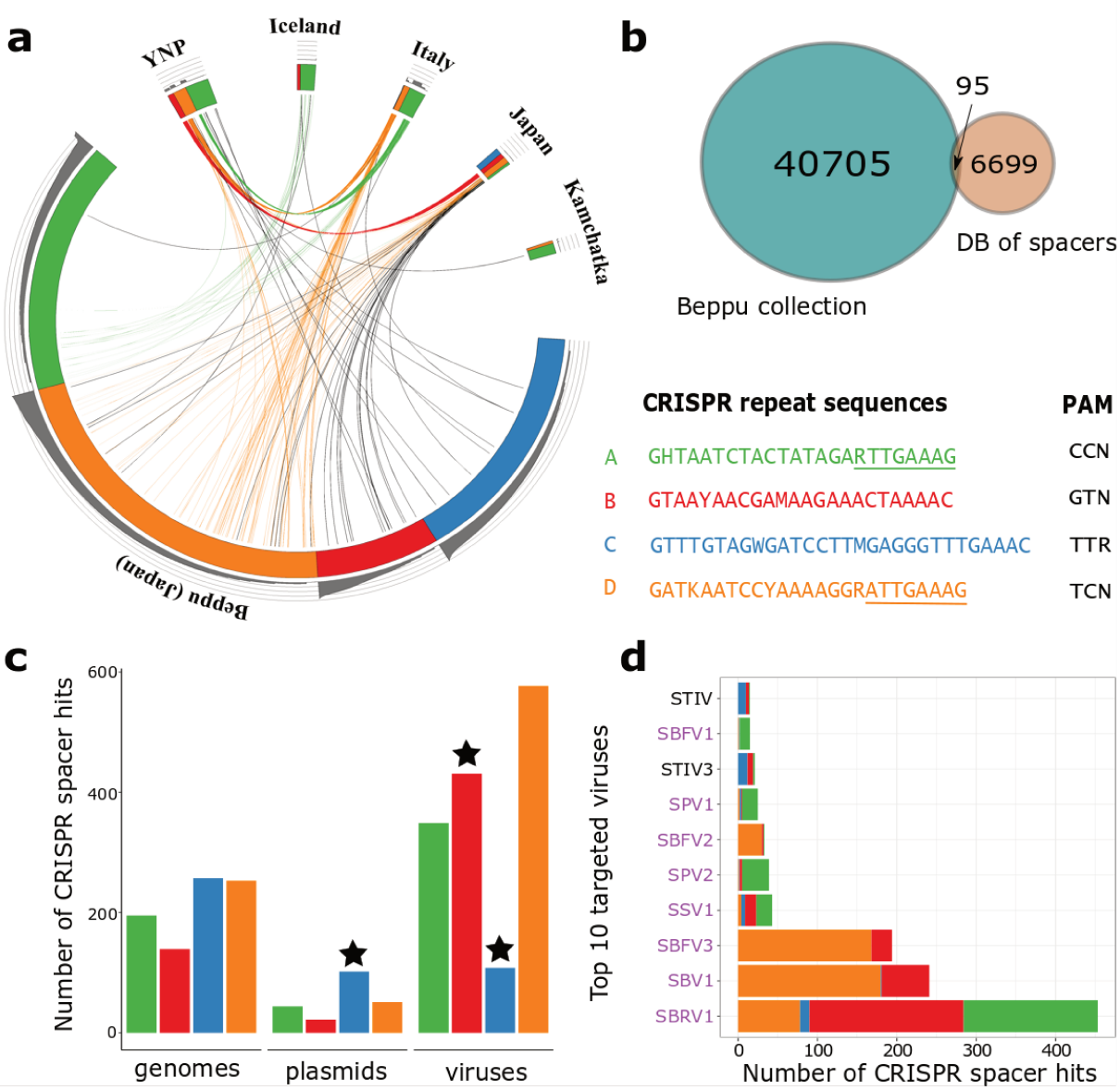


Figure 1. Characteristics of the analyzed spacer collections. A. A circular diagram of spacers amplified from the Beppu hot spring and spacers from the Sulfolobales isolates clustered by the country of origin. Spacers belonging to arrays with the four principal repeat sequences found in Sulfolobales are indicated by identical colors; the color-coded repeat sequences and the corresponding PAMs are shown in the bottom right corner of the figure (note that CRISPR repeat types A and D share the last 8 nucleotides). Spacers that differ from each other by less than 2 nucleotides are connected by lines whose colors correspond to colors indicating CRISPR repeat type. Black lines connect spacers shared by arrays of different types. The outer grey histograms represent the abundance of CRISPR spacers in log10 scale. YNP, Yellowstone National Park, United States. **B.** Intersection of the Beppu spacer collection and spacers from sequenced Sulfolobales isolates available in public databases (DB). The numbers represent the actual number of spacers. **C.** The bar plot showing the numbers of protospacers found in Sulfolobales host genomes, plasmids and viruses. The stars indicate values that differ significantly (chi square test, P-value < 0.001) from the expectation. Colors represent spacers associated with different CRISPR repeat types, as in Figure 1. **D.** The bar plot shows the numbers of protospacers found in the top 10 targeted Sulfolobales viruses. Names of viruses isolated in Beppu, Japan are highlighted with violet color.

To explore the temporal CRISPRome dynamics, we compared the distributions of frequencies of spacers across samples, including the original environmental sample and the enrichment cultures grown in Sulfolobus- and Acidianus-favoring media (Figure 2). Despite possible biases introduced by PCR amplification, CRISPR spacers sequenced from the same replicon generally get similar representation in HTS reads[36]. Therefore, the abundances of spacers show a multimodal distribution (Figure 2), likely reflecting the number of spacer-carrying Sulfolobales strains in the sample. Comparison of the temporal variation in the spacer abundances revealed significant differences between the J14 and J15 samples. Given that the strains from both samples were propagated under the same conditions, and initially displayed similar spacer composition (Supplementary Figure 4), differences in the growth dynamics for some of the strains are unlikely to result from the cultivation in the artificial setting as such, and instead might be caused by viruses present in enrichment cultures. Indeed, we have previously shown that samples J14 and J15 contain different, albeit overlapping, virus populations. Whereas J14 contains SBV1, SBFV1, SBFV3, SBRV1, and SPV2, J15 contains SBV1, SBFV1, SBFV2 and SPV1 (Supplementary Figure 1)[35]. Among these, SPV1 and SPV2 (family Portogloboviridae) are by far the most abundant in the respective samples, accounting for ~90% of all virome reads[35].

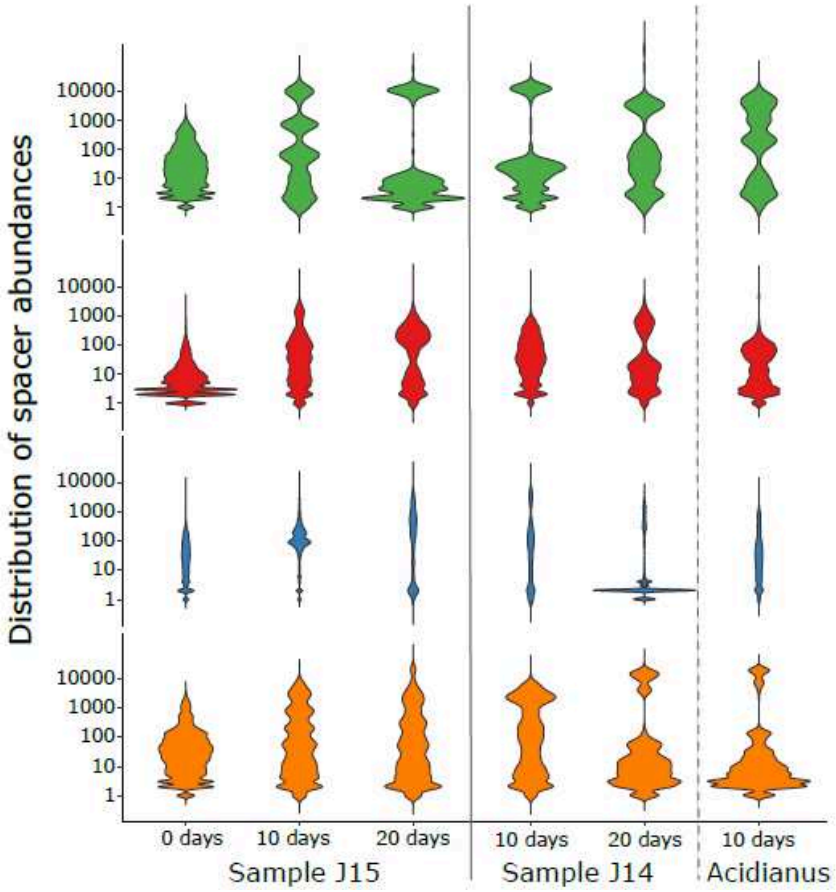


Figure 2. The violin plots show the density of the distribution of spacer abundances in the environmental sample and enrichment cultures established from samples J14 and J15. In the J14 sample, the enrichment culture established in the Acidianus-favoring medium is separated from those established in the Sulfolobus-favoring medium by a dashed line. Plots in each row represent spacers, associated with different CRISPR repeat types and color-coded as in Figure 1. Plots in each row are scaled to have the same area. Log10 scale for the abundance values was used.

To understand the reasons underlying the dominance of SPV1 and SPV2 and their exclusivity to the corresponding samples, we focused on the comparison of spacers targeting the two viruses in J15 and J14,

respectively. Notably, the genomes of SPV1 and SPV2 are 92% identical to each other[35] and mapping of the CRISPR spacers from our dataset showed that genomic location of sequence divergence between SPV1 and SPV2 specifically coincides with targeting by CRISPR spacers (P-value<0.01). Thus, CRISPR targeting is an important factor driving the genome evolution of portogloboviruses (Figure 3a).

Six spacers associated with type A CRISPR repeat targeting SPV1 and SPV2 were the most abundant in the corresponding enrichment cultures. Unexpectedly, 3 of these spacers targeting SPV1 (100% identity) were sequenced from the J15 enrichment culture dominated by SPV1, and conversely, the 3 spacers targeting SPV2 (100% identity) were sequenced from the J14 enrichment dominated by SPV2. This result was inconsistent with an a priori expectation of negative correspondence between the frequency of spacers and the targeted viruses. Furthermore, as mentioned above, abundant CRISPR spacers in our data could be assembled into long CRISPR arrays (see Supplementary Text). However, despite being among the most abundant in our dataset, the 6 SPV1- and SPV2-targeting spacers could not be reconstructed into long arrays, but instead appeared to be located within mini-CRISPR arrays each carrying 1 or 2 spacers. We found that the mini-CRISPR arrays including type A CRISPR repeats flanking the SPV-targeting spacers are encoded in intergenic regions of both SPV1 and SPV2 genomes. Thus, the 6 most abundant CRISPRome spacers originated from mini-CRISPR arrays in SPV1 and SPV2 genomes, rather than from the Sulfolobales genomes. The relative positions of the mini-CRISPR arrays containing 2 spacers in the SPV1 and SPV2 genomes were the same, but the corresponding spacers were different, implying active spacer turnover. These mini-CRISPR arrays are preceded by the promoter-containing leader sequences similar to those found in genomic Sulfolobus CRISPR arrays (Supplementary Figure 5). Unlike in the case of certain bacteriophages and integrated mobile genetic elements, which carry complete CRISPR-Cas systems[17,41], the SPV-encoded mini-CRISPR arrays are not associated with recognizable cas genes. However, given the sequence similarity of the repeats and leader sequences to the corresponding elements of the host[25], it is highly likely that new spacers can be inserted by the endogenous host-encoded adaptation modules. Consistent with this possibility, some of the genetic tools designed for Sulfolobus specifically rely on the recruitment of endogenous Cas proteins to function with artificially designed, plasmid-borne CRISPR spacers targeting a gene of interest[29,42].

Remarkably, two of the three spacers carried by SPV2 target SPV1, whereas only one of the three spacers carried by SPV1 targets SPV2, with another one targeting a pRN1-like plasmid integrated in the S. tokodaii genome (BA000023, nucleotide coordinates 328508 – 335407). Importantly, the loci orthologous to the regions targeted by spacers in the viruses carrying the spacers contain either point mutations or deletions, preventing self-targeting. Notably, the SPV1 and SPV2 spacers target regions close to the mini-CRISPR arrays, although origins and consequences of this proximity remains unclear (Figure 3b). These findings prompted us to search for additional mini-CRISPR arrays in our CRISPRome dataset, resulting in 15 more candidates (Figure 4a). Three of the mini-CRISPR arrays were confirmed to be encoded within viral genomes by analysis of the previously sequenced viromes from J14 and J15 samples. All three were found in the virome contigs containing fragments of genes orthologous to those of SPV1 and SPV2 (Figure 3b). We conclude that these additional mini-CRISPR arrays are carried by minor strains of SPV-like viruses present in the population. Of the 26 spacers carried by the 15 candidate mini-CRISPR arrays, 18 were found to target different loci within the SPV1 or SPV2 genomes (Figure 4b).

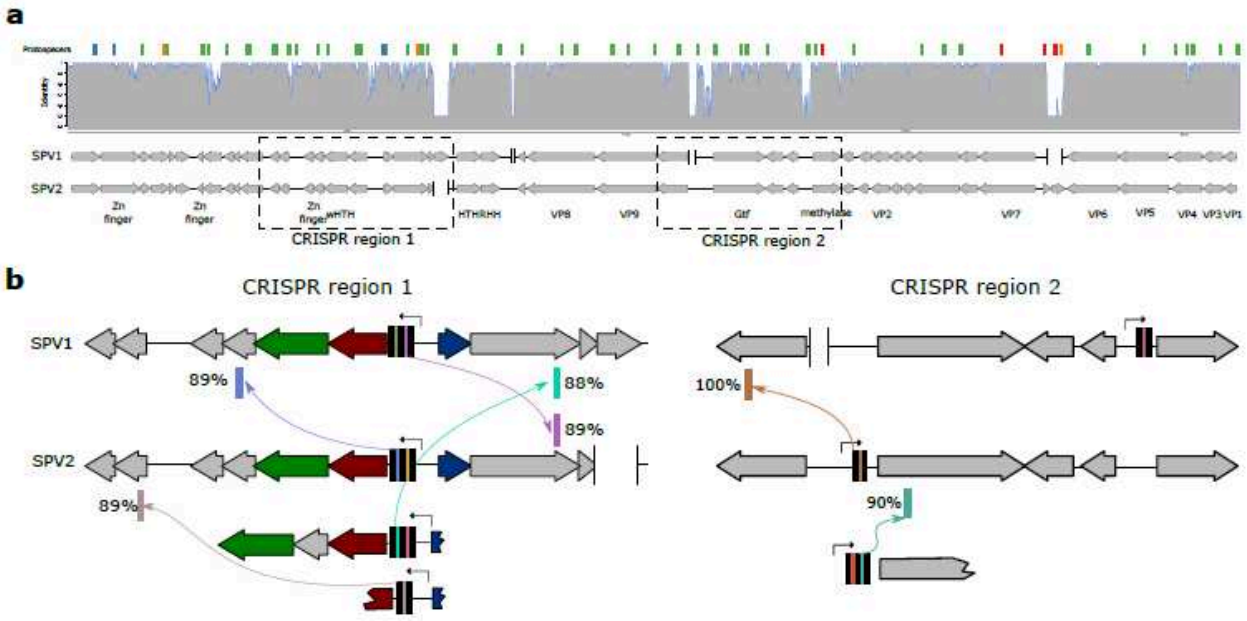
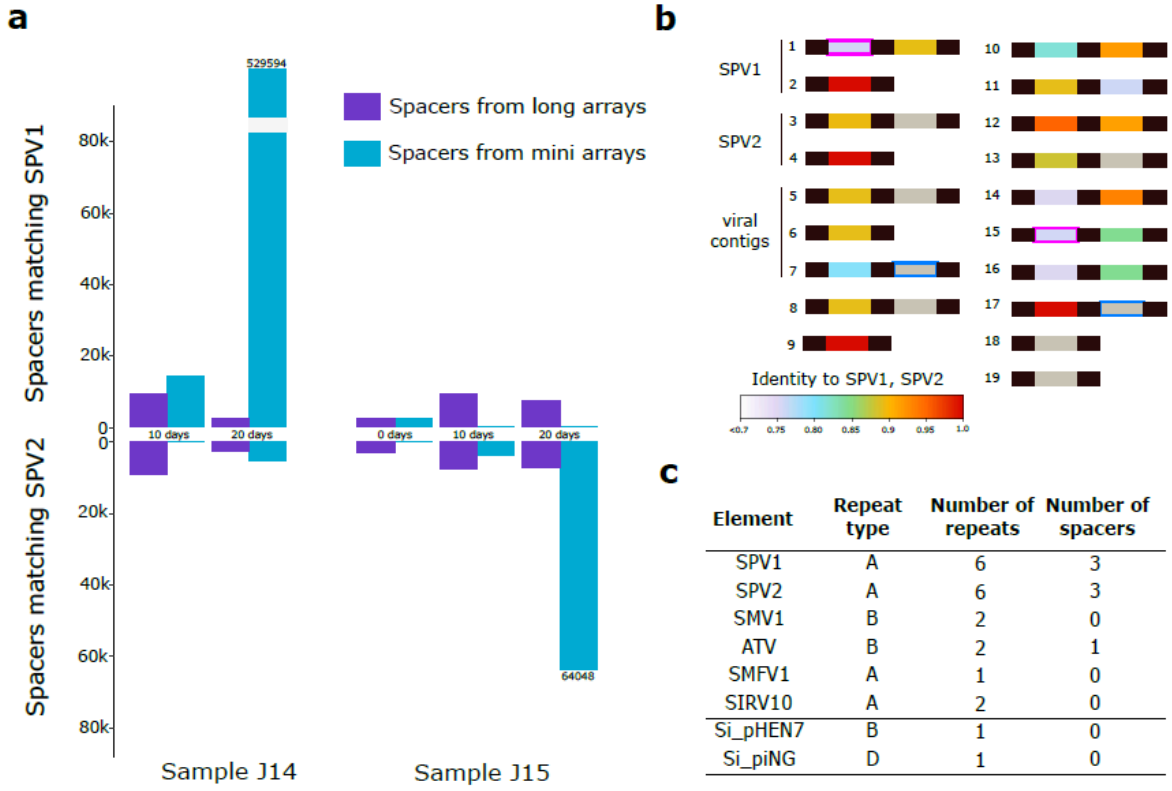


Figure 3. Mini-CRISPR arrays in SPV1 SPV2 genomes. A. Comparison of the SPV1 and SPV2 genomes. Genes are represented with arrows following the direction of transcription. Deletions in one of the two genomes with respect to the other are indicated as gaps bordered with vertical lines. Grey histogram above the genome maps shows the identity calculated in 50bp window from the SPV1-SPV2 nucleotide alignment. Locations of protospacers are showed as colored bars at the top of the figure. The regions zoomed-in in panel B are boxed. B. Zoom-in on two regions of the SPV1 and SPV2 genomes carrying mini-CRISPR arrays (CRISPR region 1 and CRISPR region 2). Black bars represent CRISPR repeat. The predicted promoters in the leader sequences are indicated with broken arrows. Positions of hits of spacers from mini-CRISPR arrays carried by SPV-like viruses are shown with colored bars and arrows link the spacers and the corresponding protospacers. Identities between spacer and protospacers are indicated next to the protospacer bars. Three mini-CRISPR arrays found in the virome data are shown below the corresponding regions of alignment.

The reciprocal CRISPR targeting by SPV1 and SPV2 strongly suggests that the virus-encoded mini-CRISPR arrays are involved in interviral conflicts and represent a distinct mechanism of superinfection exclusion, whereby a cell infected by one virus becomes resistant to the other virus. This possibility is consistent with the observation that J14 and J15 cultures contain exclusively SPV2 and SPV1, respectively (Supplementary Figure 1). Furthermore, we hypothesize that avoidance of self-targeting promotes speciation in the portoglobovirus population. In a similar fashion, it has been recently suggested that CRISPR spacers acquired during inter-species mating of halophilic archaea also influence speciation[43]. Importantly, SPV1 is a non-lytic virus, which establishes a chronic infection and is released without killing its host[44]. Therefore, the association between a non-lytic, CRISPR-bearing virus and the host is beneficial to both parties and can thus be considered a form of symbiosis.

To assess the generality of the potential CRISPR-mediated virus-virus interactions, we searched if any of the other available genomes of Sulfolobales viruses and plasmids carry CRISPR repeats of the four types. A mini-CRISPR array has been also identified in the genome of Acidianus two-tailed virus (family Bicaudaviridae). It consisted of a single spacer flanked by type B CRISPR repeats. In addition, stand-alone CRISPR repeats similar to those of the corresponding host species, albeit without spacers, were identified in the genomes of 3 other viruses and 2 conjugative plasmids (Figure 4c). However, the stand-alone repeats were not preceded by recognizable leader sequences. Whether such repeats are competent targets for spacer integration is thus unclear.



c

Element	Repeat type	Number of repeats	Number of spacers
SPV1	A	6	3
SPV2	A	6	3
SMV1	B	2	0
ATV	B	2	1
SMFV1	A	1	0
SIRV10	A	2	0
Si_pHEN7	B	1	0
Si_piNG	D	1	0

Figure 4. Mini-CRISPR arrays in viral genomes. **A.** Total abundance of SPV1 and SPV2 marching spacers from long and mini-arrays. **B**. mini-CRISPR arrays predicted from the CRISPRome data. Identity of spacers to SPV1 or SPV2 genomes is color-coded with the scale provided at the bottom of the figure. **C.** Mini-CRISPR arrays and standalone repeats in Sulfolobales viruses and plasmids.

To assess the effects of virus-mediated versus host-mediated CRISPR immunity against SPV1 and SPV2, we compared the number and abundance of the targeting spacers originating from the long CRISPR arrays (i.e., host-borne) and mini-CRISPR arrays (Figure 4a). In the initial environmental sample J15 and in 10-days enrichments, spacers from the long arrays were the major contributors to the total immunity against SPV1 and SPV2 viruses. However, after 20 days, the abundance of spacers from mini-arrays increased dramatically, whereas the number of spacers from the long arrays was decreased, possibly due to the predation of the host by SPV1 and SPV2. Moreover, spacers from the host arrays targeted SPV1 and SPV2 indiscriminately (judging from the identity between spacers and protospacers), whereas spacers from mini-arrays are specific against either SPV1 or SPV2.

Collectively, our results demonstrate the utility of the CRISPRome for understanding virus-host interactions and reveal a potential strategy used by viruses to restrict competing MGE via CRISPR-mediated superinfection exclusion. A recent, independent comparative genomic analysis of bacterial and archaeal viruses has demonstrated the presence of CRISPR mini-arrays and single-repeat units in many bacteriophage and prophage genomes as well as a few archaeal viruses including Acidianus two-tailed virus[45]. Some of the spacers in the identified mini-arrays targeted adjacent genes in closely related virus genomes but not the mini-array-carrying virus itself, in full agreement with the pattern identified in the present work. However, unlike the case of SPV1 and SPV2 described here, most of the phage mini-arrays lack the leader regions, suggesting that they might acquire spacers via recombination with host arrays rather than canonical adaptation. Interviral conflicts via virus-targeting mini-arrays and other similar strategies are likely to contribute to viral genome evolution and speciation, and further validate the ‘guns for hire’ concept under which components of various defense and counter-defense systems are commonly exchanged between viruses and their cellular hosts.

Methods

Description of samples

The enrichment cultures established from two environmental samples, J14 and J15, were dominated by members of the genus Sulfolobus (85% in J14 and 79% in J15), unclassified members of the family Sulfolobaceae (14% in J14 and 20% in J15), and genera Sulfurisphaera and Acidianus (1% in both samples)[35]. The viral component of the enrichments included populations of seven viruses belonging to four different families[35]. We were able to perform PCR amplification of the CRISPR spacers with the DNA extracted directly from the J15 sample, but not from the J14 sample, possibly, due to the insufficient biomass in the latter. The cultures propagated in the Sulfolobus-favoring medium displayed efficient growth of the biomass, whereas those propagated in the Acidianus-favoring medium grew poorly and were discontinued after 10 days of incubation. Thus, in total, we analyzed one environmental sample and five enrichments: two 10-days enrichments and two 20-days enrichments in Sulfolobus-favoring medium, and one 10-days enrichment in the Acidianus-favoring medium.

CRISPRome amplification

To amplify CRISPR arrays of Sulfolobales from total DNA samples, six pairs of primers (Supplementary table 3) were designed. Two pairs of primers, C1 and C2, were designed to cover the diversity of the type C CRISPR repeats. Amplification reactions were carried out with DreamTaq DNA polymerase (Thermo Fisher Scientific, UK) under the following conditions: initial denaturation for 5 min at 95°C, followed by 40 cycles of 30 s at 95°C, 30s at 42-53°C (depending on the Tm of specific primers), and 60s at 72°C, and a final extension at 72°C for additional 2 min. For each DNA sample with each primer pair, five individual PCR reactions were set up. No amplification was obtained with the primer pair G1. After amplification, individual reactions were pooled and processed jointly. Amplicons were visualized on 1% ethidium bromide stained agarose gels and DNA fragments of 300–1000 bp in length were purified from the gel and sequenced on MiSeq (Illumina) with paired-end 250-bp read lengths (Genomics Platform, Institut Pasteur, France).

Spacer extraction and clustering

Spacer sequences were extracted using spget program (https://github.com/zzaheridor/spget). Spget identifies degenerate sequences of CRISPR repeats and PCR primers, and extracts all sequences between them. To account for possible sequencing mistakes and natural CRISPR repeat diversity, additional 2-5 mismatches were allowed in repeat and primer sequences. Based on expected spacer lengths, extracted spacers shorter than 25 nt or longer than 60 nt were filtered out. An additional quality filter was applied – only spacers with all nucleotides sequenced with the Phred score value higher than 20 were used for further analysis. The described filtering resulted in the removal of ~25% of all spacers.

The clustering was performed by UCLUST program [46], with 85% identity threshold and zero penalties for end gaps. UCLUST algorithm was also used with 85% identity threshold to find common spacers for different sets. The coverage of spacer diversity was estimated with Good's criterion: C = 1 – (N/total number of clusters), where N is the number of sequences that occurred only once or twice. The alpha diversity (Shannon entropy) and Chao estimate of coverage were calculated using the R package "vegan" (https://cran.r-project.org/web/packages/vegan/index.html). The spacer sequences are available in Supplementary Data 1.

Reconstruction of CRISPR arrays

The procedure of CRISPR array reconstruction uses pairs of neighboring spacers obtained from NGS reads. All pairs for the sample are joined into a directed graph, where each node represents a spacer, edges connect the spacers that appeared together in a pair, and the number of found pairs in NGS reads is used as an edge weight. The PCR amplification procedure could possibly lead to the emergence of

chimeric pairs, when two independent spacers from different CRISPR arrays are falsely connected into a pair. For example, when an amplification product from one cycle (a primer-spacer-primer unit) is used as a long primer with 5' overhang for the next cycle. Assuming that chimeric pairs are rare PCR artifacts, we filtered edges in our graph based on the weight. For each edge (pair of neighboring spacers), we calculate the sum weight of outgoing edges from the first spacer in the pair and the sum weight of incoming edges for the second spacer in the pair. If the weight of tested edge was lower than 5% of the calculated sums, the tested edge was removed (see Supplementary figure 6A). One example of reconstructed graph is shown in Supplementary Figure 7. Several arrays share the same terminal, leader-distant spacer, some of the arrays are branching towards the leader end. The script for reconstruction of the CRISPR array graphs is implemented in R language (https://www.R-project.org/)[47]. Because this approach is not suitable for identification of mini-CRISPR arrays, we used the eccentricity metrics (the length of the longest path, which is going through the selected node). The eccentricity number shows the length of the longest CRISPR array, which can be reconstructed using selected spacer (see Supplementary figure 6B).

Protospacer analysis

Protospacers were searched for with BLASTN[48] (word size 8, e-value < 0.01) in local databases of Sulfolobales viruses, plasmids and cellular genomes. PAMs were identified by aligning flanking sequences of protospacers. For Figure 2A, chi-square test followed by Bonferroni correction was used to test the specificity of spacers associated with different CRISPR-Cas repeat types to different sources of protospacers (Sulfolobales host genomes, viruses or plasmids) based on total number of spacers for each CRISPR repeat type and total number of hits to a particular source.

Loss of minor spacers during cultivation

The error bars indicate the confidence interval for the proportion of lost spacers calculated as conf = z(0.975)*sqrt(lost*(1-lost)/N), where 'lost' is a fraction of lost spacers and N is the total number of spacers in each group.

Assembly of viral contigs from spacers

To reconstruct the viral contigs, we performed "all spacers against all spacers" BLASTN (word size 8, identity > 0.7). Then a graph of spacers was built, where spacers are connected if they were matched by BLAST search. The graph was decomposed, spacers from the largest subgraphs were aligned with MUSCLE[49], and the alignment was manually corrected.

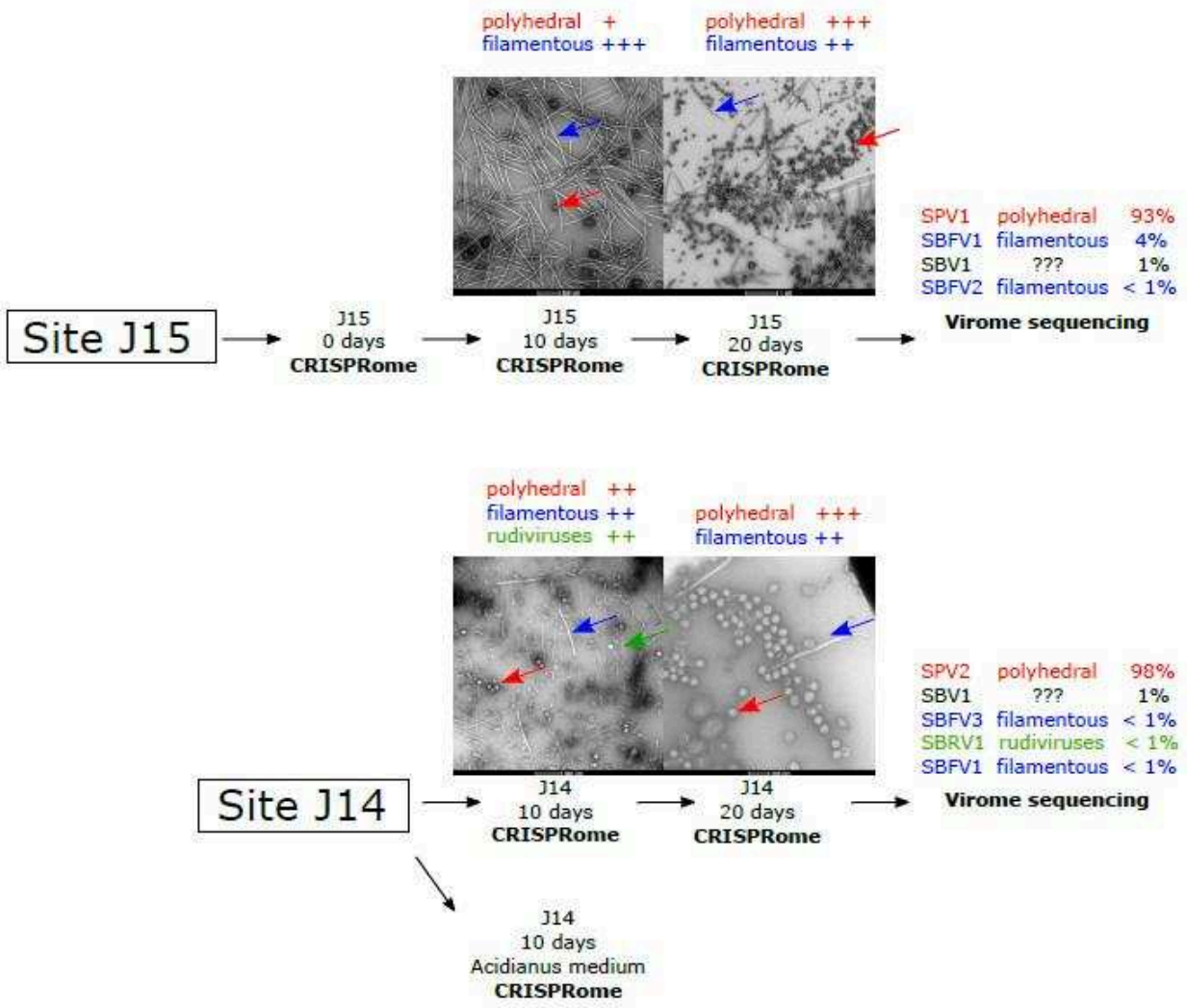
Prediction of mini-CRISPR arrays in the CRISPRome data

First, we calculated the frequency of sequencing reads with two spacers in each sample and estimate the probability (p) of spacer to be in the pair (~0.5, for J15 sample, 0 days enrichment). Assuming that all spacers are independent, we calculated the probability to observe no pairs for the spacer, which was sequenced N times: P-value = $(1-p)^N$. For the spacer sequenced 100 times, P-value was < 0.01, so we defined a threshold N for the mini-CRISPR array candidates. Similar approach was used for mini-CRISPR arrays with two spacers: the probability for a spacer to appear as first spacer in the pair is 0.5. If the spacer was sequenced in the pair 42 times, the estimated P-value to observe spacer only as a first spacer in the pair is P-value = $0.5^{42} < 0.01$.

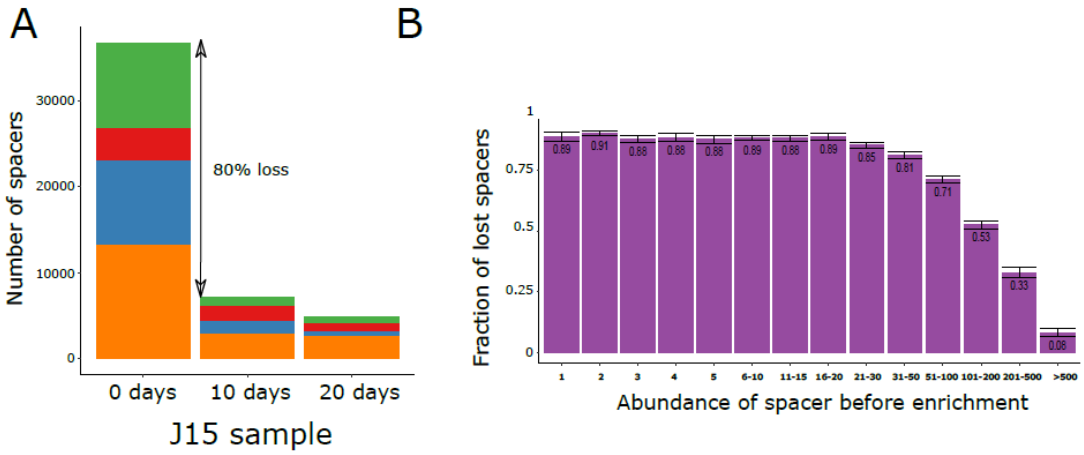
Determination of the integration sites

The precise boundaries of integration were defined based on the presence of direct repeats corresponding to attachment sites or target site duplications. The direct and inverted repeats were searched for using Unipro UGENE[50].

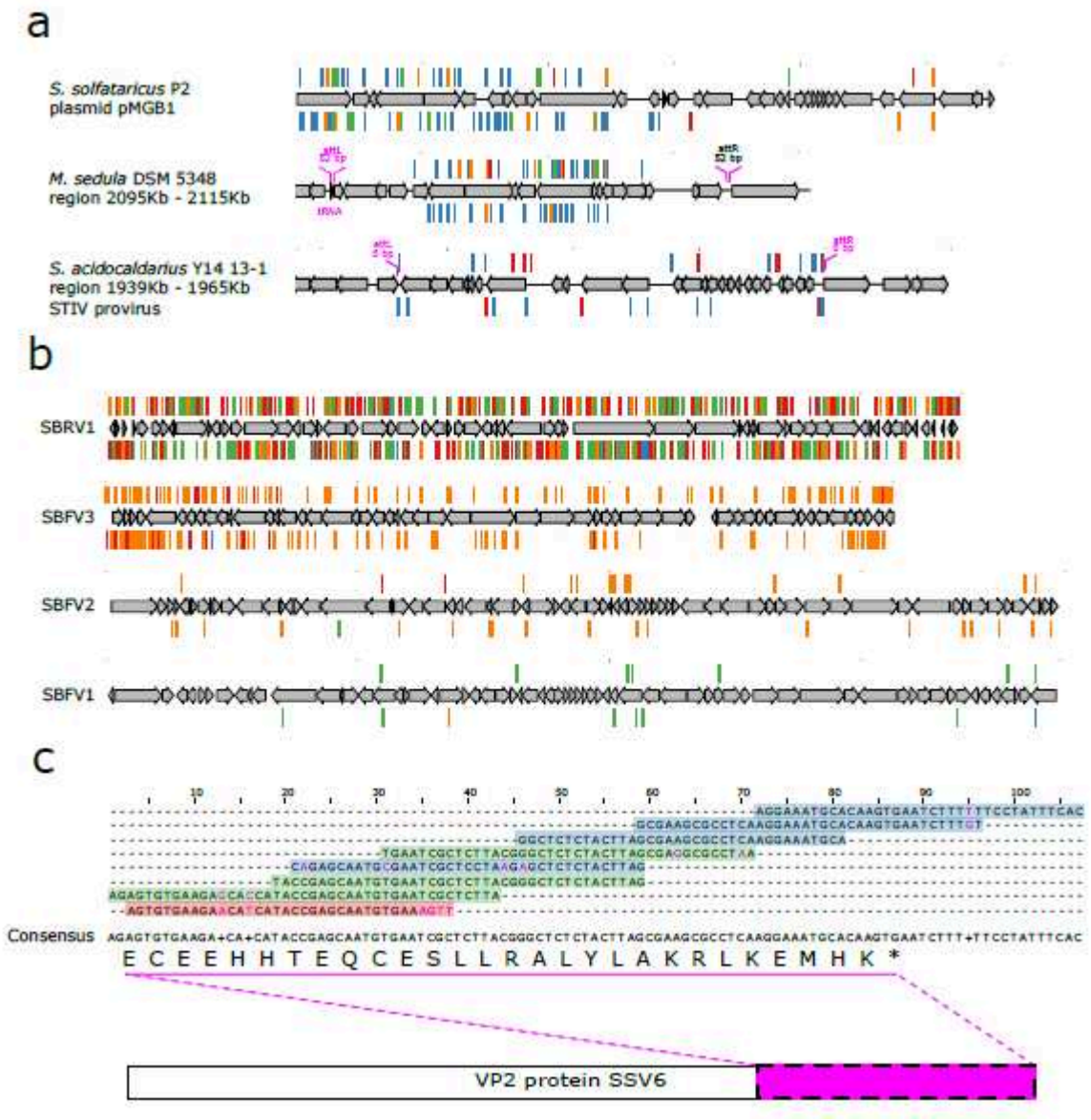
SUPPLEMENTARY FIGURES



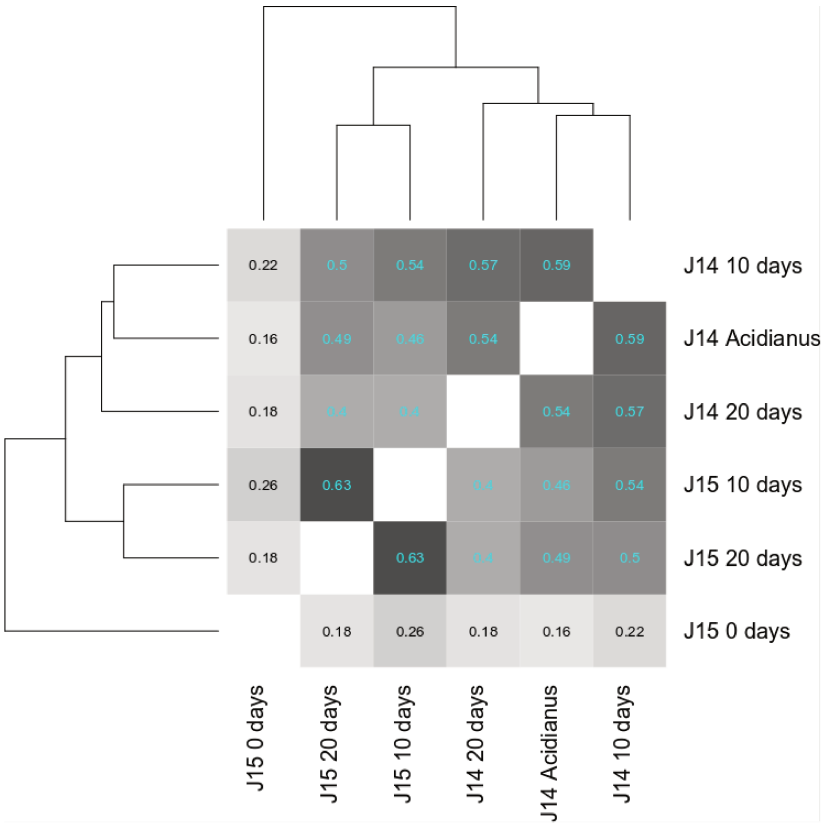
Supplementary Figure 1. Description of samples. Schematic representation of analysed samples. When availiable, images of virus diversity in enrichments are shown. Viruses belonged to different families are highlighted with arrows (polyhedral – red, filamentous – blue, rudiviruses – green).



Supplementary Figure 2. Loss of spacers during cultivation. A. Number of spacers associated with four CRISPR repeat types are shown as a barplot for J15 sample **B.** Fraction of spacers lost in all enrichment cultures is shown for groups of spacers with different abundances. Errorbars show confidence intervals for the proportion.



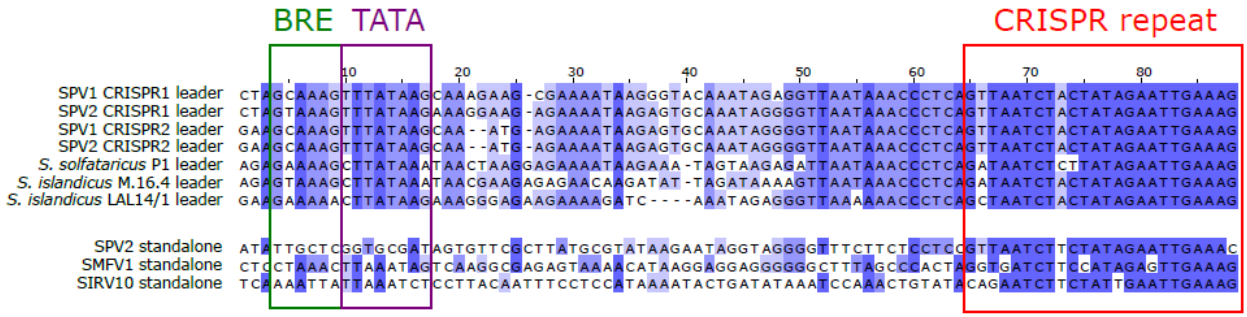
Supplementary Figure 3. A. Maps of several Sulfolobales integrated and extrachromosomal elements targeted by the Beppu CRISPR spacers. Protospacers are shown as thin bars above and below the genes (represented by grey arrows) depending on the targeted strand. The color of spacer bars corresponds to types of the CRISPR repeats. Identified attachment sites (attL and attR) for the integrated elements are shown in pink. The visualization is created by R packacge Gviz[51] **B.** Genome maps of several Sulfolobales viruses, targeted by Beppu CRISPR spacers. **C.** An example of viral contig reconstruction by overlapping spacer sequences. The color of spacers in alignment corresponds to CRISPR repeat type. Not conserved positions in the alignment are highlighted by pink color. Consensus nucleotide sequence and protein translation are shown below the alignment.



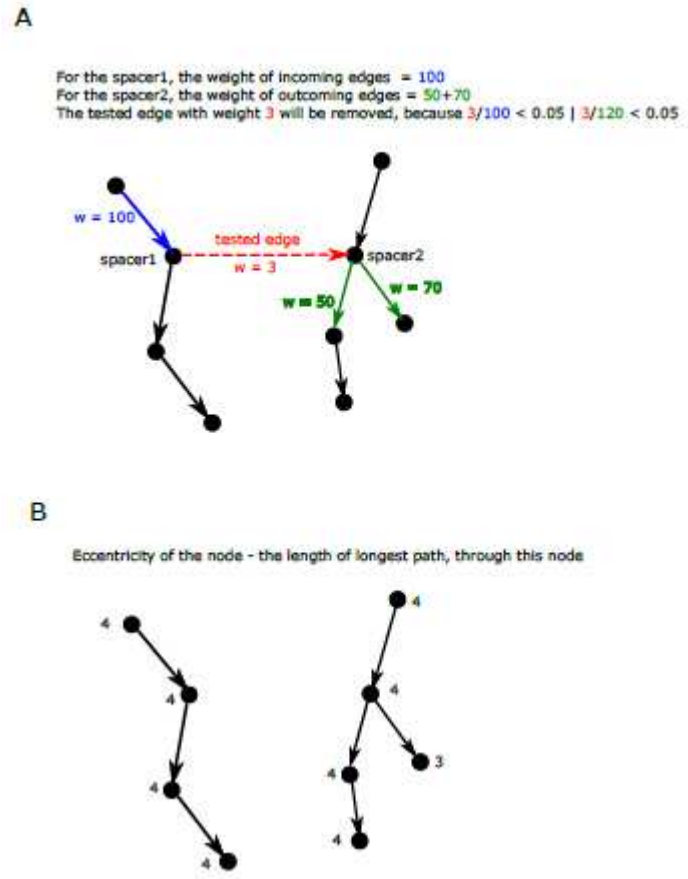
405

406 **Supplementary Figure 4.** Fraction of spacers shared between samples.

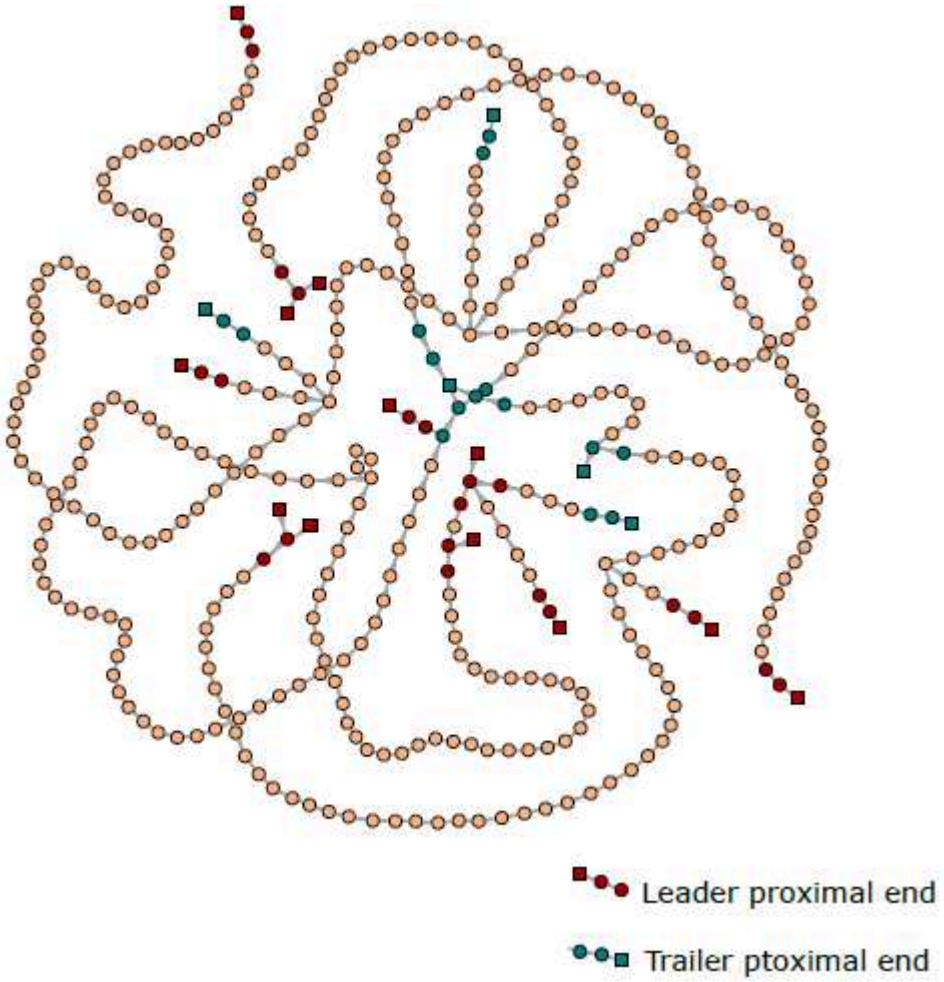
407



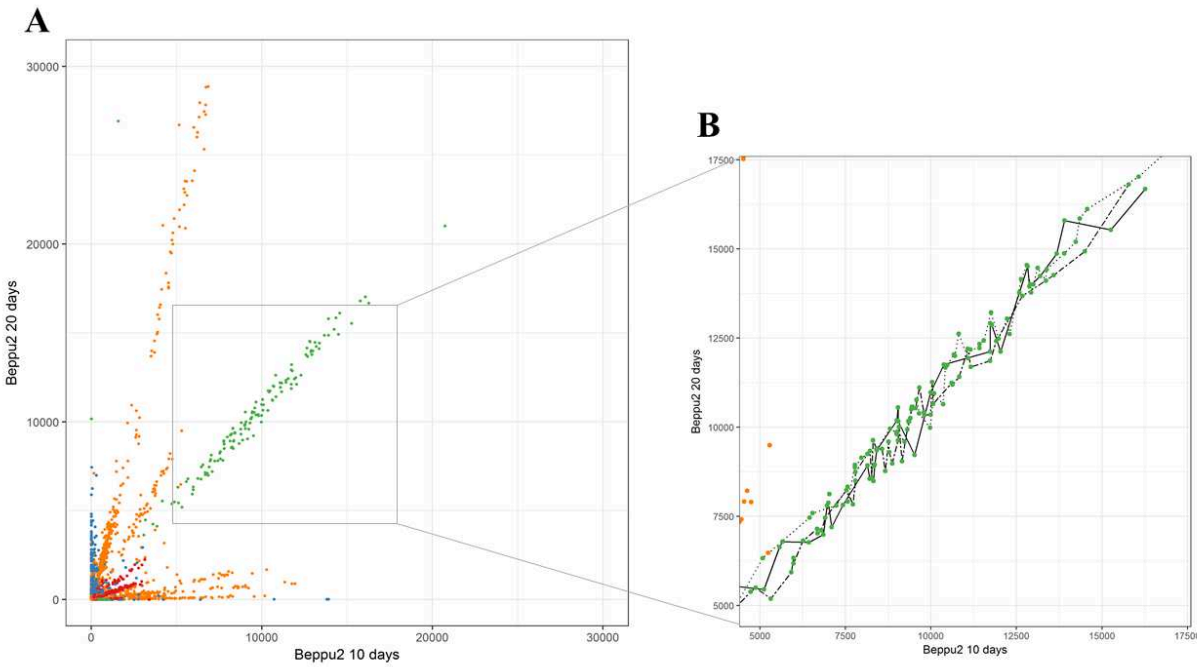
Supplementary Figure 5. Alignment of the loci including the leader sequences and CRISPR repeats associated with the Sulfolobus CRISPR arrays and virus-borne mini-CRISPR arrays. The bottom 3 sequences correspond to stand-alone CRISPR repeats. BRE and TATA elements found in the promoters of the leader sequences and the CRISPR repeats are boxed.



Supplementary Figure 6. Methods of reconstruction of long CRISPR arrays. **A.** Filtration of CRISPR array graph, by removing low abundant edges. **B.** Example of eccentricity calculation



Supplementary Figure 7. An example of reconstructed CRISPR array graph. Three spacers in leader-proximal or trailer-proximal ends are highlighted with red and green respectively.



Supplementary Figure 8. Spacers with linearly changed frequencies in 10 and 20 days enrichments of J14 sample. Dashed, dotted and solid lines represent three independent components of CRISPR arrays graph.

SUPPLEMENTARY TABLES

Supplementary table 1. Distribution of CRISPR repeat types in Sulfolobales genomes.

CRISPR repeat type	Sulfolobales genomes
A	Metallosphaera , Acidianus, Sulfolobus
B	Acidianus, Sulfolobus
C	Metallosphaera
D	Metallosphaera, Acidianus, Sulfolobus

Supplementary table 2. Diversity and coverage estimations for Beppu spacer sets.

Sample	CRISPR spacers, total	Clusters	Good's criterion	Schao	alpha diversity (Shannon)
J15 – 0 days	2 971 721	33 991	0.91	36 068 (+-97)	9.18
J15 – 10 days	5 166 123	6 155	0.88	9 431 (+-439)	7.12
J15 – 20 days	4 787 974	4 462	0.76	5 868 (+-115)	5.80
J14 – 10 days	4 129 915	6 825	0.91	8 756 (+-251)	7.16
J14 – 20 days	4 540 454	4 585	0.84	6 573 (+-229)	6.56
J14 "Acidianus" – 10 days	3 234 790	4 020	0.87	4 332 (+-35)	5.96

Supplementary table 3. Primers sequences for amplification of CRISPR arrays of Sulfolobales.

Repeat	Forward primer (5'-3')	Reverse primer (5'-3')
G1	TCGTCGGCAGCGTCAGATGTGTATAAGAG ACAGCTTTTCTCTTATGAGACTAGTAC	GTCTCGTGGGCTCGGAGATGTGTATAAGA GACAGCTAGTCTCATAAGAGAAAAGTAAT
A	TCGTCGGCAGCGTCAGATGTGTATAAGAG ACAGTAATCTACTATAGARTTGAAAG	GTCTCGTGGGCTCGGAGATGTGTATAAGA GACAGTTCAAYTCTATAGTAGATTADC
B	TCGTCGGCAGCGTCAGATGTGTATAAGAG ACAGAAYAACGAMAAGAAACTAAAAC	GTCTCGTGGGCTCGGAGATGTGTATAAGA GACAGTTTAGTTTCTTKTCGTTRTTAC
C1	TCGTCGGCAGCGTCAGATGTGTATAAGAG ACAGAACCCTCAAAGGATCACTACAA	GTCTCGTGGGCTCGGAGATGTGTATAAGA GACAGGTGATCCTTTGAGGGTTTGAAAC
C2	TCGTCGGCAGCGTCAGATGTGTATAAGAG ACAGGWGATCCTTMGAGGGTTTGAAAC	GTCTCGTGGGCTCGGAGATGTGTATAAGA GACAGACCCTCKAAGGATCWCTACAAAC
D	TCGTCGGCAGCGTCAGATGTGTATAAGAG ACAGTKAATCCYAAAAGGRATTGAAAG	GTCTCGTGGGCTCGGAGATGTGTATAAGA GACAGTTCAATYCCTTTTRGGATTMATC

Adaptor sequences are underlined.

Supplementary table 4. Identified integrated elements in Sulfolobales genomes.

Acc. number	Strain	Start	End	Size, bp	# spacers	Element type
CP000682.1	M. sedula DSM 5348	2096383	2112116	15734	58	cryptic (inactivated)
BA000023.2	S. tokodaii str. 7	~262600	~274200	11500	20	cryptic (inactivated)
BA000023.2	S. tokodaii str. 7	1310850	1355729	44880	93	conjugative
CP001399.1	S. islandicus L.S.2.15	1858852	1900333	41482	23	conjugative (inactivated)
CP001401.1	S. islandicus M.16.27	1437439	1481760	44322	20	conjugative
CP001402.1	S. islandicus M.16.4	1474356	1512307	37952	25	conjugative
CP001403.1	S. islandicus Y.G.57.14	1465198	1505472	40275	23	conjugative
CP001731.1	S. islandicus L.D.8.5	1323689	1390124	66436	27	conjugative
CP020362.1	S. acidocaldarius Y14 16-22	395173	437039	41867	20	conjugative (inactivated)
CP020362.1	S. acidocaldarius Y14 16-22	1991521	2008456	16936	27	provirus (STIV-like)
CP020363.1	S. acidocaldarius Y14 13-1	1943014	1959949	16936	27	provirus (STIV-like)

SUPPLEMENTARY TEXT

Temporal CRISPR spacer dynamics in the enrichment cultures

In the original environmental J15 sample, most spacers display similar abundances. However, after 10 days of cultivation, the community has visibly stratified into well-defined groups, each characterized by a specific frequency of spacers and likely representing a discrete strain. After 20 days of cultivation, the gap between the groups of high-abundance (≤10,000 coverage) and low-abundance (≤10 coverage) spacers increased for CRISPR types A, B and C. The moderate abundance spacers (10-1000 coverage) have largely disappeared, especially, in the case of A-type and C-type repeats, suggesting that the population became dominated by a handful of strains. The situation was different for the spacers associated with the D-type CRISPR repeats: the 4 dominant groups of populations grew in abundance and spawned a small group of extremely abundant spacers (>10,000 coverage).

A different pattern was observed with the J14 sample, where we could compare the enrichment cultures of 10 and 20 days. Whereas the population structures for the B-type repeats followed the same course as in the J15 samples, the populations bearing the D-type repeats segregated to the high-abundance and low-abundance groups. By contrast, populations with the A-type repeats showed an increase in moderate abundance spacers (opposite to the situation in the J15 sample), whereas those with the C-type repeats evolved towards collapse, with the majority of the strains displaying very low abundance.

Assembly of viral contigs and CRISPR arrays

Although this approach was complicated by short (30-36 bp) spacer lengths and inherent absence of spacers from genomic regions devoid of the protospacer adjacent motifs (PAM), we were able to reconstruct contigs of up to 200 nucleotides (Supplementary Figure 3c). Following the in silico translation, matches to viral proteins were identified, as in the example shown in Figure 3E, where the reconstructed contig encodes the fusellovirus structural protein VP2 of fuselloviruses. The reconstruction of the viral contigs from the CRISPRome data is conceptually similar to the reconstruction of plant virus genomes from small interfering RNA sequences[52].

Approximately 50% of our HTS sequencing reads include not solitary spacers but small fragments of CRISPR arrays with 2 or, less frequently, 3 spacers. The assembly of these fragments through identical spacers, theoretically, should allow reconstruction of longer CRISPR arrays. In practice, however, the spacer diversity of natural Sulfolobales population can only be represented as a graph (Supplementary Figure 7), which, in some cases, cannot be resolved into separate CRISPR arrays, due to intrinsic variations, such as deletion of spacers in the trailer end of CRISPR arrays or acquisition of new spacers at the leader end. To overcome this problem, we introduce the eccentricity metrics. The eccentricity of a spacer is the length of the longest CRISPR array, which can be reconstructed with this spacer (Supplementary Figure 6B). The longest CRISPR arrays (the maximal eccentricity) were 131, 66, 139 and 119 for spacers associated with the A-, B-, C- and D-type repeats, respectively. These length estimates agree with the average lengths of arrays in sequenced Sulfolobales isolates. The eccentricity > 3 was observed for 38% of all spacers associated with the A-type CRISPR repeats and 98% of spacers with abundances > 100. Each Sulfolobales genome usually contains more than one CRISPR array with the same CRISPR repeat sequences. We observed groups of spacers from 3 independent graph components with linearly correlated frequencies in two samples (Supplementary Figure 8), which is consistent with them being sequenced from the same genome.

Detection of integrated MGE by spacer matching

Archaeal viruses and plasmids are known to integrate into the genomes of their hosts. For many of these integrated MGE, closely related extrachromosomal relatives are not known, making their identification

cumbersome. Mapping the CRISPR spacers against the Sulfolobales chromosomes provides an efficient approach to identify integrated MGEs, both related to known plasmids and viruses as well as novel and even deteriorating ones. A threshold of 3 protospacers per kb of genomic DNA was found to be a reliable predictor for the presence of integrated MGEs. Using this approach, we predicted 11 MGEs integrated in 9 Sulfolobales genomes and subsequently validated the precise integration sites for all but one element (Supplementary Figure 3A; Supplementary table 4). These integrated MGE included 2 STIV-like proviruses, 7 integrated pNOB-like conjugative plasmids and 2 cryptic integrated elements. Some of the elements were apparently inactivated by transposon insertions and are unlikely to be mobile. Collectively, these integrated MGEs are targeted by 336 distinct spacers from our collection.

REFERENCES

1 De Sordi, L., Lourenco, M. & Debarbieux, L. The Battle Within: Interactions of Bacteriophages and Bacteria in the Gastrointestinal Tract. *Cell Host Microbe* **25**, 210-218 (2019).

2 Breitbart, M., Bonnain, C., Malki, K. & Sawaya, N. A. Phage puppet masters of the marine microbial realm. *Nat Microbiol* **3**, 754-766 (2018).

3 Fernandez, L., Rodriguez, A. & Garcia, P. Phage or foe: an insight into the impact of viral predation on microbial communities. *ISME J* **12**, 1171-1179 (2018).

4 Koonin, E. V. & Krupovic, M. The depths of virus exaptation. *Curr Opin Virol* **31**, 1-8 (2018).

5 Forterre, P. & Prangishvili, D. The great billion-year war between ribosome- and capsid-encoding organisms (cells and viruses) as the major source of evolutionary novelties. *Ann N Y Acad Sci* **1178**, 65-77 (2009).

6 Koonin, E. V. & Dolja, V. V. A virocentric perspective on the evolution of life. *Curr Opin Virol* **3**, 546-57 (2013).

7 Chow, C. E. & Suttle, C. A. Biogeography of Viruses in the Sea. *Annu Rev Virol* **2**, 41-66 (2015).

8 Makarova, K. S. *et al.* An updated evolutionary classification of CRISPR-Cas systems. *Nat Rev Microbiol* **13**, 722-36 (2015).

9 Doron, S. *et al.* Systematic discovery of antiphage defense systems in the microbial pangenome. *Science* **359** (2018).

10 Maxwell, K. L. The Anti-CRISPR Story: A Battle for Survival. *Mol Cell* **68**, 8-14 (2017).

11 Borges, A. L., Davidson, A. R. & Bondy-Denomy, J. The Discovery, Mechanisms, and Evolutionary Impact of Anti-CRISPRs. *Annu Rev Virol* **4**, 37-59 (2017).

12 van Houte, S., Buckling, A. & Westra, E. R. Evolutionary Ecology of Prokaryotic Immune Mechanisms. *Microbiol Mol Biol Rev* **80**, 745-63 (2016).

13 Samson, J. E., Magadan, A. H., Sabri, M. & Moineau, S. Revenge of the phages: defeating bacterial defences. *Nat Rev Microbiol* **11**, 675-87 (2013).

14 Koonin, E. V., Makarova, K. S. & Wolf, Y. I. Evolutionary Genomics of Defense Systems in Archaea and Bacteria. *Annu Rev Microbiol* **71**, 233-261 (2017).

15 Rostol, J. T. & Marraffini, L. (Ph)ighting Phages: How Bacteria Resist Their Parasites. *Cell Host Microbe* **25**, 184-194 (2019).

16 Koonin, E. V. & Krupovic, M. A movable defense. *Scientist* **29**, 46-53 (2015).

17 Seed, K. D., Lazinski, D. W., Calderwood, S. B. & Camilli, A. A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. *Nature* **494**, 489-91 (2013).

18 Peters, J. E., Makarova, K. S., Shmakov, S. & Koonin, E. V. Recruitment of CRISPR-Cas systems by Tn7-like transposons. *Proc Natl Acad Sci U S A* **114**, E7358-E7366 (2017).

19 Garrett, R. A. *et al.* CRISPR-based immune systems of the Sulfolobales: complexity and diversity. *Biochem Soc Trans* **39**, 51-7 (2011).

20 Prangishvili, D. *et al.* The enigmatic archaeal virosphere. *Nat Rev Microbiol* **15**, 724-739 (2017).

21 He, F. *et al.* Anti-CRISPR proteins encoded by archaeal lytic viruses inhibit subtype I-D immunity. *Nat Microbiol* **3**, 461-469 (2018).

22 Guo, T., Han, W. & She, Q. Tolerance of Sulfolobus SMV1 virus to the immunity of I-A and III-B CRISPR-Cas systems in Sulfolobus islandicus. *RNA Biol*, 1-8 (2018).

23 Athukoralage, J. S., Rouillon, C., Graham, S., Gruschow, S. & White, M. F. Ring nucleases deactivate type III CRISPR ribonucleases by degrading cyclic oligoadenylate. *Nature* **562**, 277-280 (2018).

24 Han, W. *et al.* A type III-B CRISPR-Cas effector complex mediating massive target DNA destruction. *Nucleic Acids Res* **45**, 1983-1993 (2017).

25 Rollie, C., Schneider, S., Brinkmann, A. S., Bolt, E. L. & White, M. F. Intrinsic sequence specificity of the Cas1 integrase directs new spacer acquisition. *Elife* **4** (2015).

26 Leon-Sobrino, C., Kot, W. P. & Garrett, R. A. Transcriptome changes in STSV2-infected Sulfolobus islandicus REY15A undergoing continuous CRISPR spacer acquisition. *Mol Microbiol* **99**, 719-28 (2016).

27 Erdmann, S., Le Moine Bauer, S. & Garrett, R. A. Inter-viral conflicts that exploit host CRISPR immune systems of Sulfolobus. *Mol Microbiol* **91**, 900-17 (2014).

28 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**, 406-17 (2015).

29 Zebec, Z., Manica, A., Zhang, J., White, M. F. & Schleper, C. CRISPR-mediated targeted mRNA degradation in the archaeon Sulfolobus solfataricus. *Nucleic Acids Res* **42**, 5280-8 (2014).

30 Held, N. L., Herrera, A., Cadillo-Quiroz, H. & Whitaker, R. J. CRISPR associated diversity within a population of Sulfolobus islandicus. *PLoS One* **5** (2010).

31 Munson-McGee, J. H. *et al.* A virus or more in (nearly) every cell: ubiquitous networks of virus-host interactions in extreme environments. *ISME J* **12**, 1706-1714 (2018).

32 Held, N. L., Herrera, A. & Whitaker, R. J. Reassortment of CRISPR repeat-spacer loci in Sulfolobus islandicus. *Environ Microbiol* **15**, 3065-76 (2013).

33 Bautista, M. A., Black, J. A., Youngblut, N. D. & Whitaker, R. J. Differentiation and Structure in Sulfolobus islandicus Rod-Shaped Virus Populations. *Viruses* **9** (2017).

34 Held, N. L. & Whitaker, R. J. Viral biogeography revealed by signatures in Sulfolobus islandicus genomes. *Environ Microbiol* **11**, 457-66 (2009).

35 Liu, Y. *et al.* New archaeal viruses discovered by metagenomic analysis of viral communities in enrichment cultures. *Environ Microbiol* **doi: 10.1111/1462-2920.14479** (2019).

36 Savitskaya, E. *et al.* Dynamics of Escherichia coli type I-E CRISPR spacers over 42 000 years. *Mol Ecol* **26**, 2019-2026 (2017).

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

38 Lintner, N. G. *et al.* Structural and functional characterization of an archaeal clustered regularly interspaced short palindromic repeat (CRISPR)-associated complex for antiviral defense (CASCADE). *J Biol Chem* **286**, 21643-56 (2011).

39 Shmakov, S. A. *et al.* The CRISPR Spacer Space Is Dominated by Sequences from Species-Specific Mobilomes. *MBio* **8** (2017).

40 Martin, A. *et al.* SAV 1, a temperate u.v.-inducible DNA virus-like particle from the archaebacterium Sulfolobus acidocaldarius isolate B12. *EMBO J* **3**, 2165-8 (1984).

41 Krupovic, M. *et al.* Integrated mobile genetic elements in Thaumarchaeota. *Environ Microbiol* (2019).

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

43 Turgeman-Grott, I. *et al.* Pervasive acquisition of CRISPR memory driven by inter-species mating of archaea can limit gene transfer and influence speciation. *Nature microbiology* **4**, 177-186 (2019).

44 Liu, Y. *et al.* A novel type of polyhedral viruses infecting hyperthermophilic archaea. *J Virol* **91**, e00589-17 (2017).

45 Faure, G. *et al.* CRISPR in mobile genetic elements: counter-defense and beyond. *Nat Rev Microbiol* **In press** (2019).

46 Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460-1 (2010).

47 Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695 (2006).

48 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403-10 (1990).

49 Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792-7 (2004).

50 Okonechnikov, K., Golosova, O. & Fursov, M. Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics* **28**, 1166-7 (2012).

51 Hahne, F. & Ivanek, R. Visualizing Genomic Data Using Gviz and Bioconductor. *Methods Mol Biol* **1418**, 335-51 (2016).

609 52 Pooggin, M. M. Small RNA-Omics for Plant Virus Identification, Virome Reconstruction, and
610 Antiviral Defense Characterization. *Front Microbiol* **9**, 2779 (2018).
611
612

CHAPTER V

Integrated Mobile Genetic Elements in Thaumarchaeota.

Introduction:

Following the description of CRISPR mini-arrays in SPV1 and SPV2 viruses in Chapter IV, this Chapter introduces CRISPR arrays carried by mobile genetic elements integrated in the genomes of Thaumarchaeota.

Contribution: I identified insertion sequences (transposons) in thaumarchaeal genomes (Figures 1A, 1C, 7) and analyzed spacer diversity in thaumarchaeal genomes and thaumarchaeal integrated elements (section “iMGE-encoded CRISPR arrays”).

environmental microbiology

Environmental Microbiology (2019) **00**(00), 00–00 doi:10.1111/1462-2920.14564

Integrated mobile genetic elements in Thaumarchaeota

Mart Krupovic,[1*] Kira S. Makarova,[2] Yuri I. Wolf,[2] Sofia Medvedeva,[1,3,4] David Prangishvili,[1] Patrick Forterre[1,5] and Eugene V. Koonin[2]

[1]*Institut Pasteur, Unité Biologie Moléculaire du Gène chez les Extrêmophiles, 75015 Paris, France.*

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

[3]*Center of Life Sciences, Skolkovo Institute of Science and Technology, Skolkovo, Russia.*

[4]*Sorbonne Université, Collège doctoral, 75005 Paris, France.*

[5]*Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Univ. Paris- Sud, Université Paris-Saclay, Gif-sur-Yvette cedex, Paris, France.*

Summary

To explore the diversity of mobile genetic elements (MGE) associated with archaea of the phylum Thaumarchaeota, we exploited the property of most MGE to integrate into the genomes of their hosts. Integrated MGE (iMGE) were identified in 20 thaumarchaeal genomes amounting to 2 Mbp of mobile thaumarchaeal DNA. These iMGE group into five major classes: (i) proviruses, (ii) casposons, (iii) insertion sequence-like transposons, (iv) integrative-conjugative elements and (v) cryptic integrated elements. The majority of the iMGE belong to the latter category and might represent novel families of viruses or plasmids. The identified proviruses are related to tailed viruses of the order *Caudovirales* and to tailless icosahedral viruses with the double jelly-roll capsid proteins. The thaumarchaeal iMGE are all connected within a gene sharing network, highlighting pervasive gene exchange between MGE occupying the same ecological niche. The thaumarchaeal mobilome carries multiple auxiliary metabolic genes, including multicopper oxidases and ammonia monooxygenase subunit C (AmoC), and stress response genes, such as those for universal stress response proteins (UspA). Thus, iMGE might make important contributions to the fitness and adaptation of their hosts. We identified several iMGE carrying type I-B CRISPR-Cas systems and spacers matching other thaumarchaeal iMGE, suggesting antagonistic interactions between coexisting MGE and symbiotic relationships with the ir archaeal hosts.

Received 30 November, 2018; revised 10 February, 2019; accepted 13 February, 2019. *For correspondence. E-mail krupovic@pasteur.fr; Tel. +33 1 40 61 37 22; Fax +33 1 45 68 88 34.

Introduction

Similar to bacteria and eukaryotes, archaea are associated with diverse classes of mobile genetic elements (MGE), collectively referred to as the mobilome. Based on genomic features and the mode of interaction with the host cells, the archaeal mobilome can be divided into five large classes: (i) viruses (Pietilä *et al*., 2014; Snyder *et al*., 2015; Prangishvili *et al*., 2017; Krupovic *et al*., 2018; Munson-McGee *et al*., 2018), (ii) conjugative elements (Prangishvili *et al*., 1998; Greve *et al*., 2004), (iii) small cryptic plasmids (Forterre *et al*., 2014; Wang *et al*., 2015), (iv) transposable elements closely related to bacterial insertion sequences (IS) (Filée *et al*., 2007) and (v) the more recently discovered self-synthesizing transposon-like elements called casposons which employ a homologue of the CRISPR-associated Cas1 protein as their integrase (casposase) (Krupovic *et al*., 2014; Krupovic *et al*., 2017). All five classes of MGE are also represented in bacteria, whereas eukaryotes lack conjugative elements and casposons.

Viruses infecting archaea are notoriously diverse in terms of their virion morphologies and gene contents (Pietilä *et al*., 2014; Wang *et al*., 2015; Prangishvili *et al*., 2017; Krupovic *et al*., 2018; Munson-McGee *et al*., 2018). Comparative structural and genomic studies show that the archaeal virosphere can be generally divided into two large fractions, the archaea-specific viruses and the cosmopolitan viruses (Iranzo *et al*., 2016b). The archaea-specific viruses are, by definition, unique to archaea and often display unexpected virion morphologies, such as bottle-shaped, spindle-shaped or droplet-shaped (Prangishvili *et al*., 2017). Most of these viruses are, thus far, known to infect hyperthermophiles of the phylum Crenarchaeota. Archaea-specific viruses are currently classified into 13 families that are characterized by unique gene contents that are distinct from those of viruses infecting bacteria and eukaryotes, and only minimally shared across different archaeal virus families. By contrast, the cosmopolitan fraction of the archaeal virosphere consist of viruses that display common architectural and genomic

features with viruses of bacteria and eukaryotes, and for many genes, homologues in bacterial viruses are readily detectable (Iranzo *et al*., 2016b). These include tailed dsDNA viruses representing all three major families of the order *Caudovirales* (*Myoviridae*, *Siphoviridae* and *Podoviridae*), the dominant supergroup of bacterial viruses, as well as icosahedral viruses with the double jelly-roll (DJR) and single jelly-roll (SJR) major capsid proteins (MCP) classified in the families *Turriviridae* and *Sphaerolipoviridae*, respectively (Pietilä *et al*., 2014; Prangishvili *et al*., 2017).

Representatives of all five classes of archaeal (and bacterial) MGE can integrate into the genomes of their hosts and reside as integrated MGE (iMGE). In fact, a substantial fraction of cellular genomes, across all three domains of life, consists of diverse classes of iMGE (Craig *et al*., 2015). Very often, iMGE are not merely silent passengers within the cellular genomes but can have pronounced effects on the functioning, adaptation and evolution of their host cells. In bacteria, many adaptive traits, such as various transporters, antibiotic resistance genes or toxins, are encoded by integrative-conjugative elements (ICE), pathogenicity islands and transposons which allow host bacteria to compete with other organisms for resources and colonize new ecosystems (Escudero *et al*., 2015; Johnson and Grossman, 2015; Guédon *et al*., 2017; Novick and Ram, 2017; Partridge *et al*., 2018). Indeed, pathogenicity determinants typically are carried by integrated or extrachromosomal MGE. Thus, the perception of iMGE as 'junk DNA' or 'genomic parasites' is changing to the concept of iMGE being major agents of molecular innovation and environmental adaptation of cellular organisms (Omelchenko *et al*., 2005; Frost and Koraimann, 2010; Frank and Feschotte, 2017; Jangam *et al*., 2017; Koonin and Krupovic, 2018). Typically, MGE integration leaves a molecular scar in the cellular genome which manifests as direct repeats (DR) flanking the iMGE (Grindley *et al*., 2006). In the case of integration mediated by tyrosine recombinases, the DR, known as left and right attachment sites (*attL* and *attR*), result from recombination between homologous sites present on the cellular chromosome and the MGE (Grindley *et al*., 2006). By contrast, the DR flanking transposons, as in the case of the recently described thaumarchaeal casposons (Krupovic *et al*., 2014; 2017), are referred to as target site duplication (TSD) and are generated by staggered cleavage of the target site, followed by fill-in DNA repair (Mahillon and Chandler, 1998; Béguin *et al*., 2016).

Considerable efforts have been undertaken to explore the diversity and distribution of MGE in bacterial genomes. By contrast, our understanding of the archaeal mobilome remains limited. The vast majority of archaeal viruses and plasmids have been characterized from hyperthermophiles of the phylum Crenarchaeota and halophiles of the phylum Euryarchaeota (Forterre *et al*., 2014; Pietilä *et al*., 2014; Wang *et al*., 2015; Prangishvili *et al*., 2017; Munson-McGee *et al*., 2018), whereas not a single virus or plasmid has been characterized for members of the third major phylum of cultivated archaea, the Thaumarchaeota. Thaumarchaea are among the most widely distributed archaea in the environment and are generally recognized to exert the primary control over ammonia oxidation in terrestrial, marine and geothermal habitats (Stahl and de la Torre, 2012). Due to their unusually high affinity for ammonia, this group of archaea is believed to outcompete the bacterial ammonia oxidizers in accessing ammonia and appear to determine the oxidation state of nitrogen available to associated microbial communities (Martens-Habbena *et al*., 2009). Furthermore, as autotrophs, thaumarchaea also play an important role in the fixation of inorganic carbon. For instance, in oxygenated surface deep-sea sediments, chemosynthesis largely depends on the oxidation of ammonia, with 1 mol of CO_2 fixed per 10 mol of NH_4^+ oxidized (Wuchter *et al*., 2006).

It has been demonstrated that virus-mediated turnover of thaumarchaea in surface deep-sea sediments accounts for up to one-third of the total microbial biomass killed, resulting in the release of approximately 0.3–0.5 gigatons of carbon per year globally and that turnover of thaumarchaea by viruses in the deep ocean is faster than that of bacteria (Danovaro *et al*., 2016). These findings illuminate the prominent role of thaumarchaeal viruses in the Biosphere (Danovaro *et al*., 2017). Despite the importance of thaumarchaea and their viruses in the global nitrogen and carbon cycling (Offre *et al*., 2013), only two proviruses (Krupovic *et al*., 2011; Abby *et al*., 2018) and three casposons (Krupovic *et al*., 2014; Krupovic *et al*., 2016) have been identified in the thaumarchaeal genomes. In addition, several putative thaumarchaeal virus genomes, all members of the order *Caudovirales*, have been sequenced in the course of single-cell genomics and metagenomics studies (Chow *et al*., 2015; Labonté *et al*., 2015; Ahlgren *et al*., 2019; López-Pérez *et al*., 2018), although metagenomics analyses have further hinted at an unexplored diversity of thaumarchaeal viruses (Danovaro *et al*., 2016; Roux *et al*., 2016; Vik *et al*., 2017). Furthermore, it is currently unclear whether any of the many morphologically unique viruses discovered in crenarchaea (Prangishvili *et al*., 2017) are associated with mesophilic archaea, such as thaumarchaea.

Here, we report the results of a search of the genomes of thaumarchaea isolated from diverse environments for iMGE. The identified iMGE are assigned to five classes, namely, proviruses, casposons, IS-like transposons, putative integrative-conjugative elements and cryptic integrated elements, and provide insights into the prevalence, diversity and distribution of the thaumarchaeal mobilome.

Results

iMGE detection in thaumarchaeal genomes

The genomes of 21 species representative of the taxonomic diversity and environmental distribution of the phylum Thaumarchaeota were analysed for the presence of iMGE (Supporting Information Table S1). The analysed genomes belong to four thaumarchaeal orders, namely, Cenarchaeales, Nitrosopumilales, Nitrososphaerales and *Candidatus* Nitrosocaldales, as well as four proposed unassigned genera, including *Ca.* Nitrosotalea, *Ca.* Nitrosotenuis, *Ca.* Nitrosopelagicus and *Ca.* Caldiarchaeum. The latter genus includes a single representative, *Ca.* Caldiarchaeum subterraneum, which in phylogenetic analyses forms a sister group to Thaumarchaeota and is usually assigned to a distinct archaeal phylum, the Aigarchaeota (Nunoura *et al*., 2011). However, in the GenBank database it is affiliated to the phylum Thaumarchaeota and was, thus, retained in our analysis. The analysed organisms were isolated from a wide range of environments, including a subsurface gold mine (Nunoura *et al*., 2011), thermal springs (Spang *et al*., 2012; Lebedeva *et al*., 2013; Abby *et al*., 2018; Daebeler *et al*., 2018), wastewater treatment plant (Li *et al*., 2016), marine waters (Santoro *et al*., 2015; Bayer *et al*., 2016; Ahlgren *et al*., 2017) and sediments (Park *et al*., 2014) and various soil samples (Kim *et al*., 2011; Lehtovirta-Morley *et al*., 2011; Tourna *et al*., 2011; Zhalnina *et al*., 2014; Lehtovirta-Morley *et al*., 2016; Herbold *et al*., 2017). Although most of these organisms are mesophiles, some are psychrophilic (Hallam *et al*., 2006), thermophilic (Nunoura *et al*., 2011; Spang *et al*., 2012; Lebedeva *et al*., 2013; Abby *et al*., 2018; Daebeler *et al*., 2018) or acidophilic (Lehtovirta-Morley *et al*., 2011).

We employed three different strategies to search for the iMGEs (see Materials and Methods for details). Specifically, the genomes were analysed for the presence of (i) loci enriched in ORFans and uncharacterized genes; (ii) genes encoding signature proteins typical of different archaeal MGE groups; (iii) genes encoding integrases of the tyrosine recombinase superfamily. For detailed analysis and annotation, we considered only those loci that displayed typical features of site-specific integration and/or contained signature MGE genes surrounded by additional virus- or plasmid-related genes. In total, 74 iMGEs were predicted with high confidence in 20 thaumarchaeal genomes (Supporting Information Table S2), with the number of iMGE per genome ranging from 1 to 8 (median = 3). Only one of the analysed thaumarchaeal species, *Ca.* Nitrosopelagicus brevis CN25 (Santoro *et al*., 2015), lacked identifiable iMGEs. In addition to the multigene iMGE, 20 of the 21 analysed thaumarchaeal genomes were found to contain transposons closely related to bacterial insertion sequences (IS) (Mahillon and Chandler, 1998; Filée *et al*., 2007). The number of IS-like transposons per genome varied from 0 in *Cenarchaeum symbiosum* A to 83 in *Ca.* Nitrososphaera gargensis Ga9_2 (Supporting Information Table S1). Thaumarchaea isolated from soil samples generally have larger genomes (p value = 0.093) and more iMGE per genome (p value = 0.072) than those inhabiting aquatic environments (Fig. 1A), whereas freshwater and marine thaumarchaea have similar numbers of iMGE. Consistently, *Ca.* Nitrosopelagicus brevis CN25, which does not carry identifiable iMGE, has the smallest genome (1.23 Mbp) among the sequenced thaumarchaea. Thus, the number of iMGE appears to scale close to linearly with the host genome size although, given the limited dataset, the two values show relatively weak positive correlation (R = 0.469, p value = 0.031; Fig. 1B). The number of the-more abundant IS-like transposons showed stronger correlation with the genome size (R = 0.738, p value = 0.00013; Fig. 1C). No statistically significant differences were observed in the number of iMGE or transposons between mesophiles and thermophiles.

Targets and molecular features of MGE integration

The putative *att*/TSD sites could be determined for 68 of the 74 elements (Supporting Information Table S2). Of the six iMGE for which *att*/TSD could not be confidently predicted, five are proviruses and one is a cryptic integrated element. These might be either inactivated iMGE or their recombination sites could be too short for unambiguous identification without additional sequence information from closely related strains. The DR flanking the thaumarchaeal elements were considerably shorter than those characteristic of iMGEs from other archaea. The majority of thaumarchaeal *att* sites were shorter than 26 bp (as short as 8 bp, median length of 17 bp); only seven iMGEs had *att* sites longer than 25 bp (Fig. 2A). By contrast, the *att* sites characterized for MGEs integrated in crenarchaeal genomes ranged from 29 to 69 bp (median length of 45 bp) (She *et al*., 2002). Similarly to the case of bacteria, archaeal MGEs often integrate into tRNA genes (Williams, 2002; She *et al*., 2004; Krupovic *et al*., 2010b; Béguin *et al*., 2016; Cossu *et al*., 2017; Wang *et al*., 2018a). However, other integration targets, including protein-coding genes and intergenic regions, have also been reported (Krupovic *et al*., 2010a; 2014; Shah *et al*., 2012;Anderson *et al*., 2017). Among the 68 thaumarchaeal iMGEs for which precise integration sites could be defined, 39 used tRNA genes as integration targets, 15 were found in the intergenic regions and 14 integrated into the 3′-distal regions of protein-coding genes (Supporting Information Table S2). There was no apparent relationship between the type of integration target used and the host organism or the type of iMGE.

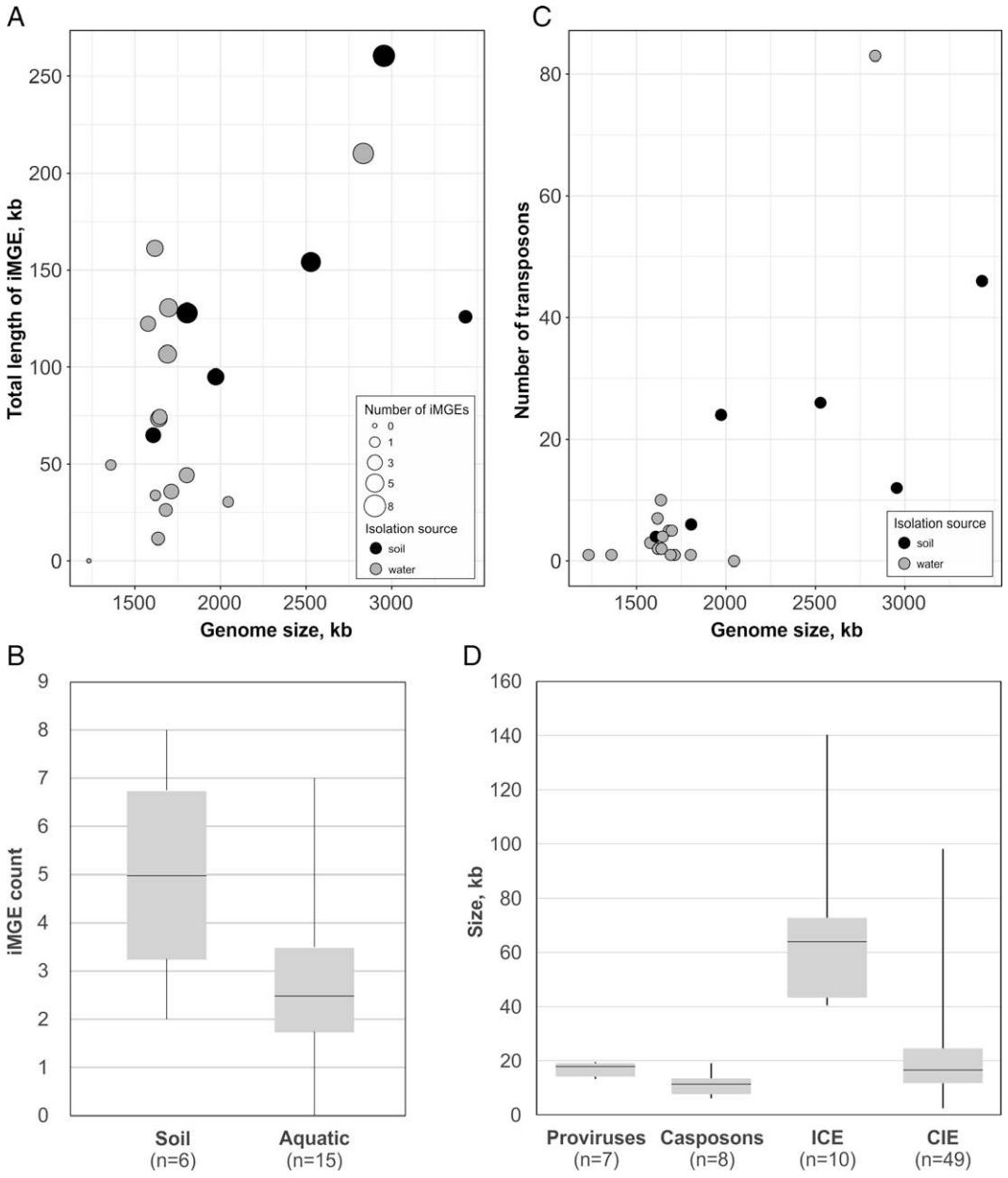


Fig. 1. Characteristics of thaumarchaeal iMGE.
A. Correspondence between the cumulative size of the iMGEs in the genome and the total genome size. Grey and black circles represent iMGEs present in the genomes of thaumarchaea isolated from aquatic and soil samples, respectively, with the diameter of the circles corresponding to the number of iMGEs per genome.
B. Box plot shows the frequency of iMGE in genomes of thaumarchaea isolated from soil and aquatic (marine and freshwater) environments.
C. Correspondence between the number of IS-like transposons in the genome and the total genome size. Grey and black circles denote the IS identified in the genomes of thaumarchaea isolated from aquatic and soil samples, respectively.
D. Box plot show size distribution in the four iMGE classes. Each box represents the middle 50th percentile of the data set and is derived using the lower and upper quartile values. The median value is displayed by a horizontal line inside the box. Whiskers represent the maximum and minimum values.

Several thaumarchaea hosted iMGEs which occupied all three types of target sites within the same genome (Supporting Information Table S2).

Integration into tRNA genes. Thirty-nine iMGE integrations (57%) were identified in genes encoding tRNAs with 22 anticodons corresponding to 14 amino acids (Supporting Information Table S2). Notably, insertions occurred within both intron-less (n = 29) and intron-containing (n = 10) tRNA genes (Fig. 2B). *Ca.* Nitrosotalea okcheonensis CS contained four different elements integrated in distinct tRNA genes, whereas in *Ca.* Nitrosotenuis

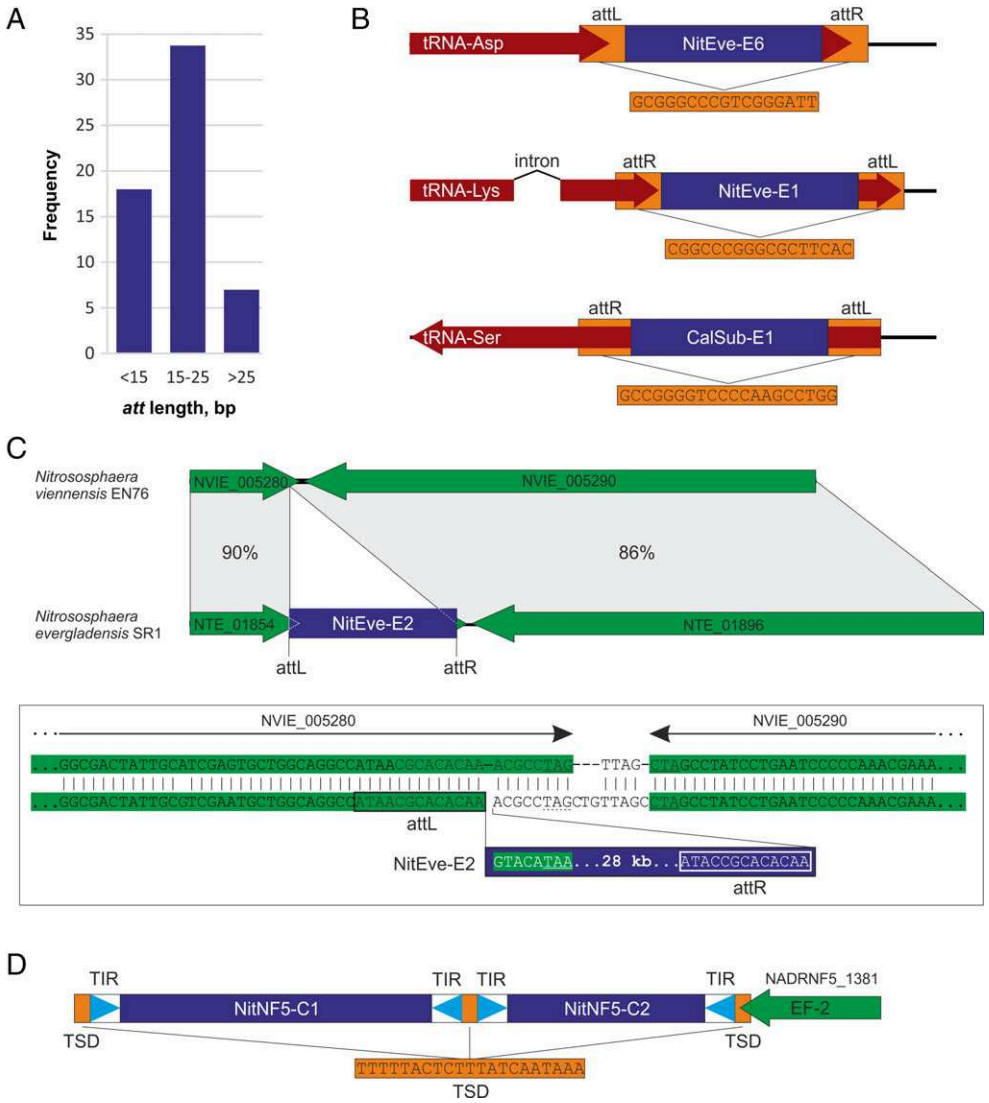


Fig. 2. Properties of site-specific MGE integration in thaumarchaea.
A. Frequency of iMGE integration in different target sites.
B. Integrations in tRNA genes. iMGE are indicated by blue rectangles; tRNA genes are shown as red arrows; attachment (att) sites are highlighted in orange.
C. Integrations in protein-coding genes. The protein coding genes are shown with green arrows, whereas the iMGE is shown as a blue rectangle. The figure compares an empty site in the genome of *Nitrososphaera viennensis* EN76 and an iMGE-occupied site in the genome of *Nitrososphaera evergladensis* SR1. The box shows a zoom-in on the corresponding integration sites in the two species. The original stop codon is underlined, whereas the one introduced by the iMGE is indicated with a broken line. Attachment sites are boxed.
D. Tandem integration of two casposons into a protein-coding gene. Terminal inverted repeats (TIR) are shown with light blue triangles, whereas target site duplications (TSD) are shown as orange rectangles.

sp. AQ6f, four tRNA genes accommodated five different elements.

In bacteria and archaea, MGEs targeting tRNA genes typically recombine with the 3′-distal region of the gene (Williams, 2002; She *et al*., 2004), whereas recombination with the 5′-distal region is considerably less frequent (Zhao and Williams, 2002; Krupovic and Bamford, 2008b; Krupovic *et al*., 2010b; Gaudin *et al*., 2014; Cossu *et al*., 2017). All but one thaumarchaeal tRNA-targeting iMGEs were found to be integrated into 3′-distal regions of the tRNA genes. However, in the genome of *Ca.* C. subterraneum, CalSub-E1 apparently recombined with the 5′-distal region of the tRNA-Ser gene (Fig. 2B).

Ns. evergladensis SR1 genome carries a curious chimeric iMGE that appears to result from integration of a smaller element, NitEve-E7, into the genome of a larger one, NitEve-E6. The latter is inserted into a tRNA gene, whereas the integration site of the former element, in the absence of sequences from closely related species, could be defined only approximately. Such piggybacking

might be particularly beneficial for MGEs that do not encode specialized devices for intercellular transfer (e.g. conjugative pili). Integration into other MGEs might ensure wider horizontal spread of such elements. This strategy of dissemination is indeed widely employed by various insertion sequences which commonly integrate into larger MGE and has also been observed for casposons in *Methanosarcina* (Krupovic *et al*., 2016). Notably, seven thaumarchaeal iMGE from four different species carry transposon insertions.

Integration into protein-coding genes. Fourteen iMGEs used protein-coding genes for integration. The genes that are exploited by the MGE as integration targets encode a Zn-finger protein conserved in different species of *Nitrososphaera* (AIF83914), AsnC family transcriptional regulator (AFU58629), dihydroxy-acid dehydratase (ABX12782), diphthamide biosynthesis protein (CUR52689), phosphoribosylamine-glycine ligase (CUR51614), glucosamine-1-phosphate N-acetyltransferase (WP_075054010), elongation factor 2 (WP_014964994, WP_014963048, WP_048116371, CUR52052) and several conserved hypothetical proteins (WP_014962442, AJM91735, AJM92436). Notably, the orthologous genes for hypothetical proteins in *Ca.* Nitrosopumilus piranensis D3C (AJM91735) and *Ca.* Np. koreensis AR1 (WP_014962442) are targeted by two unrelated iMGEs, whereas in *Np. maritimus* SCM1 and *Ca.* Np. adriaticus NF5, the corresponding genes are free of MGE integrations.

Due to the fact that *att*/TSD sites of thaumarchaeal elements are generally short (Fig. 2A), their unambiguous identification was challenging, particularly when integration occurred within unorthodox targets such as protein-coding genes. In all cases, the putative integration sites were meticulously verified by comparison of the corresponding genomic loci from closely related organisms with and without MGE insertions. An example of such analysis is shown in Fig. 2C. In the *Ns. evergladensis* SR1 genome, NitEve-E2 is inserted into the 3′-distal region of a gene encoding a Zn-finger protein (AIF83914). Although, the predicted *att* site is only 13 bp-long, comparison with the corresponding region in *Ns. viennensis* EN76 provided unequivocal support for the prediction site. Interestingly, NitEve-E2 insertion replaced a eight nucleotide sequence of the target gene including the stop codon (TAG) with a non-homologous MGE-derived sequence which contains an alternative stop codon (TAA), reconstituting the open reading frame (Fig. 2C).

A gene encoding elongation factor 2 (EF-2), a GTPase involved in the translocation step of the ribosome during protein synthesis, seems to serve as the most common target for integration of thaumarchaeal casposons (Krupovic *et al*., 2014). The integration of the casposons NitAR1-C1 and NitAR2-C1 in the genomes of *Ca.* Np. koreensis AR1 and *Ca.* Np. sediminis AR2, respectively, has been described previously (Krupovic *et al*., 2014). In the present study, we identified two new casposons, NitNF5-C1 and NitNF5-C2 (see below for description), which use the same cellular gene for integration, in the genome of *Ca.* Np. adriaticus NF5. The two elements are inserted in tandem into the same *ef-2* gene (Fig. 2D). Such tandem integrations have been previously described in the case of family 2 casposons in *Methanosarcina* sp. (Krupovic *et al*., 2016), but have not been observed for thaumarchaeal family 1 casposons. Notably, archaeal and bacterial MGEs that use tyrosine recombinases for integration are also known to form arrays of integrated elements by re-using the same integration site (Krupovic and Bamford, 2008b; Krupovic *et al*., 2010b; Das *et al*., 2013). *Ca.* Nt. devanaterra contains two family 1 casposons as well. One of these is also integrated in the *ef-2* gene, whereas the other one is inserted into the 3′-distal region of a gene encoding phosphoribosylamine-glycine ligase. Finally, the NitEve-C1 casposon identified in the *Ns. evergladensis* SR1 genome does not target any protein-coding genes but is inserted into an intergenic region. These new observations indicate that *ef-2* is not the universal target for thaumarchaeal casposons, even within the genus *Nitrosopumilus.*

Five major classes of thaumarchaeal MGE

Based on the gene content analysis, the thaumarchaeal iMGE could be broadly grouped into five major classes: (i) proviruses, (ii) casposons, (iii) putative integrative-conjugative elements (ICE), (iv) cryptic integrated elements (CIE) and (v) IS-like transposons. The first four classes include complex, multigene mobile elements, whereas IS-like transposons typically consist of 1 or 2 genes, one of which encodes a transposase. Hereafter, we reserve the term iMGE for the complex elements. The majority (n = 48) of the identified iMGE belong to the CIE category and might represent novel families of viruses or plasmids. The identified iMGE greatly vary in size, spanning nearly three orders of magnitude from 2.6 to 140 kb (median size of 16.8 kb; Fig. 1D). Collectively, the 74 iMGE amount to 1 938 724 bp of mobile thaumarchaeal DNA. Proviruses and casposons are rather uniform in size, all smaller than 20 kb, whereas ICE and CIE are more variable and reach 140 and 98 kb, respectively (Fig. 1D). Below we characterize all five classes of thaumarchaeal MGE in more detail.

Proviruses. Two groups of putative proviruses were identified in thaumarchaeal genomes: proviruses related to tailed bacterial and archaeal viruses of the order *Caudovirales*, and those related to viruses encoding the double jelly-roll (DJR) major capsid proteins (MCP). Searches

initiated with the sequences of the large terminase subunit (TerL), a signature protein of the *Caudovirales*, yielded five hits in thaumarchaeal genomes. Two of these hits were to the previously reported putative proviruses Nvie-Pro1 and NCAV2-Pro1 in the genomes of *Ns. viennensis* EN76 (Krupovic *et al*., 2011) and *Ca.* Nitrosocaldus cavascurensis SCU2 (Abby *et al*., 2018), respectively. The three new hits were in the genomes of *Ca.* C. subterraneum, *Ca.* Np. koreensis AR1 and *Ca.* Nitrosocaldus islandicus 3F. The latter element was identical to NCAV2-Pro1 from *Ca.* Nc. cavascurensis SCU2. In Nvie-Pro1 and NCAV2-Pro1, potential recombination sites and, consequently, the exact boarders of the elements could not be detected (Krupovic *et al*., 2011; Abby *et al*., 2018). Similarly, the boarders of CalSub-Pro in the genome of *Ca.* C. subterraneum could be determined only approximately. However, analysis of the gene content in the vicinity of *terL* in Nvie-Pro1, NCAV2-Pro1 and CalSub-Pro identify genes for all components necessary for the morphogenesis of full-fledged tailed virions. In CalSub-Pro, we identified gene homologues of the HK97-like MCP, the portal protein as well as the major and minor tail proteins, including the baseplate, head to tail connector, tail tape measure and tail fibre proteins (Fig. 3A). CalSub-Pro also contains a gene for the putative capsid maturation protease. Whereas Nvie-Pro1 encodes a chymotrypsin-like protease fused to the MCP (Krupovic *et al*., 2011), CalSub-Pro carries a gene for the typical S78-family caudoviral prohead protease (Pfam id: PF04586) located immediately upstream of the MCP gene, a typical gene order in *Caudovirales.* NCAV2-Pro1 (and NitIsl-Pro1) also encode a typical caudoviral prohead protease; however, unlike in CalSub-Pro but similar to Nvie-Pro1, the protease domain is fused to the MCP (Fig. 3A), highlighting the fluidity of the morphogenetic module in thaumarchaeal head-tail proviruses. Interestingly, neither of the proviruses contains identifiable genes for genome replication proteins. Given the lack of identifiable *att* sites and genome replication apparatus, on the one hand, and the presence of an apparently functional virion morphogenesis module on the other hand, there is a distinct possibility that the corresponding loci represent domesticated *Caudovirales*-derived elements, akin to the gene transfer agents (GTA) operating in some bacteria and euryarchaea (Lang *et al*., 2012; Lang *et al*., 2017; Koonin and Krupovic, 2018). Alternatively, these loci could be remnants of inactivated proviruses although conservation of the morphogenetic modules argues against this possibility. Notably, despite the shared gene contents, the three head-tail virus-derived elements described above are highly divergent and appear to be derived from distinct members of the *Caudovirales.*

Analysis of the *Ca.* Np. koreensis AR1 genome showed that the TerL homologue is indeed encoded within a putative iMGE, NitKor-E1. However, the only other identifiable *Caudovirales*-like gene in this elements was that for the small terminase subunit (TerS), located immediately upstream of the TerL-encoding gene, a typical location for this gene. All other genes in this element, although typical of MGE, could not be attributed to *Caudovirales* or any other group of viruses and included a VapBC toxin-antitoxin system, PD-(D/E)XK family restriction endonuclease and tyrosine integrase (Fig. 3A). The terminase complex is highly specific to viruses of the orders *Caudovirales* and *Herpesvirales*, and so far has not been identified in nonviral MGE. Thus, its function in NitKor-E1 remains enigmatic but likely is a relic from a past integration of a head-tailed virus. However, in the absence of other viral signature genes and experimental evidence of virion formation, we classify NitKor-E1 as a CIE rather than a provirus.

Viruses with the DJR MCPs infect hosts in all three domains of life (Krupovic and Bamford, 2008a; Krupovic and Koonin, 2015). In addition to the DJR MCP, these viruses share a specific genome packaging ATPase of the FtsK-HerA superfamily (Iyer *et al*., 2004) which is unrelated to TerL proteins of *Caudovirales* and *Herpesvirales.* The genes for the capsid protein and the packaging ATPases are typically encoded close to each other and appear to be inherited as a single module. In archaea, this supergroup of viruses is represented by *Sulfolobus* turreted icosahedral viruses, STIV and STIV2, two members of the family *Turriviridae* (Rice *et al*., 2004; Happonen *et al*., 2010). However, several other integrated and extrachromosomal MGE encoding both signature proteins have been described in euryarchaea and crenarchaea (Krupovic and Bamford, 2008b; Bernick *et al*., 2012; Gaudin *et al*., 2014; Rensen *et al*., 2015). The viral nature of these MGE has not been confirmed. However, a provirus closely related to STIV and STIV2 is integrated in the genome of certain *S. acidocaldarius* strains (Anderson *et al*., 2017; Mao and Grogan, 2017), suggesting that the euryarchaeal iMGE also represent functional viruses. Recently, homologues of DJR MCP have been reported also in thaumarchaea, but the exact boundaries of the putative proviruses have not been defined (Yutin *et al*., 2018). Searches seeded with the sequence of the STIV MCP yielded hits to three proteins in thaumarchaea: two identical proteins are encoded in the genome of *Ca.* Ns. gargensis Ga9_2 and the third one in the genome of *Ca.* Ns. evergladensis SR1.

The two identical MCP homologues in *Ca.* Ns. gargensis Ga9_2 genome are encoded within two nearly identical proviruses, NitGar-E1 and NitGar-E2, tandemly integrated into the same target site within an intergenic region. The most notable difference between the two elements is the presence of an IS*H3* family insertion sequence in NitGar-E1 (Fig. 3B). NitEve-E7 of *Ca.* Ns. evergladensis SR1 is

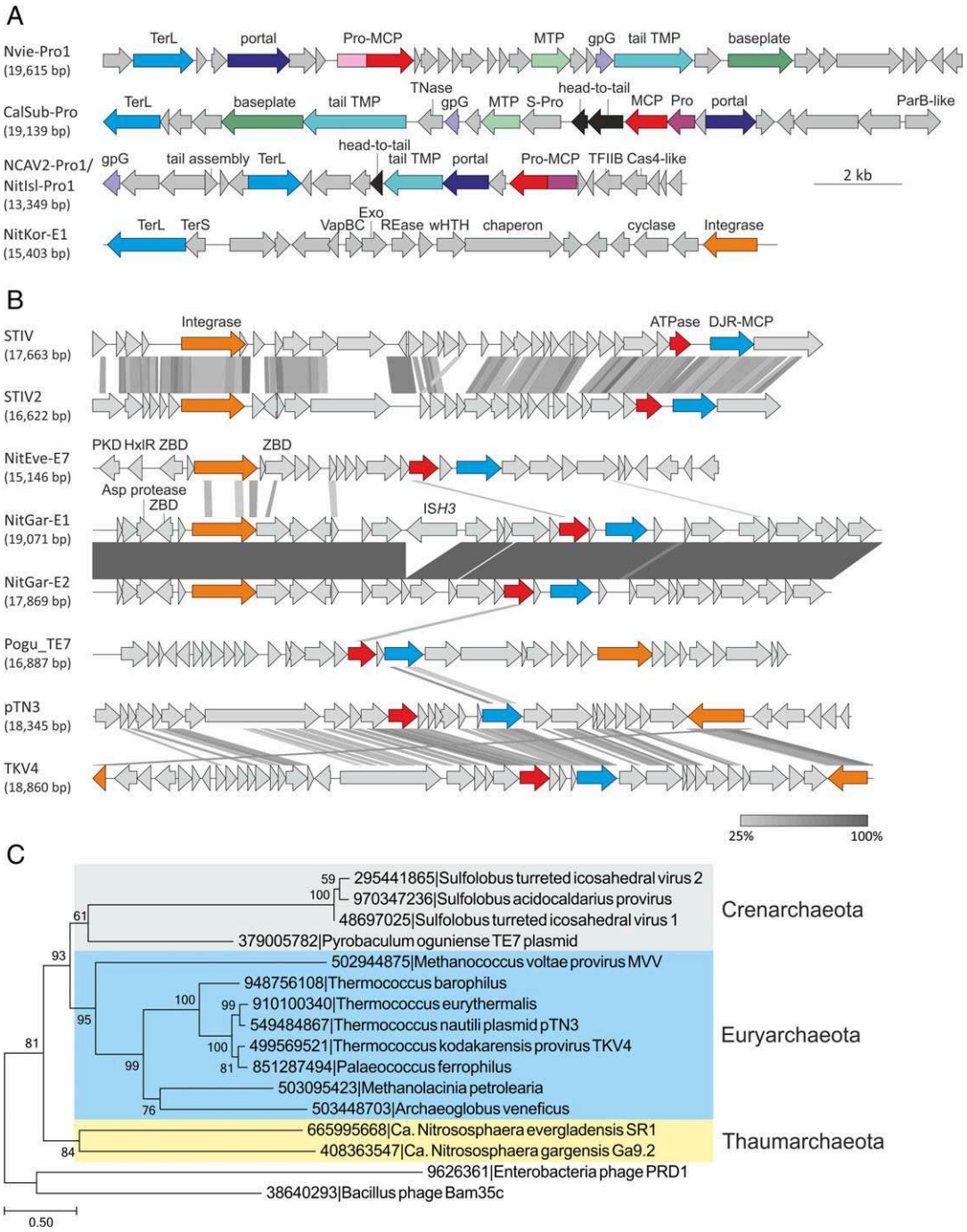


Fig. 3. Comparison of thaumarchaeal proviruses.

A. Genome maps of proviruses related to tailed bacterial and archaeal viruses of the order *Caudovirales.* Functionally equivalent genes are shown using the same colours.

Abbreviations: TerS and TerL, small and large subunits of the terminase, respectively; Pro, prohead maturation protease; S-Pro, serine protease; MCP, major capsid protein; MTP, major tail protein; TMP, tape measure protein; Exo, exonuclease; REase, restriction endonuclease; wHTH, winged helix-turn-helix.

B. Genome maps of archaeal viruses and proviruses encoding the DJR MCPs. Functionally equivalent genes are shown using the same colours.

Abbreviations: ATPase, A32-like genome packaging ATPase; ZBD, zinc-binding domain-containing protein; HxlR, HxlR family DNA-binding transcriptional regulator; PKD, PKD (Polycystic Kidney Disease) domain-containing protein; IS*H3*, IS*H3* family insertion sequence. For more detailed annotation see Supporting Information data file 1.

C. Maximum likelihood phylogeny of concatenated A32-like ATPase and DJR-MCP proteins. The tree was constructed using the automatic optimal model selection (RtREV +G6 + I + F) and is rooted with bacterial tectiviruses. The scale bar represents the number of substitution per site.

only distantly related to the proviruses of Ns. gargensis Ga9_2. As aforementioned, NitEve-E7 is integrated into NitEve-E6, an integrative-conjugative element (see below), suggesting that NitEve-E7 piggybacks NitEve-E6 to be transferred between cells via conjugation. Genomic context analysis shows that the MCP genes are encoded in the vicinity of a predicted genome packaging ATPases, as is the case for other archaeal viruses and proviruses of this supergroup (Fig. 3B). Besides the MCP and ATPase, the proviruses also share divergent integrases of the tyrosine recombinase superfamily. To better understand the evolutionary relationships among archaeal DJR MCP-encoding proviruses, we constructed a maximum likelihood phylogeny of concatenated ATPase and MCP proteins, two signature proteins shared by all elements, from representative (pro)viruses associated with crenarchaea, euryarchaea and thaumarchaea. Note that although all proviruses also encode integrases, these do not appear to be orthologous and seem to have been independently acquired or replaced in different virus lineages. The phylogenetic tree rooted with bacterial tectiviruses revealed three clades corresponding to 3 different archaeal phyla, Crenarchaeota, Euryarchaeota and Thaumarchaeota, respectively (Fig. 3C). This result suggests deep association and co-evolution of DJR MCP-encoding viruses with their archaeal hosts or distinct origins of these viruses in different archaeal phyla. Many more representatives of this virus supergroup from different archaeal phyla will be needed to distinguish between the two possibilities.

Casposons. Previously, we described 3 distinct thaumarchaeal casposons which were classified into family 1 (Krupovic *et al*., 2014). Differently from casposons from families 2, 3 and 4, family 1 casposons encode family B DNA polymerases (PolB) that shows the closest sequence similarity to protein-primed PolBs (pPolB) of archaeal viruses (Krupovic *et al*., 2014). Here, we identified five distinct family 1 casposons in the genomes of *Ca.* Ns. evergladensis SR1, *Ca.* Np. adriaticus NF5 and *Ca.* Nt. devanaterra. The latter two species each contain two casposons. Whereas the two casposons in *Ca.* Np. adriaticus NF5 are tandemly integrated into the same target site (Fig. 2D), those in *Ca.* Nt. devanaterra are inserted into different protein-coding genes. Notably, the five casposons are not closely related to each other or to those described previously (Fig. 4A).

Besides the genes for Cas1 and pPolB, family 1 casposons share 3 or 4 uncharacterized genes encoded immediately upstream of the *pPolB* gene. In addition, each casposon carries element-specific genes (Fig. 4A). The new casposons encode several nucleases that have not been previously observed in family 1, including OLD family nucleases (in NitDev-C1 and NitNF5-C1), NERD domain-containing nuclease related to Holliday junction

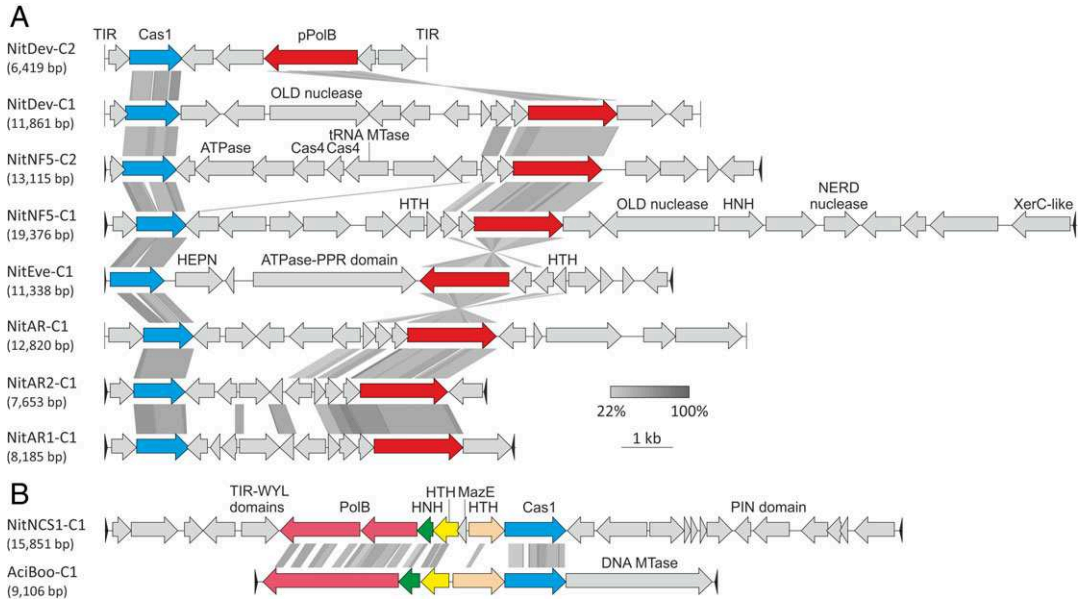


Fig. 4. Comparison of thaumarchaeal casposons.
A. Family 1 casposons.
B. Comparison of the family 2 casposons from *Ca.* Nitrosotalea okcheonensis CS (NitNCS1-C1) and *Aciduliprofundum boonei* (AciBoo-C1). Homologous genes are shown using the same colours.
Abbreviations: TIR, terminal inverted repeats; (p)PolB, (protein-primed) family B DNA polymerase; OLD, OLD (overcome lysogenization defect) family nuclease; HTH, helix-turn-helix; HNH, HNH family nuclease; MTase, methyltransferase. For detailed annotation see Supporting Information data file 1.

resolvases (NitNF5-C1) and HNH nuclease (NitNF5-C1). Most notably, NitNF5-C2 encodes two homologues of the Cas4 nuclease, which is involved in the adaptation process in many CRISPR-Cas systems (Hudaiberdiev *et al*., 2017; Kieper *et al*., 2018; Lee *et al*., 2018; Shiimori *et al*., 2018), and might participate in casposon integration, which is mechanistically closely similar to CRISPR spacer integration (Béguin *et al*., 2016; Krupovic *et al*., 2017). Both Cas4 copies display closest sequence similarity to Cas4 homologues from different *Clostridia.* Furthermore, NitEve-C1 encodes a HEPN nuclease, a member of an expansive nuclease family that is typically associated with various microbial defence systems, including toxin-antitoxin, abortive infection, restriction-modification as well as type III and type VI CRISPR-Cas systems (Anantharaman *et al*., 2013; Shmakov *et al*., 2015).

Finally, we identified a new casposon, NitNCS1-C1, in *Ca.* Nitrosotalea okcheonensis CS, which does not belong to family 1. It shares highest sequence similarity to the family 2 casposon AciBoo-C1 from *Aciduliprofundum boonei* (phylum Euryarchaeota), the only experimentally studied casposon thus far (Hickman and Dyda, 2015; Béguin *et al*., 2016). NitNCS1-C1 encodes a conserved set of proteins typical of family 2 casposons, including a distinct PolB, Cas1, HNH nuclease and 2 helix-turn-helix proteins (Fig. 4B). Notably, it also encodes a protein containing a WYL domain that is often found in regulators of the CRISPR-Cas systems (Makarova *et al*., 2014b; Yan *et al*., 2018). The PolB gene of NitNCS1-C1 appears to be fragmented, and it remains unclear whether the two fragments constitute a functional protein or the element is inactivated. Similar to AciBoo-C1 but unlike all other thaumarchaeal casposons, NitNCS1-C1 is inserted into a tRNA-Pro gene. Accordingly, NitNCS1-C1 is the first family 2 casposon in Thaumarchaeota.

Integrative-conjugative elements. The third type of identified thaumarchaeal iMGE are potential ICEs. ICEs are the largest among the four iMGE categories (median size of 64 kb; Fig. 5A). Two ICEs, NCAV2-E1 and NCAV2-E2, have been recently described in the genome of *Ca.* Nc. cavascurensis SCU2 (Abby *et al*., 2018). Here, we identified eight additional ICEs (Supporting Information Table S2). Similar to NCAV2-Pro1, orthologs of NCAV2-E1 and NCAV2-E2 are present in the genome of a closely related (ANI = 99.9%) species *Ca.* Nc. islandicus 3F (Daebeler *et al*., 2018). Notably, however, *Ca.* Nc. islandicus 3F harbours an additional ICE, NitIsl-E3, compared to *Ca.* Nc. cavascurensis SCU2, which instead has an empty site (Fig. 5A), confirming the recent mobility of NitIsl-E3. Figure 5B shows the regions of thaumarchaeal ICEs containing genes encoding components of the predicted conjugation/secretion systems. Similar to conjugative plasmids of *Sulfolobus* (Prangishvili *et al*., 1998; Greve *et al*., 2004), most of the thaumarchaeal ICEs carry a pair of signature genes for the homologues of VirB4/TrbE and VirD4/TraG ATPases which energize type IV secretion systems (Wallden *et al*., 2010). Other conserved components include homologues of the integral membrane proteins VirB6, VirB3 and TadC; FlaI and PilT ATPases; prepilin peptidase and pilins (Fig. 5B). Furthermore, all identified thaumarchaeal ICEs encode homologues of transcription factor IIB (TFIIB) which, in most elements, are located immediately upstream of the genes for the ParB-like partitioning protein, likely, in the same operon. Notably, TFIIB homologues have been previously detected in the vicinity of genes encoding type IV secretion systems in other archaea (Makarova *et al*., 2016). However, coupling with ParB appears to be specific to thaumarchaeal ICEs. Overall, the conserved genes were not syntenic (except in the orthologous ICEs; Fig. 5B), suggesting extensive recombination within the putative conjugation module. We did not detect candidates for relaxases which generate a single-stranded copy of ICE DNA prior to transfer in bacteria (Johnson and Grossman, 2015). However, typical relaxases are also absent in the bona fide conjugative plasmids of *Sulfolobus*, consistent with the suggestion that the archaeal conjugation machinery is distinct from that of bacteria and might transfer dsDNA as the substrate (Greve *et al*., 2004).

The predicted DNA replication modules of the thaumarchaeal ICEs also show considerable differences. Only NitEve-E6, the largest identified ICE, encodes its own DNA polymerase (PolB) that is more closely related to the PolBs from family 2 casposons (Krupovic *et al*., 2014) (hit to NitNCS1-C1 casposon, E = 3e-38, 41% identity), rather than to cellular replicative polymerases which were not recovered even after several PSI-BLAST iterations. NitGar-E3 and NitVie-E3 encode homologues of the Cdc6/Orc1 replication initiator, whereas NitVie-E3 and NitNCS1-E3 encode UvrD-like superfamily 1 helicases. NCAV2-E2 (and orthologous NitIsl-E2) carry genes for type IA topoisomerases which could also participate in their replication. NCAV2-E1 (and orthologous NitIsl-E1) and NitIsl-E3 encode MGE-specific replication proteins containing an N-terminal archaeo-eukaryotic primase (AEP) domain (also referred to as the primpol domain) and a C-terminal superfamily 3 helicase (S3H) domain, an organization commonly found in replication proteins of various MGE and viruses (Iyer *et al*., 2005; Lipps, 2011; Kazlauskas *et al*., 2018). The diversity of genome replication modules associated with thaumarchaeal ICEs suggests distinct origins and evolutionary histories of these elements.

Cryptic integrated elements. The CIE vary in size from 2.6 kb to 98 kb but the majority are smaller than 20 kb (median = 17 kb; Fig. 1D). There are no discernible signature genes that would be specific to thaumarchaeal

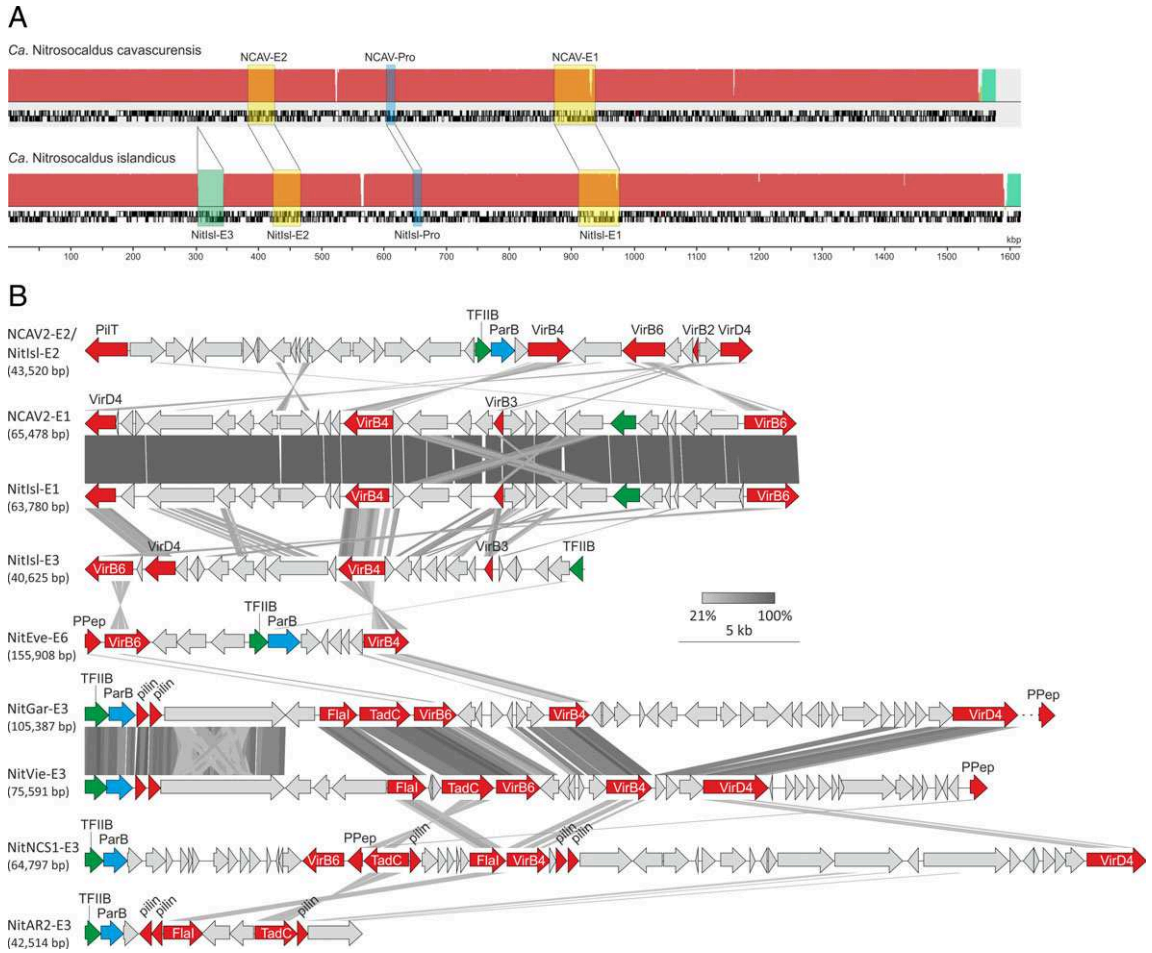


Fig. 5. Comparison of thaumarchaeal integrative-conjugative elements.
A. Comparison of the genomes of two closely related *Nitrosocaldus* strains, *Ca.* Nc. cavascurensis SCU2 and *Ca.* Nc. islandicus 3F. Shared ICEs and proviruses are indicated with transparent yellow and blue boxes, whereas the ICE element unique to *Ca.* Nc. islandicus 3F is shown highlighted with a green box.
B. Thaumarchaeal integrative-conjugative elements. Only regions including the genes encoding the predicted components of the conjugation apparatus are depicted (highlighted in red). Genes for the ParB-like segregation protein and TFIIB transcription initiation factor are shown in blue and green, respectively. PPep, prepilin peptidase. For detailed annotation see Supporting Information data file 1.

CIE. By definition, the most conserved protein, although belonging to different arCOGs, is the integrase. Interestingly, NitEve-E3 encodes an SSV1-like integrase which is split into two fragments upon integration of the MGE although no other homologues of viral genes were identified in this element. Similar to ICE, CIE encode diverse genome replication proteins, including those specific to MGEs (Fig. 6). ThaMY3-E2, the largest of the identified CIE (98.3 kb), encodes homologues of PolB and archaeal replicative helicase MCM, whereas NitGar-E6 and NitEve-E3 encode MCM but not PolB. The MCM helicases have been previously found to be frequently recruited from the host as the main replication proteins of various crenarchaeal and euryarchaeal MGEs, including viruses and plasmids (Krupovic *et al*., 2010b; Kazlauskas *et al*., 2016). By contrast, NitDev-E3 and NitAR2-E2 encode a superfamily 2 helicase and a homologue of the Cch helicase (AAA+ ATPase superfamily) from a *Staphylococcus aureus* mobile genomic island (Mir-Sanchis *et al*., 2016), respectively. NitAQ6f-E1 encodes a homologue of the Cdc6/Orc1 replication initiator, a distant homologue of the MCM helicases. Presumably, both the MCM helicases and Orc1 recruit the cellular replisome for the MGE replication. Some CIE, such as CalSub-E1, NitKor_MY1-E1 and NitAQ6f-E4, encode primpols. In the corresponding NitKor_MY1-E1 and NitAQ6f-E4 proteins, the primpol domain is fused to the S3H domain. By contrast, in CalSub-E1, the primpol domain, the α-helical PriCT-1 linker domain and the S3H domain are encoded by separate genes (Fig. 6). We also identified one thaumarchaeal CIE, NitAQ6f-E2, encoding a rolling-circle replication initiation endonuclease homologous to those of haloarchaeal sphaerolipovirus SNJ1 and several euryarchaeal plasmids (Wang *et al*., 2018b), suggesting that

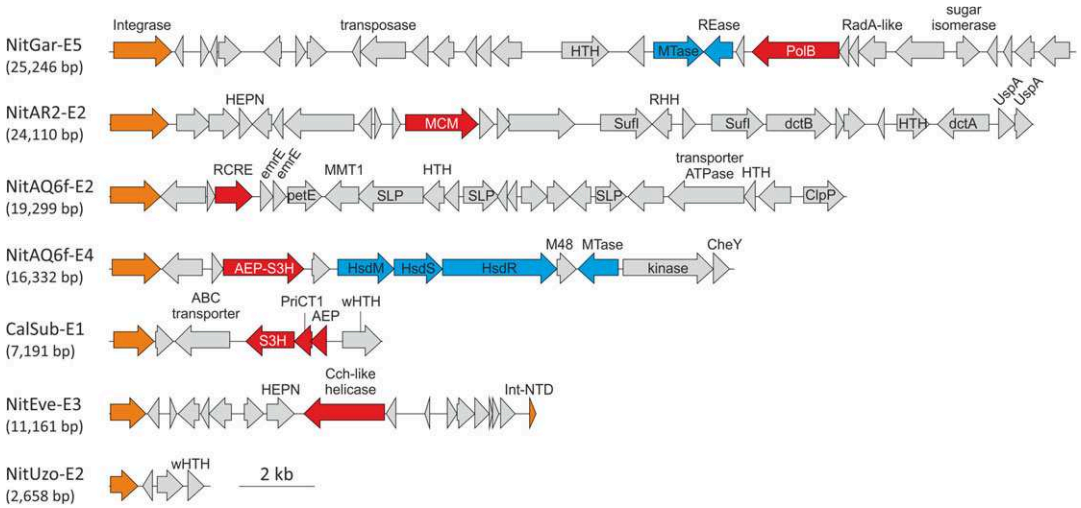


Fig. 6. Genome maps of selected thaumarchaeal cryptic integrated elements.
Integrase genes are highlighted in orange, gene encoding diverse replication-associated proteins are shown in red and components of the restriction-modification systems are in blue.
Abbreviations: UspA, UspA family nucleotide-binding protein; dctA, C4-dicarboxylic acids transport protein (Na+/H+ dicarboxylate symporter); dctB, C4-dicarboxylate transport sensor protein; SufI, multicopper oxidase; SLP, S-layer protein with immunoglobulin domain; PetE, Plastocyanin/azurin/halocyanin family protein; MMT1, Co/Zn/Cd cation transporter; HsdM/S/R, type I restriction-modification system methyltransferase/specificity/restriction subunits; MTase, methyltransferase; Mod: Adenine-specific DNA methyltransferase; REase, restriction endonuclease; M48, M48 family peptidase; CheY, chemotaxis protein receiver domain; EmrE, membrane transporter of cations and cationic drugs; RHH, ribbon-helix–helix domain-containing protein; (w)HTH, (winged) helix-turn-helix; RCRE, rolling circle replication initiation endonuclease; AEP, archaeo-eukaryotic primase; S3H, superfamily 3 helicase; MCM, minichromosome maintenance helicase.

NitAQ6f-E2 replicates by the rolling-circle mechanism. Finally, NitGar-E5 carries an operon consisting of a PolB gene, two copies of a gene encoding a small uncharacterized protein (arCOG08101), and an inactivated RadA homologue (Fig. 6). Similar operons have been previously identified in archaeal genomes and proposed to be involved in DNA repair or regulation of replication (Makarova *et al*., 2014a).

For many CIEs, we could not identify obvious candidates for replication proteins. For instance, the smallest identified CIE, NitUzo-E2 (2.6 kb), encodes only four predicted proteins, including an integrase, a winged helix-turn-helix (wHTH) protein and two hypothetical proteins (Fig. 6). The replication of this element might be initiated by the wHTH protein, as in the case of Reps from the IncP-1 family plasmids (Konieczny *et al*., 2014). However, given that wHTH proteins also are likely to be involved in transcription regulation, functional assignment without experimental verification appears premature. Overall, the replication modules of CIEs closely resemble those of ICEs, suggesting frequent transitions between the two types of iMGE. As a case in point, NitVie-E4 encodes a VirB6 homologue but no other recognizable proteins involved in conjugation, suggesting that this element evolved from an ICE ancestor via the loss of the conjugation apparatus which is consistent with the twice-smaller size of this element (20.2 kb) compared to that of ICE.

Insertion sequences. Although, previous comprehensive analysis of the IS diversity in archaea did not include representatives from the Thaumarchaeota (Filée *et al*., 2007), similar to many other archaea and bacteria, thaumarchaeal genomes are extensively parasitized by IS-like transposons. We identified 244 IS belonging to 13 families across 20 thaumarchaeal genomes (Fig. 7, Supporting Information Table S1). The majority of thaumarchaeal IS encode transposases of the DDE superfamily (11 IS families), whereas transposases of the HUH and serine recombinase superfamilies are characteristic of the IS*200*/IS*605* and IS*607* families, respectively. Notably, IS*150* family elements have not been previously described in archaea (Filée *et al*., 2007). There is considerable variation in both the copy number and diversity of IS elements among thaumarchaeal species (Fig. 7). Whereas most thaumarchaea carry only a few IS per genome, six species contain ten or more copies of different transposons (Fig. 1C, Supporting Information Table S1). The highest number of IS elements is found in *Ca.* Ns. gargensis Ga9_2 which carries 83 IS from 11 different families, with IS*200*/IS*605* being the dominant one (Fig. 7). There are signs of transposon proliferation and expansion for certain IS families. For instance, IS*5* elements in *Ca.* Nitrocosmicus oleophilus MY3 are found in 43 copies per chromosome, the largest for any thaumarchaeal IS family, whereas in all other species, they are present in low copy numbers or are lacking altogether. Some of the IS families are restricted to a single

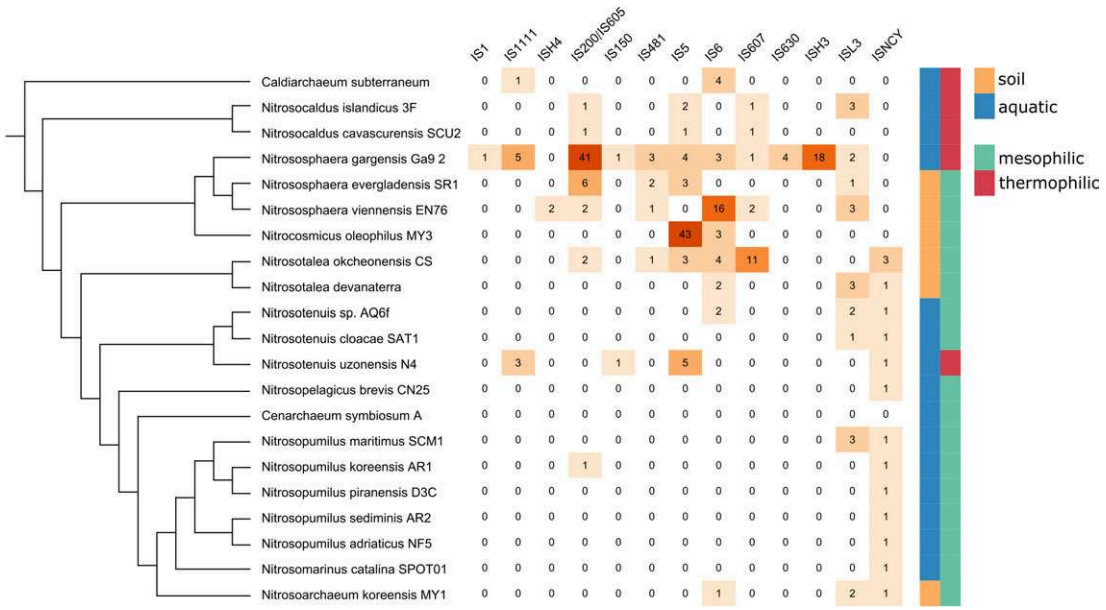


Fig. 7. Diversity and distribution of thaumarchaeal insertion sequences.
On the left is the schematic cladogram representing the relationships among thaumarchaeal species. The source of isolation is indicated on the right of the figure. The abundance of identified IS elements in each species is shown as a heatmap, with the exact numbers indicated within the corresponding cells.

thaumarchaeal species (IS*1*, IS*4*, IS*630*, IS*H3*; Fig. 7), suggesting a recent horizontal acquisition, but the sources of these transfers remain to be investigated.

iMGE-encoded CRISPR arrays

Four iMGE, namely, 2 ICE (NCAV2-E1 and NitIsl-E1) and 2 CIE (NitVie-E4 and NitEve-E4), were found to carry CRISPR arrays (Fig. 8A). In the two CIEs, the CRISPR arrays are adjacent to complete suites of Type-IB *cas* genes, including apparently functional adaptation and effector modules. By contrast, in the ICEs, the CRISPR arrays are not accompanied by *cas* genes. As aforementioned, NCAV2-E1 and NitIsl-E1 are closely related (Fig. 5A), and the major differences between the two ICEs involve the corresponding CRISPR arrays (Fig. 8A). Despite identical repeat sequences, the number of CRISPR spacers is different between the two elements (96 in NCAV2-E1 versus 69 in NitIsl-E1). Furthermore, only 43 spacers are shared between NCAV2-E1 and NitIsl-E1, whereas the rest of the spacers were apparently divergently acquired following the diversification of the two *Nitrosocaldus* strains, suggesting active exposure to distinct MGEs. For such *in trans* insertion of spacers by the host adaptation machinery to occur, the repeats in the iMGE should be (nearly) identical to those in the host CRISPR array. This is indeed the case, as the repeat sequences of NCAV2-E1/NitIsl-E1 are identical to those of the endogenous CRISPR array #3 of *Ca.* Nc. cavascurensis SCU2 which is accompanied by an apparently functional Type I-B *cas* genes, including the adaptation module (Abby *et al*., 2018). Notably, the repeat sequence of NitVie-E4 is closely related to that of NCAV2-E1/NitIsl-E1 (Fig. 8B), despite the lack of shared spacers and presence of the *cas* genes in NitVie-E4. Although the repeat sequence of NitEve-E4 is more divergent, its comparison with the repeat sequences from the other iMGEs (Fig. 8B) indicates that they all might be related.

To gain insight into the provenance of the iMGE-encoded CRISPR-Cas systems, we assessed the positions of the corresponding Cas1 proteins, the signature proteins of the CRISPR-Cas systems, in the global Cas1 phylogeny (Makarova *et al*., 2018). The Cas1 from NitVie-E4 was nested among bacterial Cas1 homologues from Type I-B systems, whereas Cas1 from NitEve-E4 forms a clade with homologues from *Ns. viennensis* EN7 and *Nitrosopumilus* sp. LS, which was nested among Cas1 associated with Type-III CIRSPR-Cas systems (Makarova *et al*., 2018). This phylogenetic position suggests that the Type I-B CRISPR-Cas systems carried by the two thaumarchaeal iMGE have been independently acquired from distinct sources. Furthermore, the similarity between the repeat sequences of the iMGE-carried stand-alone CRISPR arrays and the host array accompanied by *cas* genes suggests that the former evolved from the latter through the loss of the *cas* genes.

To investigate potential interplay between thaumarchaeal iMGE and CRISPR-Cas systems, we first examined if any of the cellular CRISPR spacers target the identified iMGE. Two spacers in the genome of *Ns. viennensis* EN7

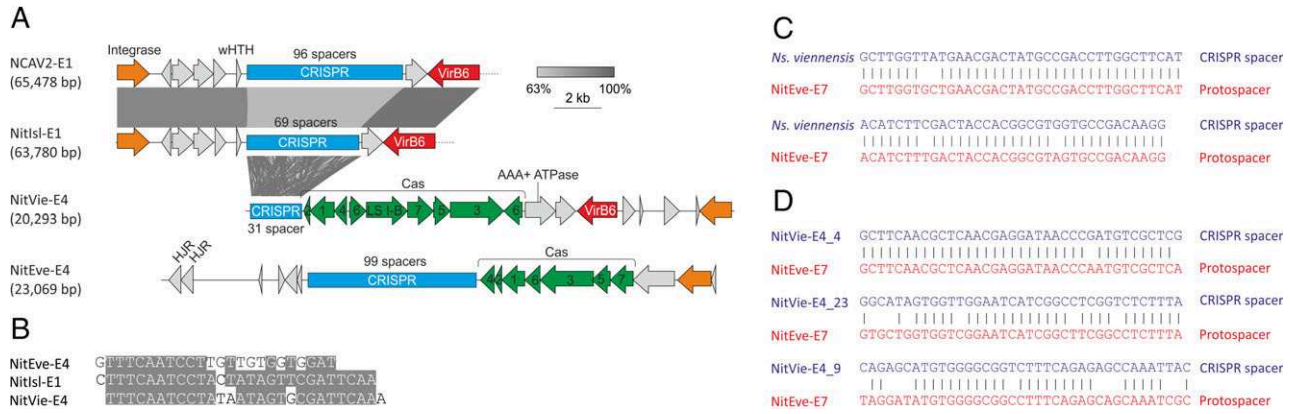


Fig. 8. CRISPR arrays carried by thaumarchaeal iMGE.
A. Loci of iMGE-carried stand-alone CRISPR arrays and CRISPR-Cas systems. CRISPR arrays are shown as blue rectangles with the number of spacers indicated. *cas* genes are shown in green and indicated with the corresponding numbers. LS, large subunit; HJR, Holliday junction resolvase; wHTH, winged helix-turn-helix.
B. Alignment of the CRISPR repeat sequences from NitIsl-E1/NCAV2-E1, NitVie-E4 and NitEve-E4 iMGE.
C. Matches between the chromosomal CRISPR spacers (blue) and iMGE (red).
D. Matches between the iMGE-carried CRISPR spacers (blue) and iMGE (red).

produced significant matches (95% and 94% identity, respectively) to the provirus NitEve-E7 (Fig. 8C). Notably, both spacers targeted different regions of the gene for the DJR MCP. Next, we analysed if the CRISPR spacers encoded by the four iMGEs target other iMGEs. Three spacers from the NitVie-E4 were found to match (95% [*E* = 2.5e-12], 79% [*E* = 1.1e-05] and 74% [*E* = 1.35e-04] identity, respectively) the NitEve-E7 provirus, with one of the spacers (NitVie-E4_4) targeting the DJR MCP gene (Fig. 8D) at a different region than the two spacers from the bona fide chromosomal *Ns. viennensis* EN7 CRISPR array. The similarities between the NitVie-E4_23 and NitVie-E4_9 spacers and their targets are at the boarder of significance. Thus, as a control, BLASTN search (word size 8, identity over full length of spacer > 70% and *E*-value <0.001) of spacer matches was performed against the *Escherichia coli* genome, which is of a similar size and GC content as our thaumarchaeal iMGE database. No spacer hits with the same thresholds were found in the control search. Furthermore, given that all five spacers (two from the host CRISPR array and three from NitVie-E4) with identifiable protospacers target the same provirus, it appears likely that these two matches are true positives. Finally, *Ns. viennensis* and *Ns. evergladensis* are both soil-dwellers (Tourna *et al*., 2011; Zhalnina *et al*., 2014). These observations suggest that the mobile CRISPR loci mediate conflicts between different iMGE competing in the same environment. Obviously, experimental validation is needed to corroborate this conjecture and assess its generality.

Functional potential of thaumarchaeal iMGE

To study the distribution and diversity of functions encoded by different classes of thaumarchaeal iMGE, the 2105 iMGE-encoded proteins were classified into functional arCOG categories (Makarova *et al*., 2015) (Supporting Information data file 1) and further segregated into five broader group (Fig. 9A). These include

i. 'Metabolism and transport' (arCOG categories C, E, F, G, H, I, P and Q);

ii. 'Cellular processes and signaling' (arCOG categories D, M, N, O and T);

iii. 'Information storage and processing' (arCOG categories J, K and L);

iv. 'Unknowns' (arCOG categories R and S, and hypothetical proteins which could not be ascribed to arCOGs);

v. 'Mobilome' (arCOG categories X, U and V; note that categories 'U': 'Intracellular trafficking, secretion and vesicular transport' and 'V': 'Defence mechanisms' containing the conjugation apparatus and various restriction-modification systems, respectively, are herein included into the 'Mobilome' group).

All 21 functional categories recognized in the arCOG database (Makarova *et al*., 2015) were represented among the iMGE proteins. As is typical of archaeal MGE (Makarova *et al*., 2014c), the majority (63%–82%) of proteins from all four iMGE classes lack functional annotation and fall into the 'Unknowns' group, with the highest number of such proteins found in proviruses (Fig. 9A). By contrast, the proteins typical of MGE, such as structural virion proteins, integrases, genome packaging ATPases, transposases and other proteins from the 'Mobilome' category, represented a core of less than 20% (less than 10% for proviruses and ICE) of the total protein content in each iMGE class. Notably, proviruses and casposons were relatively depleted in proteins of the groups 'Information storage and processing' and 'Cellular processes and signaling', whereas ICE

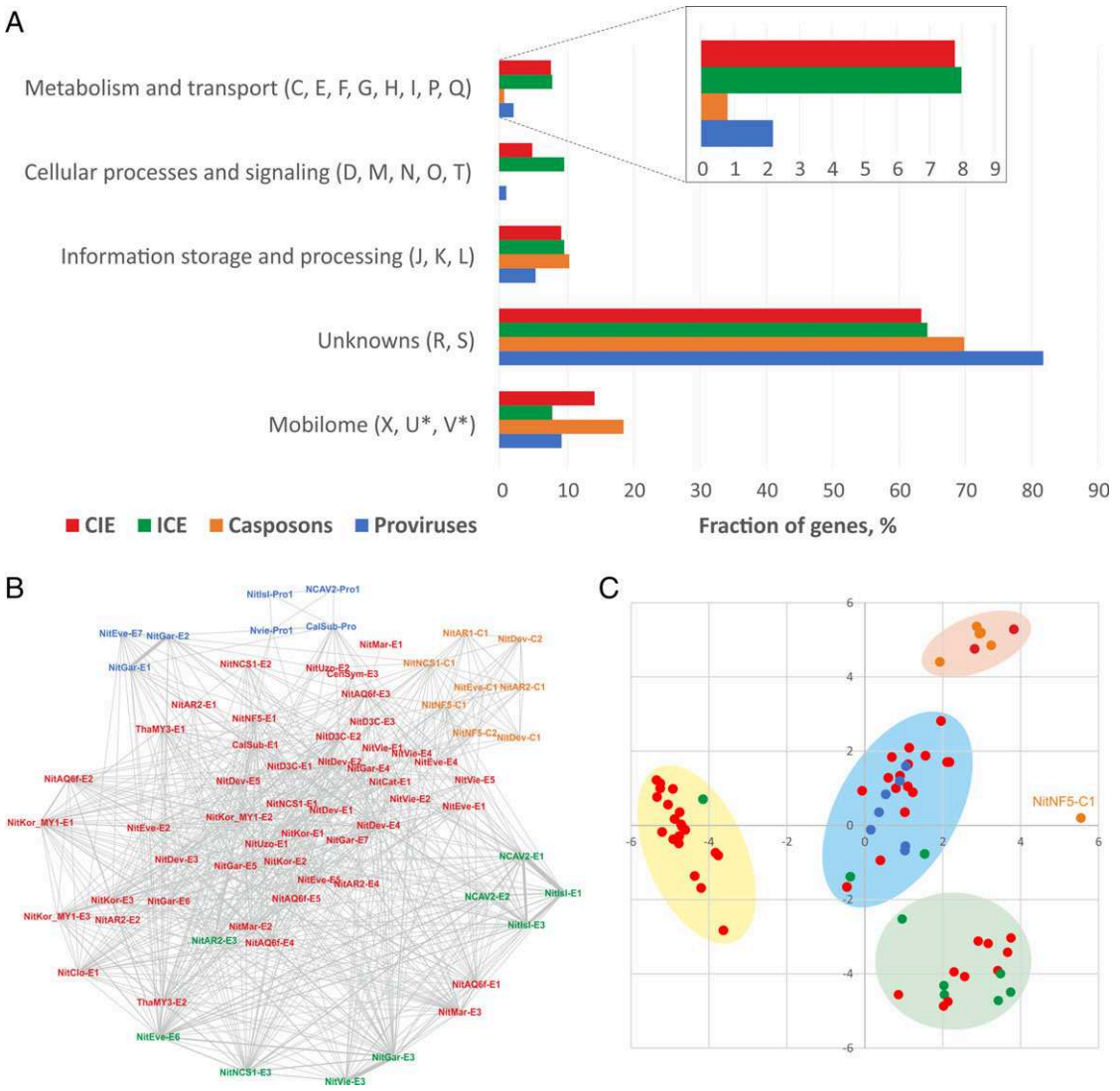


Fig. 9. Comparative genomics of thaumarchaeal iMGE.
A. Classification of genes from the four classes of iMGE into arCOG functional categories. Note that arCOG categories U (Intracellular trafficking, secretion and vesicular transport) and V (Defence mechanisms) are herein included into the 'Mobilome' category.
B. Network of thaumarchaeal iMGE based on the shared arCOGs. The nodes correspond to iMGE, whereas the connecting edges represent shared arCOGs. The four iMGE classes are colour-coded and the key is provided in panel A.
C. Classical multidimensional scaling analysis of iMGE. The four iMGE classes are colour-coded and the key is provided in panel A.

and CIE carry greater numbers of the so-called auxiliary metabolic genes (AMG) involved in metabolism and transport compared to proviruses and casposons (Fig. 9A, inset). For instance, many elements encode multicopper oxidases, which have been suggested to assist in the process of ammonia oxidation by producing NO (Schleper and Nicol, 2010; Kozlowski *et al*., 2016). In addition, one element, NitEve-E6, encodes an ammonia monooxygenase subunit C (AmoC; hit to PFAM profile PF04896.12, HHpred probability = 100%) and two iMGE encode nitrogen regulatory protein PII (HHpred probabilities > 99%), and might actively participate in nitrogen cycling in soil environments, as has been recently proposed for putative AmoC-encoding marine thaumarchaeal viruses assembled from metagenomic data (Ahlgren *et al*., 2019). In addition, iMGE were found to encode various dehydrogenases, stress response proteins, different membrane transporters of cations and drugs, chemotaxis protein receiver domains and many more (Supporting Information data file 1). The discovery of this diverse protein repertoire suggests that conjugative and cryptic elements play important roles in host adaptation and affect the fitness and survival of their hosts.

All thaumarchaeal iMGE are connected in a gene sharing network

Comparison of the gene (arCOG) content across the four classes of iMGE shows that all elements are connected to each other within a gene sharing network (Fig. 9B), indicating that some iMGE carry genes with broad distribution across different iMGE classes. Nevertheless, the two subgroups of proviruses (*Caudovirales* and DJR MCP-encoding proviruses, respectively) and casposons formed discernible clusters within this network, suggesting that, in the case of iMGE with relatively small genomes, a small set of core genes is sufficient to hold the (sub)classes together. By contrast, CIE and ICE were largely intermixed. Embedding the iMGE distance matrix into a 2-D space using Classical Multidimensional Scaling (CMDS) analysis (Borg and Groenen, 2005), revealed four clusters of elements (Fig. 9C). However, these clusters were not homogeneous with respect to the four iMGE classes. For instance, CIEs were distributed across all four clusters, whereas ICEs were present in three clusters. Notably, NitNF5-C1, the largest of the identified casposons (Fig. 4A), did not cluster with other casposons but was an outlier (Fig. 9C). This is not surprising, given that this casposon, besides the casposon-specific proteins, encodes several other proteins, including XerC-like tyrosine recombinase, that are shared with many other iMGE.

Table 1. Top 20 most common arCOGs from the thaumarchaeal iMGE.

Count	arCOG	Category	Annotation
33	arCOG01245	X	XerD/XerC family integrase
17	arCOG01242	X	XerD/XerC family integrase
16	arCOG02053	T	UspA family nucleotide-binding protein
13	arCOG00606	R	CBS domain
12	arCOG08677	S	Zn-ribbon domain containing protein
11	arCOG02626	V	Type I restriction-modification system, S subunit
10	arCOG01452	V	CRISPR-associated protein Cas1
9	arCOG08805	V	CopG/RHH family DNA binding protein
9	arCOG03914	Q	Multicopper oxidase
9	arCOG00602	R	CBS domain containing protein
9	arCOG00608	K	Predicted transcriptional regulator with C-terminal CBS domains
9	arCOG02632	V	Type I restriction-modification system, methyltransferase subunit
8	arCOG01471	R	Hemerythrin HHE cation binding domain containing protein
8	arCOG01981	K	Transcription initiation factor TFIIB
7	arCOG15271	X	Casposon associated protein-primed PolB family polymerase
7	arCOG04559	P	Membrane transporter of cations and cationic drugs
7	arCOG02868	O	Protein-disulfide isomerase
7	arCOG07844	S	VirB6/TrbL; membrane protein associated with conjugation system
6	arCOG14992	S	Uncharacterized protein conserved in casposons
6	arCOG00878	V	Type I restriction-modification system, restriction subunit

Analysis of the iMGE gene content revealed several protein families broadly distributed in iMGE (Table 1) which provide connectivity within the network. These include not only the XerC/XerD and Cas1 family integrases which, primarily, the former family, are essential for mobility and, thus, carried by the vast majority of iMGE, but also different families of transcription regulators, components of restriction modification and conjugation systems and several protein families potentially contributing to the host fitness and adaptation. For instance, 16 iMGE encode universal stress response proteins of the UspA family (Table 1). The proteins of the UspA family have been shown to play regulatory and protective roles to enable microbial adaptation and survival under various environmental stresses, such as nutrient starvation, drought, extreme temperatures, high salinity, the presence of antibiotics and heavy metals and other forms of stress (Vollmer and Bark, 2018). The connectivity of the iMGE network and the extent of gene sharing suggest that the thaumarchaeal mobilome has been shaped by three major processes, namely, (1) horizontal gene exchange, (2) independent acquisition of homologous genes from the host and (3) evolutionary transitions between different iMGE classes, in particular, between the CIE and ICE.

Discussion

Based on functional considerations and mode of propagation, thaumarchaeal iMGE can be categorized into five classes, namely, proviruses, casposons, ICE, CIE and the short IS-like transposons. Whereas IS-like transposons generally consist of 1 or 2 genes, those of the other four classes encompass multiple genes and display great diversity in terms of genomic complexity and functional content. All five classes of iMGE found in thaumarchaea are also present in other archaea (e.g. phylum Euryarchaeota) and bacteria although some of the classes have not been thus far identified in certain archaeal and bacterial lineages. For instance, casposons and viruses of the order *Caudovirales* have not been detected in members of the phylum Crenarchaeota. This might be due to insufficient sampling or to genuine lack of these elements in this archaeal phylum. By contrast, bacteria are known to

contain additional classes of iMGE that have not been detected in archaea, including thaumarchaea. These include composite DNA transposons which, in addition to the transposase genes, carry diverse passenger genes, such as those for antibiotic resistance (Nicolas *et al*., 2015); various pathogenicity islands and phage-inducible chromosomal islands that are induced upon phage infection and hijack the virus particle for intercellular transmission (Novick and Ram, 2016; 2017); mobile integrons, complex genetic platforms that allow bacteria to evolve rapidly through the acquisition, excision and shuffling of genes found in mobile elements known as cassettes (Escudero *et al*., 2015); or pipolins, a recently characterized group of bacterial iMGE encoding primer-independent DNA polymerases (Redrejo-Rodríguez *et al*., 2017). However, given our limited understanding on the archaeal mobilome and especially the diversity of iMGE, it cannot be ruled out that counterparts to some of these bacterial iMGE classes in thaumarchaea are awaiting discovery. The CIE class is particularly enigmatic and might include functionally distinct classes of iMGE.

In addition to proviruses related to tailed viruses of the order *Caudovirales*, which have been previously observed in thaumarchaeal genomes and also detected by several metagenomics studies (Chow *et al*., 2015; Labonté *et al*., 2015; Ahlgren *et al*., 2019; López-Pérez *et al*., 2018), we identified proviruses encoding the DJR MCP, one of the most widely distributed and diverse groups of dsDNA viruses in all three domains of life (Krupovic and Bamford, 2008a; Krupovic and Koonin, 2015; Yutin *et al*., 2018). Although the number of identified archaeal viruses with the DJR MCP is small, phylogenetic analysis suggests a coevolution of this virus group with the major archaeal lineages, including Thaumarchaeota. If validated by broader studies, this conclusion would parallel the apparently ancient evolutionary association of the *Caudovirales* with thaumarchaea (Krupovic *et al*., 2011). Thus, at least these two groups of viruses can be confidently traced to the last common ancestor of the archaea and, in all likelihood, to the last universal cellular ancestor. We did not identify any iMGE related to the archaea-specific virus groups associated with other archaeal phyla, and whether any of these extend to Thaumarchaeota, remain to be determined. Potentially, some or even many of the CIE, which comprise the majority of the identified thaumarchaeal iMGE (65%), represent novel families of archaeal viruses and plasmids. Systematic experimental induction of the replication of CIE and ICE could be a rewarding exercise, not only from a fundamental standpoint, but also to develop replicons that might serve as much-needed genetic tools in thaumarchaea. Identification of iMGE in thaumarchaea from diverse environments provides a broad choice of potential replicons that potentially could be tailored for different model organisms. Given their circular topology, CIE and ICE elements with smaller genome sizes (3–12 kbp) appear to be best suited for the development of shuttle vectors for facile genetic manipulation in *Escherichia coli.*

Gene content analysis revealed an extensive pangenome of thaumarchaeal iMGE. The MGE-specific genes, such as those encoding capsid proteins, viral genome packaging ATPases, conjugation proteins, integrases and so forth, constitute but a small fraction of their gene complements (10%–20% of genes). The vast majority of the iMGE genes encode proteins of unknown function. Nevertheless, a substantial fraction of genes represents auxiliary metabolic genes and stress response genes which are likely to play important roles in the adaption of their hosts to new environments, coping with stressful conditions and boosting their metabolic potential. For instance, multicopper oxidases, AmoC and nitrogen regulatory protein PII encoded by iMGE might modulate nitrogen metabolism, whereas UspA family proteins could boost the adaptation and survival of the host cells under various environmental stress conditions. The identification of functionally diverse metabolic and signalling genes in the thaumarcaheal iMGE parallels observations on the gene repertoires of some of the tailed bacterial viruses (Anantharaman *et al*., 2014; Hurwitz and U'Ren, 2016; Roux *et al*., 2016; Roitman *et al*., 2018), in particular, cyanophages that carry photosystem genes and substantially contribute to the host metabolism (Sharon *et al*., 2009; Thompson *et al*., 2011; Fridman *et al*., 2017). Taken together, these observations indicate that, at least, in the case of iMGEs with larger genomes, these elements should be considered more as symbionts of their hosts than simple genomic parasites or 'junk DNA'.

Although metabolism-related genes appear to be more prevalent in CIE and ICE, all four classes of iMGE share a substantial fraction of genes. Accordingly, the evolutionary relationships between these iMGE are most adequately represented as a gene-sharing network similar to those that have been previously constructed for double-stranded DNA viruses (Jachiet *et al*., 2014; Iranzo *et al*., 2016a,b; Bolduc *et al*., 2017). The extensive gene sharing can be explained by three nonmutually exclusive scenarios, including (1) horizontal gene exchange, (2) independent acquisition of homologous genes from various sources and (3) evolutionary transitions between different iMGE classes. Gene content similarity suggests that such transitions indeed occurred on multiple occasions between CIE and ICE, and involved the loss/acquisition of the genes encoding the conjugative apparatus.

The vast majority of known CRSIPR-Cas systems are encoded by cellular organisms and deployed to counter the replication of MGE, but some MGE also carry

functional CRISPR-Cas systems. For instance, CRISPR-Cas systems and stand-alone CRISPR arrays have been identified in a number of prophages (Hargreaves *et al*., 2014; Chénard *et al*., 2016; Zheng *et al*., 2016; Garneau *et al*., 2018) and in the case of a *Vibrio*-infecting bacteriophage have been shown to target for destruction a pathogenicity island integrated in the host genome (Seed *et al*., 2013). By contrast, a subgroup of Tn7-like transposons has been hypothesized to employ the encoded CRISPR-Cas system for CRISPR-guided transposition (Peters *et al*., 2017). We identified four iMGE carrying CRISPR arrays, which in two cases were accompanied by complete suites of *cas* genes. The majority of spacers did not match any known viruses, mostly likely, due to the current lack of data on the thaumarchaeal mobilome. Interestingly, however, several spacers carried by a CIE matched one of the proviruses, apparently, indicative of an antagonistic interaction between iMGE residing in the same habitat. Consequently, the CRISPR-carrying CIE and the host cell appear to coexist in a symbiotic relationship, whereby the CIE provides a protection against a presumably more harmful provirus. Identification of the CRISPR loci in MGE described here and elsewhere are consistent with the 'guns-for-hire' concept whereby MGE capture and repurpose various host defence systems (Koonin and Krupovic, 2015). Collectively, our results provide insights into the diversity and evolution of the thaumarchaeal mobilome and illuminate its potential impact on the functioning and adaptation of the host cells.

Experimental procedures

Identification of iMGE

Complete or near-complete thaumarchaeal genomes were downloaded from the NCBI database. We employed three different strategies to search for the iMGEs. (i) The genomes were analysed for the presence of gene clusters, previously denoted as 'dark matter' islands, enriched in ORFans and uncharacterized genes with a very narrow phyletic distribution (Makarova *et al*., 2014c). (ii) The second approach was based on identification of genes encoding signature proteins typical of different archaeal MGE groups. These included major capsid and genome packaging proteins representing different families of archaeal viruses, protein-primed family B DNA polymerases, rolling-circle replication initiation endonucleases and SSV-type DnaA-like AAA+ ATPase. Whenever a homologue of the signature MGE gene was identified in the cellular genome, the search was repeated with the identified thaumarchaeal homologue and its genomic context was analysed for the presence of additional MGE-derived genes using blastp. (iii) The third strategy involved systematic genome context analysis of genes encoding for integrases of the tyrosine recombinase superfamily. The searches were performed against the dataset of thaumarchaeal genomes using tblastn and integrase sequences from each newly identified thaumarchaeal iMGE as queries. The three approaches produced overlapping, yet complimentary results. In the next step, the potential iMGEs were analysed for the presence of signatures of site-specific recombination.

Identification of insertion sequences

IS elements were predicted and classified into families using the ISsaga platform (Varani *et al*., 2011). The 'probable false-positive' predicted by ISsaga were excluded from the final results. Exact coordinates for all identified IS elements are provided in Supporting Information Table S3.

Determination of the integration sites

The precise boundaries of integration were defined based on the presence of direct repeats corresponding to attachment sites or target site duplications. The direct and inverted repeats were searched for using Unipro UGENE (Okonechnikov *et al*., 2012). Whenever possible, additional validation of the MGE integration sites was obtained by comparing sequences of genomes containing the putative iMGEs with those of closely related genomes that do not contain such insertions using blastn algorithm.

Annotation of the iMGE genes

For each analysed gene, the functional annotations were assigned using the PSI-BLAST program with position specific scoring matrixes derived from arCOG alignments (Altschul *et al*., 1997). To detect remote homology, additional searches were performed using PSI-BLAST (Altschul *et al*., 1997) against the nonredundant protein database at NCBI and HHpred against the PDB, CDD, SCOPe and Pfam databases available through the MPI Bioinformatics Toolkit (Zimmermann *et al*., 2018).

Network analysis

The number of distinct arCOGs shared between a pair of elements (S_{ij}) was counted in annotated iMGEs. In the network representation the thickness of the line, connecting two iMGE is proportional to S_{ij}. The distance between two elements with the respective numbers of genes X_i and X_j is calculated as $-\ln(S_{ij}/\mathrm{sqrt}(X_iX_j))$. The iMGE distance matrix was embedded into a 2-D space using the classical multidimensional scaling (*cmdscale* function in R).

Phylogenetic analysis

For phylogenetic analysis, MCP and ATPase sequences from each (pro)virus were concatenated and aligned using MUSCLE (Edgar, 2004). Poorly aligned (low information content) positions were removed using the Gappyout function of Trimal (Capella-Gutierrez *et al*., 2009). The final alignment contained 462 positions. The maximum likelihood phylogenetic tree was constructed using the PhyML program (Guindon *et al*., 2010) with the automatic selection of the best-fit substitution model for a given alignment. The best model identified by PhyML was RtREV +G6 + I + F. The tree was rooted with sequences of bacterial tectiviruses. The branch support was assessed using aBayes implemented in PhyML.

Genome comparisons

The genomes of iMGE were compared and visualized using EasyFig v2.1 with tblastx algorithm (Sullivan *et al*., 2011). The complete genomes of closely related *Nitrosocaldus* strains, *Ca.* Nc. cavascurensis SCU2 and *Ca.* Nc. islandicus 3F were compared using progressiveMauve with default parameters (Darling *et al*., 2010).

Acknowledgements

This work was supported by the European Union's Horizon 2020 research and innovation program under grant agreement 685778 (project VIRUS-X), l'Agence Nationale de la Recherche (project ENVIRA, #ANR-17-CE15-0005-01) and European Research Council (ERC) grant from the European Union's Seventh Framework Program (FP/2007-2013)/Project EVOMOBIL-ERC Grant Agreement no. 340440. Y.I.W., K.S.M. and E.V.K. are supported through the intramural program of the U.S. National Institutes of Health. S.M. was supported by Vernadski fellowship from Campus France, RSF 14-14-00988 and Skoltech SBI program grant to Konstantin Severinov.

References

Abby, S.S., Melcher, M., Kerou, M., Krupovic, M., Stieglmeier, M., Rossel, C., *et al*. (2018) *Candidatus* Nitrosocaldus cavascurensis, an ammonia oxidizing, extremely thermophilic archaeon with a highly Mobile genome. *Front Microbiol* **9**: 28.

Ahlgren, N.A., Fuchsman, C.A., Rocap, G., and Fuhrman, J. A. (2019) Discovery of several novel, widespread, and ecologically distinct marine Thaumarchaeota viruses that encode *amoC* nitrification genes. *ISME J* **13**: 618–631.

Ahlgren, N.A., Chen, Y., Needham, D.M., Parada, A.E., Sachdeva, R., Trinh, V., *et al*. (2017) Genome and epigenome of a novel marine Thaumarchaeota strain suggest viral infection, phosphorothioation DNA modification and multiple restriction systems. *Environ Microbiol* **19**: 2434–2452.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389–3402.

Anantharaman, K., Duhaime, M.B., Breier, J.A., Wendt, K.A., Toner, B.M., and Dick, G.J. (2014) Sulfur oxidation genes in diverse deep-sea viruses. *Science* **344**: 757–760.

Anantharaman, V., Makarova, K.S., Burroughs, A.M., Koonin, E.V., and Aravind, L. (2013) Comprehensive analysis of the HEPN superfamily: identification of novel roles in intra-genomic conflicts, defense, pathogenesis and RNA processing. *Biol Direct* **8**: 15.

Anderson, R.E., Kouris, A., Seward, C.H., Campbell, K.M., and Whitaker, R.J. (2017) Structured populations of *Sulfolobus acidocaldarius* with susceptibility to mobile genetic elements. *Genome Biol Evol* **9**: 1699–1710.

Bayer, B., Vojvoda, J., Offre, P., Alves, R.J., Elisabeth, N.H., Garcia, J.A., *et al*. (2016) Physiological and genomic characterization of two novel marine thaumarchaeal strains indicates niche differentiation. *ISME J* **10**: 1051–1063.

Béguin, P., Charpin, N., Koonin, E.V., Forterre, P., and Krupovic, M. (2016) Casposon integration shows strong target site preference and recapitulates protospacer integration by CRISPR-Cas systems. *Nucleic Acids Res* **44**: 10367–10376.

Bernick, D.L., Karplus, K., Lui, L.M., Coker, J.K., Murphy, J. N., Chan, P.P., *et al*. (2012) Complete genome sequence of *Pyrobaculum oguniense*. *Stand Genomic Sci* **6**: 336–345.

Bolduc, B., Jang, H.B., Doulcier, G., You, Z.Q., Roux, S., and Sullivan, M.B. (2017) vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect archaea and bacteria. *PeerJ* **5**: e3243.

Borg, I., and Groenen, P. (2005) *Modern Multidimensional Scaling: Theory and Applications*. New York: Springer-Verlag.

Capella-Gutierrez, S., Silla-Martinez, J.M., and Gabaldon, T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**: 1972–1973.

Chénard, C., Wirth, J.F., and Suttle, C.A. (2016) Viruses infecting a freshwater filamentous cyanobacterium (*Nostoc* sp.) encode a functional CRISPR array and a proteobacterial DNA polymerase B. *mBio* **7**: e00667-16.

Chow, C.E., Winget, D.M., White, R.A., III, Hallam, S.J., and Suttle, C.A. (2015) Combining genomic sequencing methods to explore viral diversity and reveal potential virus-host interactions. *Front Microbiol* **6**: 265.

Cossu, M., Badel, C., Catchpole, R., Gadelle, D., Marguet, E., Barbe, V., *et al*. (2017) Flipping chromosomes in deep-sea archaea. *PLoS Genet* **13**: e1006847.

Craig, N.L., Chandler, M., Gellert, M., Lambowitz, A.M., Rice, P.A., and Sandmeyer, S.B. (2015) *In Mobile DNA III*. Washington, DC: ASM Press.

Daebeler, A., Herbold, C.W., Vierheilig, J., Sedlacek, C.J., Pjevac, P., Albertsen, M., *et al*. (2018) Cultivation and genomic analysis of "*Candidatus* Nitrosocaldus islandicus," an Obligately thermophilic, ammonia-oxidizing Thaumarchaeon from a hot spring biofilm in Graendalur Valley, Iceland. *Front Microbiol* **9**: 193.

Danovaro, R., Rastelli, E., Corinaldesi, C., Tangherlini, M., and Dell'Anno, A. (2017) Marine archaea and archaeal viruses under global change. *F1000Res* **6**: 1241.

Danovaro, R., Dell'Anno, A., Corinaldesi, C., Rastelli, E., Cavicchioli, R., Krupovic, M., *et al*. (2016) Virus-mediated archaeal hecatomb in the deep seafloor. *Sci Adv* **2**: e1600492.

Darling, A.E., Mau, B., and Perna, N.T. (2010) progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* **5**: e11147.

Das, B., Martinez, E., Midonet, C., and Barre, F.X. (2013) Integrative mobile elements exploiting Xer recombination. *Trends Microbiol* **21**: 23–30.

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**: 113.

Escudero, J.A., Loot, C., Nivina, A., and Mazel, D. (2015) The Integron: adaptation on demand. *Microbiol Spectr* **3**: MDNA3-0019-2014.

Filée, J., Siguier, P., and Chandler, M. (2007) Insertion sequence diversity in archaea. *Microbiol Mol Biol Rev* **71**: 121–157.

Forterre, P., Krupovic, M., Raymann, K., and Soler, N. (2014) Plasmids from Euryarchaeota. *Microbiol Spectr* **2**: PLAS-0027-2014.

Frank, J.A., and Feschotte, C. (2017) Co-option of endogenous viral sequences for host cell function. *Curr Opin Virol* **25**: 81–89.

Fridman, S., Flores-Uribe, J., Larom, S., Alalouf, O., Liran, O., Yacoby, I., *et al*. (2017) A myovirus encoding both photosystem I and II proteins enhances cyclic electron flow in infected *Prochlorococcus* cells. *Nat Microbiol* **2**: 1350–1357.

Frost, L.S., and Koraimann, G. (2010) Regulation of bacterial conjugation: balancing opportunity with adversity. *Future Microbiol* **5**: 1057–1071.

Garneau, J.R., Sekulovic, O., Dupuy, B., Soutourina, O., Monot, M., and Fortier, L.C. (2018) High prevalence and genetic diversity of large phiCD211 (phiCDIF1296T)-like prophages in *Clostridioides difficile*. *Appl Environ Microbiol* **84**: e02164-17.

Gaudin, M., Krupovic, M., Marguet, E., Gauliard, E., Cvirkaite-Krupovic, V., Le Cam, E., *et al*. (2014) Extracellular membrane vesicles harbouring viral genomes. *Environ Microbiol* **16**: 1167–1175.

Greve, B., Jensen, S., Brugger, K., Zillig, W., and Garrett, R. A. (2004) Genomic comparison of archaeal conjugative plasmids from *Sulfolobus*. *Archaea* **1**: 231–239.

Grindley, N.D., Whiteson, K.L., and Rice, P.A. (2006) Mechanisms of site-specific recombination. *Annu Rev Biochem* **75**: 567–605.

Guédon, G., Libante, V., Coluzzi, C., Payot, S., and Leblond-Bourget, N. (2017) The obscure world of integrative and Mobilizable elements, highly widespread elements that pirate bacterial conjugative systems. *Genes (Basel)* **8**: E337.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**: 307–321.

Hallam, S.J., Konstantinidis, K.T., Putnam, N., Schleper, C., Watanabe, Y., Sugahara, J., *et al*. (2006) Genomic analysis of the uncultivated marine crenarchaeote *Cenarchaeum symbiosum*. *Proc Natl Acad Sci USA* **103**: 18296–18301.

Happonen, L.J., Redder, P., Peng, X., Reigstad, L.J., Prangishvili, D., and Butcher, S.J. (2010) Familial relationships in hyperthermo- and acidophilic archaeal viruses. *J Virol* **84**: 4747–4754.

Hargreaves, K.R., Flores, C.O., Lawley, T.D., and Clokie, M. R. (2014) Abundant and diverse clustered regularly interspaced short palindromic repeat spacers in *Clostridium difficile* strains and prophages target multiple phage types within this pathogen. *mBio* **5**: e01045-13.

Herbold, C.W., Lehtovirta-Morley, L.E., Jung, M.Y., Jehmlich, N., Hausmann, B., Han, P., *et al*. (2017) Ammonia-oxidising archaea living at low pH: insights from comparative genomics. *Environ Microbiol* **19**: 4939–4952.

Hickman, A.B., and Dyda, F. (2015) The casposon-encoded Cas1 protein from *Aciduliprofundum boonei* is a DNA integrase that generates target site duplications. *Nucleic Acids Res* **43**: 10576–10587.

Hudaiberdiev, S., Shmakov, S., Wolf, Y.I., Terns, M.P., Makarova, K.S., and Koonin, E.V. (2017) Phylogenomics of Cas4 family nucleases. *BMC Evol Biol* **17**: 232.

Hurwitz, B.L., and U'Ren, J.M. (2016) Viral metabolic reprogramming in marine ecosystems. *Curr Opin Microbiol* **31**: 161–168.

Iranzo, J., Krupovic, M., and Koonin, E.V. (2016a) The double-stranded DNA Virosphere as a modular hierarchical network of gene sharing. *mBio* **7**: e00978-16.

Iranzo, J., Koonin, E.V., Prangishvili, D., and Krupovic, M. (2016b) Bipartite network analysis of the archaeal Virosphere: evolutionary connections between viruses and Capsidless Mobile elements. *J Virol* **90**: 11043–11055.

Iyer, L.M., Makarova, K.S., Koonin, E.V., and Aravind, L. (2004) Comparative genomics of the FtsK-HerA superfamily of pumping ATPases: implications for the origins of chromosome segregation, cell division and viral capsid packaging. *Nucleic Acids Res* **32**: 5260–5279.

Iyer, L.M., Koonin, E.V., Leipe, D.D., and Aravind, L. (2005) Origin and evolution of the archaeo-eukaryotic primase superfamily and related palm-domain proteins: structural insights and new members. *Nucleic Acids Res* **33**: 3875–3896.

Jachiet, P.A., Colson, P., Lopez, P., and Bapteste, E. (2014) Extensive gene remodeling in the viral world: new evidence for nongradual evolution in the mobilome network. *Genome Biol Evol* **6**: 2195–2205.

Jangam, D., Feschotte, C., and Betran, E. (2017) Transposable element domestication as an adaptation to evolutionary conflicts. *Trends Genet* **33**: 817–831.

Johnson, C.M., and Grossman, A.D. (2015) Integrative and conjugative elements (ICEs): what they do and how they work. *Annu Rev Genet* **49**: 577–601.

Kazlauskas, D., Krupovic, M., and Venclovas, C. (2016) The logic of DNA replication in double-stranded DNA viruses: insights from global analysis of viral genomes. *Nucleic Acids Res* **44**: 4551–4564.

Kazlauskas, D., Sezonov, G., Charpin, N., Venclovas, C., Forterre, P., and Krupovic, M. (2018) Novel families of Archaeo-eukaryotic primases associated with Mobile

genetic elements of bacteria and archaea. *J Mol Biol* **430**: 737–750.

Kieper, S.N., Almendros, C., Behler, J., McKenzie, R.E., Nobrega, F.L., Haagsma, A.C., *et al*. (2018) Cas4 facilitates PAM-compatible spacer selection during CRISPR adaptation. *Cell Rep* **22**: 3377–3384.

Kim, B.K., Jung, M.Y., Yu, D.S., Park, S.J., Oh, T.K., Rhee, S. K., and Kim, J.F. (2011) Genome sequence of an ammonia-oxidizing soil archaeon, "*Candidatus* Nitrosoarchaeum koreensis" MY1. *J Bacteriol* **193**: 5539–5540.

Konieczny, I., Bury, K., Wawrzycka, A., and Wegrzyn, K. (2014) Iteron plasmids. *Microbiol Spectr* **2**: PLAS-0026-2014.

Koonin, E.V., and Krupovic, M. (2015) A movable defense. *Scientist* **29**: 46–53.

Koonin, E.V., and Krupovic, M. (2018) The depths of virus exaptation. *Curr Opin Virol* **31**: 1–8.

Kozlowski, J.A., Stieglmeier, M., Schleper, C., Klotz, M.G., and Stein, L.Y. (2016) Pathways and key intermediates required for obligate aerobic ammonia-dependent chemolithotrophy in bacteria and Thaumarchaeota. *ISME J* **10**: 1836–1845.

Krupovic, M., and Bamford, D.H. (2008a) Virus evolution: how far does the double beta-barrel viral lineage extend? *Nat Rev Microbiol* **6**: 941–948.

Krupovic, M., and Bamford, D.H. (2008b) Archaeal proviruses TKV4 and MVV extend the PRD1-adenovirus lineage to the phylum Euryarchaeota. *Virology* **375**: 292–300.

Krupovic, M., and Koonin, E.V. (2015) Polintons: a hotbed of eukaryotic virus, transposon and plasmid evolution. *Nat Rev Microbiol* **13**: 105–115.

Krupovic, M., Forterre, P., and Bamford, D.H. (2010a) Comparative analysis of the mosaic genomes of tailed archaeal viruses and proviruses suggests common themes for virion architecture and assembly with tailed viruses of bacteria. *J Mol Biol* **397**: 144–160.

Krupovic, M., Béguin, P., and Koonin, E.V. (2017) Casposons: mobile genetic elements that gave rise to the CRISPR-Cas adaptation machinery. *Curr Opin Microbiol* **38**: 36–43.

Krupovic, M., Gribaldo, S., Bamford, D.H., and Forterre, P. (2010b) The evolutionary history of archaeal MCM helicases: a case study of vertical evolution combined with hitchhiking of mobile genetic elements. *Mol Biol Evol* **27**: 2716–2732.

Krupovic, M., Spang, A., Gribaldo, S., Forterre, P., and Schleper, C. (2011) A thaumarchaeal provirus testifies for an ancient association of tailed viruses with archaea. *Biochem Soc Trans* **39**: 82–88.

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

Krupovic, M., Shmakov, S., Makarova, K.S., Forterre, P., and Koonin, E.V. (2016) Recent mobility of Casposons, self-synthesizing transposons at the origin of the CRISPR-Cas immunity. *Genome Biol Evol* **8**: 375–386.

Krupovic, M., Cvirkaite-Krupovic, V., Iranzo, J., Prangishvili, D., and Koonin, E.V. (2018) Viruses of archaea: structural, functional, environmental and evolutionary genomics. *Virus Res* **244**: 181–193.

Labonté, J.M., Swan, B.K., Poulos, B., Luo, H., Koren, S., Hallam, S.J., *et al*. (2015) Single-cell genomics-based analysis of virus-host interactions in marine surface bacterioplankton. *ISME J* **9**: 2386–2399.

Lang, A.S., Zhaxybayeva, O., and Beatty, J.T. (2012) Gene transfer agents: phage-like elements of genetic exchange. *Nat Rev Microbiol* **10**: 472–482.

Lang, A.S., Westbye, A.B., and Beatty, J.T. (2017) The distribution, evolution, and roles of gene transfer agents in prokaryotic genetic exchange. *Annu Rev Virol* **4**: 87–104.

Lebedeva, E.V., Hatzenpichler, R., Pelletier, E., Schuster, N., Hauzmayer, S., Bulaev, A., *et al*. (2013) Enrichment and genome sequence of the group I.1a ammonia-oxidizing archaeon "*Ca.* Nitrosotenuis uzonensis" representing a clade globally distributed in thermal habitats. *PLoS One* **8**: e80835.

Lee, H., Zhou, Y., Taylor, D.W., and Sashital, D.G. (2018) Cas4-dependent prespacer processing ensures high-Fidelity programming of CRISPR arrays. *Mol Cell* **70**: e5.

Lehtovirta-Morley, L.E., Stoecker, K., Vilcinskas, A., Prosser, J.I., and Nicol, G.W. (2011) Cultivation of an obligate acidophilic ammonia oxidizer from a nitrifying acid soil. *Proc Natl Acad Sci USA* **108**: 15892–15897.

Lehtovirta-Morley, L.E., Sayavedra-Soto, L.A., Gallois, N., Schouten, S., Stein, L.Y., Prosser, J.I., and Nicol, G.W. (2016) Identifying potential mechanisms enabling Acidophily in the ammonia-oxidizing archaeon "*Candidatus* Nitrosotalea devanaterra". *Appl Environ Microbiol* **82**: 2608–2619.

Li, Y., Ding, K., Wen, X., Zhang, B., Shen, B., and Yang, Y. (2016) A novel ammonia-oxidizing archaeon from wastewater treatment plant: its enrichment, physiological and genomic characteristics. *Sci Rep* **6**: 23747.

Lipps, G. (2011) Structure and function of the primase domain of the replication protein from the archaeal plasmid pRN1. *Biochem Soc Trans* **39**: 104–106.

López-Pérez, M., Haro-Moreno, J.M., de la Torre, J.R., and Rodriguez-Valera, F. (2018) Novel Caudovirales associated with marine group I Thaumarchaeota assembled from metagenomes. *Environ Microbiol* (In press).

Mahillon, J., and Chandler, M. (1998) Insertion sequences. *Microbiol Mol Biol Rev* **62**: 725–774.

Makarova, K.S., Krupovic, M., and Koonin, E.V. (2014a) Evolution of replicative DNA polymerases in archaea and their contributions to the eukaryotic replication machinery. *Front Microbiol* **5**: 354.

Makarova, K.S., Wolf, Y.I., and Koonin, E.V. (2015) Archaeal clusters of orthologous genes (arCOGs): an update and application for analysis of shared features between Thermococcales, Methanococcales, and Methanobacteriales. *Life* **5**: 818–840.

Makarova, K.S., Koonin, E.V., and Albers, S.V. (2016) Diversity and evolution of type IV pili systems in archaea. *Front Microbiol* **7**: 667.

Makarova, K.S., Wolf, Y.I., and Koonin, E.V. (2018) Classification and nomenclature of CRISPR-Cas systems: where from here? *CRISPR J* **1**: 325–336. https://doi.org/10.1089/crispr.2018.0033.

Makarova, K.S., Anantharaman, V., Grishin, N.V., Koonin, E. V., and Aravind, L. (2014b) CARF and WYL domains: ligand-binding regulators of prokaryotic defense systems. *Front Genet* **5**: 102.

Makarova, K.S., Wolf, Y.I., Forterre, P., Prangishvili, D., Krupovic, M., and Koonin, E.V. (2014c) Dark matter in archaeal genomes: a rich source of novel mobile elements, defense systems and secretory complexes. *Extremophiles* **18**: 877–893.

Mao, D., and Grogan, D.W. (2017) How a genetically stable extremophile evolves: modes of genome diversification in the archaeon *Sulfolobus acidocaldarius*. *J Bacteriol* **199**: e00177-17.

Martens-Habbena, W., Berube, P.M., Urakawa, H., de la Torre, J.R., and Stahl, D.A. (2009) Ammonia oxidation kinetics determine niche separation of nitrifying Archaea and Bacteria. *Nature* **461**: 976–979.

Mir-Sanchis, I., Roman, C.A., Misiura, A., Pigli, Y.Z., Boyle-Vavra, S., and Rice, P.A. (2016) Staphylococcal SCCmec elements encode an active MCM-like helicase and thus may be replicative. *Nat Struct Mol Biol* **23**: 891–898.

Munson-McGee, J.H., Snyder, J.C., and Young, M.J. (2018) Archaeal viruses from high-temperature environments. *Genes (Basel)* **9**: E128.

Nicolas, E., Lambin, M., Dandoy, D., Galloy, C., Nguyen, N., Oger, C.A., and Hallet, B. (2015) The Tn3-family of replicative transposons. *Microbiol Spectr* **3**: MDNA3-0060-2014.

Novick, R.P., and Ram, G. (2016) The floating (pathogenicity) Island: a genomic dessert. *Trends Genet* **32**: 114–126.

Novick, R.P., and Ram, G. (2017) Staphylococcal pathogenicity islands-movers and shakers in the genomic firmament. *Curr Opin Microbiol* **38**: 197–204.

Nunoura, T., Takaki, Y., Kakuta, J., Nishi, S., Sugahara, J., Kazama, H., *et al*. (2011) Insights into the evolution of archaea and eukaryotic protein modifier systems revealed by the genome of a novel archaeal group. *Nucleic Acids Res* **39**: 3204–3223.

Offre, P., Spang, A., and Schleper, C. (2013) Archaea in biogeochemical cycles. *Annu Rev Microbiol* **67**: 437–457.

Okonechnikov, K., Golosova, O., and Fursov, M. (2012) Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics* **28**: 1166–1167.

Omelchenko, M.V., Wolf, Y.I., Gaidamakova, E.K., Matrosova, V.Y., Vasilenko, A., Zhai, M., *et al*. (2005) Comparative genomics of *Thermus thermophilus* and *Deinococcus radiodurans*: divergent routes of adaptation to thermophily and radiation resistance. *BMC Evol Biol* **5**: 57.

Park, S.J., Ghai, R., Martin-Cuadrado, A.B., Rodriguez-Valera, F., Chung, W.H., Kwon, K., *et al*. (2014) Genomes of two new ammonia-oxidizing archaea enriched from deep marine sediments. *PLoS One* **9**: e96449.

Partridge, S.R., Kwong, S.M., Firth, N., and Jensen, S.O. (2018) Mobile genetic elements associated with antimicrobial resistance. *Clin Microbiol Rev* **31**: e00088-17.

Peters, J.E., Makarova, K.S., Shmakov, S., and Koonin, E.V. (2017) Recruitment of CRISPR-Cas systems by Tn7-like transposons. *Proc Natl Acad Sci USA* **114**: E7358–E7366.

Pietilä, M.K., Demina, T.A., Atanasova, N.S., Oksanen, H. M., and Bamford, D.H. (2014) Archaeal viruses and bacteriophages: comparisons and contrasts. *Trends Microbiol* **22**: 334–344.

Prangishvili, D., Bamford, D.H., Forterre, P., Iranzo, J., Koonin, E.V., and Krupovic, M. (2017) The enigmatic archaeal virosphere. *Nat Rev Microbiol* **15**: 724–739.

Prangishvili, D., Albers, S.V., Holz, I., Arnold, H.P., Stedman, K., Klein, T., *et al*. (1998) Conjugation in archaea: frequent occurrence of conjugative plasmids in *Sulfolobus*. *Plasmid* **40**: 190–202.

Redrejo-Rodríguez, M., Ordóñez, C.D., Berjón-Otero, M., Moreno-González, J., Aparicio-Maldonado, C., Forterre, P., *et al*. (2017) Primer-independent DNA synthesis by a family B DNA polymerase from self-replicating Mobile genetic elements. *Cell Rep* **21**: 1574–1587.

Rensen, E., Krupovic, M., and Prangishvili, D. (2015) Mysterious hexagonal pyramids on the surface of *Pyrobaculum* cells. *Biochimie* **118**: 365–367.

Rice, G., Tang, L., Stedman, K., Roberto, F., Spuhler, J., Gillitzer, E., *et al*. (2004) The structure of a thermophilic archaeal virus shows a double-stranded DNA viral capsid type that spans all domains of life. *Proc Natl Acad Sci USA* **101**: 7716–7720.

Roitman, S., Hornung, E., Flores-Uribe, J., Sharon, I., Feussner, I., and Beja, O. (2018) Cyanophage-encoded lipid desaturases: oceanic distribution, diversity and function. *ISME J* **12**: 343–355.

Roux, S., Brum, J.R., Dutilh, B.E., Sunagawa, S., Duhaime, M.B., Loy, A., *et al*. (2016) Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537**: 689–693.

Santoro, A.E., Dupont, C.L., Richter, R.A., Craig, M.T., Carini, P., McIlvin, M.R., *et al*. (2015) Genomic and proteomic characterization of "*Candidatus* Nitrosopelagicus brevis": an ammonia-oxidizing archaeon from the open ocean. *Proc Natl Acad Sci USA* **112**: 1173–1178.

Schleper, C., and Nicol, G.W. (2010) Ammonia-oxidising archaea--physiology, ecology and evolution. *Adv Microb Physiol* **57**: 1–41.

Seed, K.D., Lazinski, D.W., Calderwood, S.B., and Camilli, A. (2013) A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. *Nature* **494**: 489–491.

Shah, S.A., Vestergaard, G., and Garrett, R.A. (2012)CRISPR/Cas and CRISPR/Cmr immune systems of Archaea . In *Regulatory RNAs in Prokaryotes*, Hess, W.R., and Marchfelder, A. (eds). Vienna: Springer-Verlag, pp. 163–181.

Sharon, I., Alperovitch, A., Rohwer, F., Haynes, M., Glaser, F., Atamna-Ismaeel, N., *et al*. (2009) Photosystem I gene cassettes are present in marine virus genomes. *Nature* **461**: 258–262.

She, Q., Brugger, K., and Chen, L. (2002) Archaeal integrative genetic elements and their impact on genome evolution. *Res Microbiol* **153**: 325–332.

She, Q., Shen, B., and Chen, L. (2004) Archaeal integrases and mechanisms of gene capture. *Biochem Soc Trans* **32**: 222–226.

Shiimori, M., Garrett, S.C., Graveley, B.R., and Terns, M.P. (2018) Cas4 nucleases define the PAM, length, and orientation of DNA fragments integrated at CRISPR loci. *Mol Cell* **70**: e6.

Shmakov, S., Abudayyeh, O.O., Makarova, K.S., Wolf, Y.I., Gootenberg, J.S., Semenova, E., *et al*. (2015) Discovery and functional characterization of diverse class 2 CRISPR-Cas systems. *Mol Cell* **60**: 385–397.

Snyder, J.C., Bolduc, B., and Young, M.J. (2015) 40 years of archaeal virology: expanding viral diversity. *Virology* **479-480**: 369–378.

Spang, A., Poehlein, A., Offre, P., Zumbragel, S., Haider, S., Rychlik, N., *et al*. (2012) The genome of the ammonia-oxidizing *Candidatus* Nitrososphaera gargensis: insights into metabolic versatility and environmental adaptations. *Environ Microbiol* **14**: 3122–3145.

Stahl, D.A., and de la Torre, J.R. (2012) Physiology and diversity of ammonia-oxidizing archaea. *Annu Rev Microbiol* **66**: 83–101.

Sullivan, M.J., Petty, N.K., and Beatson, S.A. (2011) Easyfig: a genome comparison visualizer. *Bioinformatics* **27**: 1009–1010.

Thompson, L.R., Zeng, Q., Kelly, L., Huang, K.H., Singer, A. U., Stubbe, J., and Chisholm, S.W. (2011) Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. *Proc Natl Acad Sci USA* **108**: E757–E764.

Tourna, M., Stieglmeier, M., Spang, A., Konneke, M., Schintlmeister, A., Urich, T., *et al*. (2011) *Nitrososphaera viennensis*, an ammonia oxidizing archaeon from soil. *Proc Natl Acad Sci USA* **108**: 8420–8425.

Varani, A.M., Siguier, P., Gourbeyre, E., Charneau, V., and Chandler, M. (2011) ISsaga is an ensemble of web-based methods for high throughput identification and semi-automatic annotation of insertion sequences in prokaryotic genomes. *Genome Biol* **12**: R30.

Vik, D.R., Roux, S., Brum, J.R., Bolduc, B., Emerson, J.B., Padilla, C.C., *et al*. (2017) Putative archaeal viruses from the mesopelagic ocean. *PeerJ* **5**: e3428.

Vollmer, A.C., and Bark, S.J. (2018) Twenty-five years of investigating the universal stress protein: function, structure, and applications. *Adv Appl Microbiol* **102**: 1–36.

Wallden, K., Rivera-Calzada, A., and Waksman, G. (2010) Type IV secretion systems: versatility and diversity in function. *Cell Microbiol* **12**: 1203–1212.

Wang, H., Peng, N., Shah, S.A., Huang, L., and She, Q. (2015) Archaeal extrachromosomal genetic elements. *Microbiol Mol Biol Rev* **79**: 117–152.

Wang, J., Liu, Y., Du, K., Xu, S., Wang, Y., Krupovic, M., and Chen, X. (2018a) A novel family of tyrosine integrases encoded by the temperate pleolipovirus SNJ2. *Nucleic Acids Res* **46**: 2521–2536.

Wang, Y., Chen, B., Cao, M., Sima, L., Prangishvili, D., Chen, X., and Krupovic, M. (2018b) Rolling-circle replication initiation protein of haloarchaeal sphaerolipovirus SNJ1 is homologous to bacterial transposases of the IS91 family insertion sequences. *J Gen Virol* **99**: 416–421.

Williams, K.P. (2002) Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. *Nucleic Acids Res* **30**: 866–875.

Wuchter, C., Abbas, B., Coolen, M.J., Herfort, L., van Bleijswijk, J., Timmers, P., *et al*. (2006) Archaeal nitrification in the ocean. *Proc Natl Acad Sci USA* **103**: 12317–12322.

Yan, W.X., Chong, S., Zhang, H., Makarova, K.S., Koonin, E.V., Cheng, D.R., and Scott, D.A. (2018) Cas13d is a compact RNA-targeting type VI CRISPR effector positively modulated by a WYL-domain-containing accessory protein. *Mol Cell* **70**: e5.

Yutin, N., Backstrom, D., Ettema, T.J.G., Krupovic, M., and Koonin, E.V. (2018) Vast diversity of prokaryotic virus genomes encoding double jelly-roll major capsid proteins uncovered by genomic and metagenomic sequence analysis. *Virol J* **15**: 67.

Zhalnina, K.V., Dias, R., Leonard, M.T., Dorr de Quadros, P., Camargo, F.A., Drew, J.C., *et al*. (2014) Genome sequence of *Candidatus* Nitrososphaera evergladensis from group I.1b enriched from Everglades soil reveals novel genomic features of the ammonia-oxidizing archaea. *PLoS One* **9**: e101648.

Zhao, S., and Williams, K.P. (2002) Integrative genetic element that reverses the usual target gene orientation. *J Bacteriol* **184**: 859–860.

Zheng, Z., Bao, M., Wu, F., Chen, J., and Deng, X. (2016) Predominance of single prophage carrying a CRISPR/cas system in "*Candidatus* Liberibacter asiaticus" strains in southern China. *PLoS One* **11**: e0146422.

Zimmermann, L., Stephens, A., Nam, S.Z., Rau, D., Kubler, J., Lozajic, M., *et al*. (2018) A completely Reimplemented MPI bioinformatics toolkit with a new HHpred server at its Core. *J Mol Biol* **430**: 2237–2243.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Appendix S1. Supporting information
Table S1. Supporting information
Table S2. Supporting information
Table S3. Supporting information

CHAPTER VI

Avoidance of Trinucleotide Corresponding to Consensus Protospacer Adjacent Motif Controls the Efficiency of Prespacer Selection during Primed Adaptation.

Introduction:

Analysis of spacers from the CRISPRome data in Chapters I-IV revealed some sequence features discussed in the Annex. One of the possible mechanisms underlying the spacer-specific features is described in Chapter VI.

Contribution:

I obtained preliminary results of PAM avoidance in spacer sequences during primed adaptation experiments with different plasmids (the data was later reanalyzed by the second author), in spacers from sequenced genomes (Figure 4C) and lack of PAM avoidance in the mammoth CRISPRome data.

AMERICAN SOCIETY FOR MICROBIOLOGY | mBio®

RESEARCH ARTICLE
Molecular Biology and Physiology

Avoidance of Trinucleotide Corresponding to Consensus Protospacer Adjacent Motif Controls the Efficiency of Prespacer Selection during Primed Adaptation

Olga Musharova,[a,b] Danylo Vyhovskyi,[a] Sofia Medvedeva,[a] Jelena Guzina,[c] Yulia Zhitnyuk,[a] Marko Djordjevic,[c] Konstantin Severinov,[a,b,d] Ekaterina Savitskaya[a,b]

[a]Center for Life Sciences, Skolkovo Institute of Science and Technology, Moscow, Russia
[b]Institute of Molecular Genetics, Russian Academy of Sciences, Moscow, Russia
[c]Institute of Physiology and Biochemistry, Faculty of Biology, University of Belgrade, Belgrade, Serbia
[d]Waksman Institute, Rutgers, The State University of New Jersey, Piscataway, New Jersey, USA

ABSTRACT CRISPR DNA arrays of unique spacers separated by identical repeats ensure prokaryotic immunity through specific targeting of foreign nucleic acids complementary to spacers. New spacers are acquired into a CRISPR array in a process of CRISPR adaptation. Selection of foreign DNA fragments to be integrated into CRISPR arrays relies on PAM (protospacer adjacent motif) recognition, as only those spacers will be functional against invaders. However, acquisition of different PAM-associated spacers proceeds with markedly different efficiency from the same DNA. Here, we used a combination of bioinformatics and experimental approaches to understand factors affecting the efficiency of acquisition of spacers by the *Escherichia coli* type I-E CRISPR-Cas system, for which two modes of CRISPR adaptation have been described: naive and primed. We found that during primed adaptation, efficiency of spacer acquisition is strongly negatively affected by the presence of an AAG trinucleotide—a consensus PAM—within the sequence being selected. No such trend is observed during naive adaptation. The results are consistent with a unidirectional spacer selection process during primed adaptation and provide a specific signature for identification of spacers acquired through primed adaptation in natural populations.

IMPORTANCE Adaptive immunity of prokaryotes depends on acquisition of foreign DNA fragments into CRISPR arrays as spacers followed by destruction of foreign DNA by CRISPR interference machinery. Different fragments are acquired into CRISPR arrays with widely different efficiencies, but the factors responsible are not known. We analyzed the frequency of spacers acquired during primed adaptation in an *E. coli* CRISPR array and found that AAG motif was depleted from highly acquired spacers. AAG is also a consensus protospacer adjacent motif (PAM) that must be present upstream from the target of the CRISPR spacer for its efficient destruction by the interference machinery. These results are important because they provide new information on the mechanism of primed spacer acquisition. They add to other previous evidence in the field that pointed out to a "directionality" in the capture of new spacers. Our data strongly suggest that the recognition of an AAG PAM by the interference machinery components prior to spacer capture occludes downstream AAG sequences, thus preventing their recognition by the adaptation machinery.

KEYWORDS CRISPR spacers, CRISPR-Cas, naïve adaptation, primed adaptation

Prokaryotic CRISPR-Cas systems consisting of CRISPR arrays containing identical repeats separated by unique spacers and associated *cas* genes protect cells from invading nucleic acids (1–3). CRISPR-Cas systems function by first acquiring fragments

Received 8 October 2018 **Accepted** 29 October 2018 **Published** 4 December 2018

Citation Musharova O, Vyhovskyi D, Medvedeva S, Guzina J, Zhitnyuk Y, Djordjevic M, Severinov K, Savitskaya E. 2018. Avoidance of trinucleotide corresponding to consensus protospacer adjacent motif controls the efficiency of prespacer selection during primed adaptation. mBio 9:e02169-18. https://doi.org/10.1128/mBio.02169-18.

Editor Alexander Idnurm, University of Melbourne

Address correspondence to Konstantin Severinov, severik@waksman.rutgers.edu.

O.M. and D.V. contributed equally to this article.

This article is a direct contribution from a Fellow of the American Academy of Microbiology. Solicited external reviewers: Luciano Marraffini, Rockefeller University; Blake Widenheft, University of Montana.

of invading nucleic acids, prespacers, and integrating them into CRISPR arrays as spacers, thus forming hereditable immunological memory (4). DNA of genetic invaders containing "memorized" fragments is recognized by Cas protein complexes and spacer-containing CRISPR RNAs (crRNAs) and targeted for destruction in a process called CRISPR interference (5). The recognition is achieved through complementary interaction between crRNA spacer and the target sequence, named the protospacer, and is also dependent on a specific short protospacer adjacent motif (PAM) (6–10).

CRISPR-Cas systems developed diverse mechanisms to avoid autoimmunity that should arise from targeting spacers in CRISPR array. Most of these mechanisms are based on a requirement for PAM, which is not complementary to crRNA but is specifically recognized by Cas proteins from the interfering complex (11, 12). The PAM is absent from the CRISPR repeat sequence adjoining the spacer. The separation of CRISPR defense into spacer acquisition and target interference stages and the requirement for PAM means that new spacers need to arise from sequences (prespacers) associated with PAM. Otherwise, they will not be able to perform their protective function.

For a well-studied type I-E CRISPR-Cas system from *Escherichia coli*, two modes of spacer adaptation have been described (13–15). The naive adaptation requires the Cas1 and Cas2 proteins and a CRISPR array (15). About 40% of spacers acquired during the naive adaptation arise from prespacers associated with the consensus AAG PAM; the majority of other acquired spacers are not expected to be functional in interference (15). In addition to Cas1 and Cas2, primed adaptation requires all the components of the interference stage: in *E. coli* they are the complex Cascade, the Cas3 nuclease-helicase, and a crRNA, which recognizes foreign DNA (13). Primed adaptation is much more efficient than naive adaptation, and almost 100% of prespacers chosen contain a consensus AAG PAM (16). The requirement for specific crRNA indicates that primed adaptation is triggered by the recognition of the target by the Cascade-crRNA effector complex. The site of recognition is referred to as a "priming protospacer." Upon target recognition by the effector complex, localized melting of the protospacer occurs. Melting initiates close to the PAM, in the so called "seed" region of the protospacer, and then extends further downstream (17). One protospacer strand, referred to as the "target strand," forms a heteroduplex with crRNA spacer sequence. The other, nontarget, strand is displaced, forming an R-loop. A specific feature of primed adaptation is a very strong strand bias in the orientation of selected prespacers (13, 14, 16). Upstream of the priming site, more than 90% of prespacers are oriented the same way as the priming protospacer: i.e., they map on the nontarget strand. The orientation of downstream prespacers is an opposite one: i.e., they map to the target strand. The efficiency of prespacer acquisition decreases with increasing distance from the priming site (18). No such biases are apparent during naive adaptation, and acquired spacers map to both strands of foreign DNA. It was shown that naive adaptation is affected by RecBCD activity, and acquired spacers tend to originate from regions with double-stranded breaks or replication fork stalling (19, 20).

While the presence of an AAG PAM at a prespacer side is strictly required for its selection by the adaptation machinery during primed adaptation and makes a strong contribution during naive adaptation, it alone does not determine the efficiency of prespacer usage (21, 22). Thus, in an *E. coli* culture undergoing primed adaptation of spacers from a plasmid, it is commonly observed that certain prespacers with an AAG PAM are acquired by many cells, while others are acquired rarely or not at all (22). The former are referred to as "hot" prespacers, while the latter are "cold." The pattern of hot and cold prespacers and their relative efficiencies are highly reproducible. The reasons behind the observed differential use of prespacers during adaptation are not known. In this work, we performed bioinformatics and experimental analysis that led us to conclude that a presence of an AAG trinucleotide within the prespacer has a strong negative effect on the frequency of its use during primed adaptation.

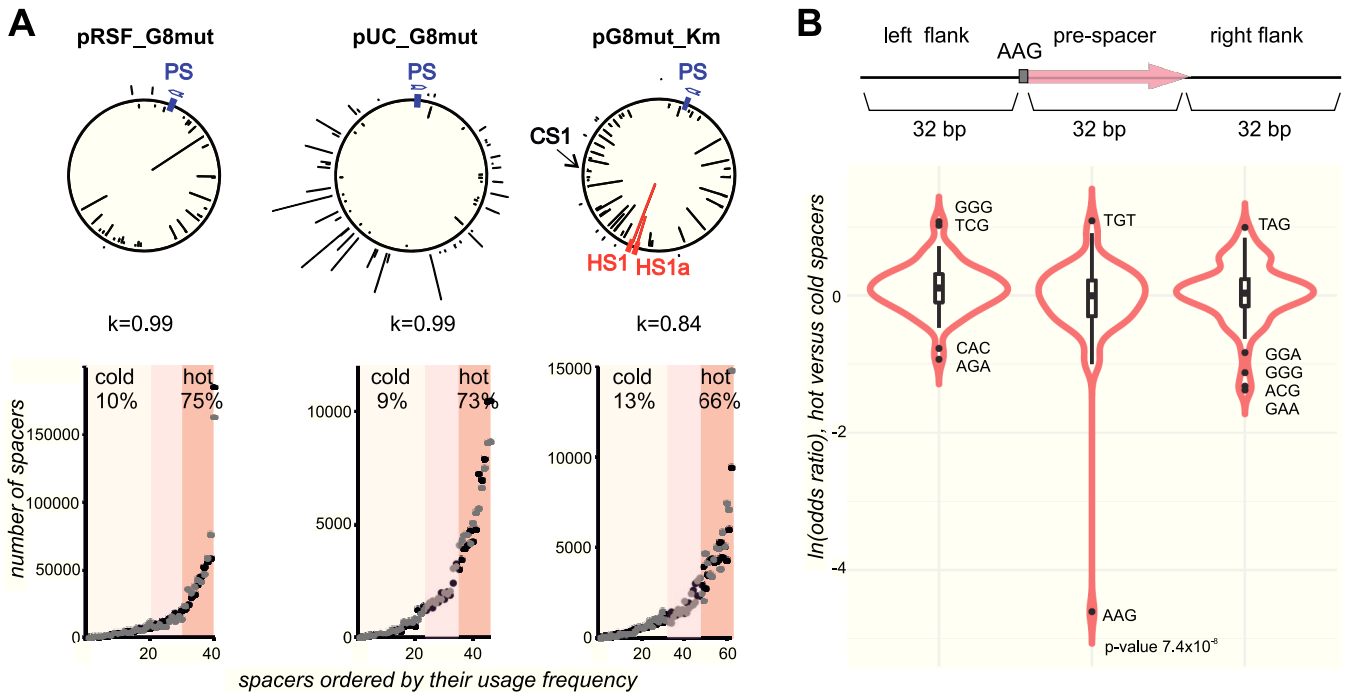


FIG 1 Prespacers actively used during primed adaptation are depleted in the AAG trinucleotide. (A) At the top, a graphical representation of spacers acquired in the course of primed adaptation from plasmids pRSF_G8mut, pUC_G8mut, and pG8mut_Km is presented. The position of the priming protospacer G8 (PS) in each plasmid is indicated by a blue rectangle. Arrows indicate the orientation of the priming protospacer (same in pRSF_G8mut and pG8mut_Km and opposite in pUC_G8mut). Spacers acquired from each plasmid are shown by black lines, with line heights indicating relative frequency of reads corresponding to different spacers. Lines projecting inside and outside the plasmid circles represent spacers mapping on opposite strands of plasmid DNA. Spacers originating from hot spot 1 (HS1) and HS1a prespacers (see the text for details) are highlighted in red. "CS1" shows the position of the cold prespacer (see the text for details). Below, Pearson correlation coefficients for mapping of spacers acquired from each plasmid in two independent experiments are given. At the bottom, spacers acquired from each plasmid were ranked according to their occurrence in Illumina reads. Each dot represents one spacer (corresponding to lines protruding from plasmid maps at the top). Dots colored black and gray represent results from two independent experiments. Spacers in the lower half of the distribution were considered cold. The top 25% of most common spacers were considered hot. The mean total percentage of cold and hot spacers from two experiments for each plasmid is given. (B) Violin plots showing odds ratio of trinucleotides in hot versus cold prespacers and their flanking sequences. The *P* value for AAG depletion in hot prespacers is shown.

RESULTS

Spacers efficiently acquired during primed adaptation have distinct nucleotide composition. To reveal possible causes of unequal acquisition efficiency of prespacers during primed adaptation, previously reported data sets of spacers acquired by *E. coli* KD263 cells transformed with plasmids pRSF_G8mut and pUC_G8mut (23, 24) were analyzed (see Table S1 in the supplemental material). In addition, new data sets of spacers acquired by KD263 cells in the presence of pG8mut_Km plasmid (Materials and Methods) were used. In each case, adaptation was initiated from a plasmid-borne G8mut priming protospacer partially matching the spacer segment of KD263 crRNA. The backbones of pRSF_G8mut, pUC_G8mut, and pG8mut-Km are sufficiently different so that most spacers of each data set do not overlap. For each sample, data sets corresponding to two biological replicates were analyzed. As expected for primed adaptation, most spacers in each culture were acquired from plasmid (99.7%) rather than the bacterial genome, and 86.35% of plasmid-derived spacers mapped to the DNA strand that was not targeted by G8 crRNA (Fig. 1A; Table S1). A total of 98.4% of plasmid spacers originated from prespacers preceded by an AAG PAM. The distribution of frequencies of spacers was highly reproducible for each plasmid, with a Pearson correlation of 0.84 or higher. While it has been observed that regions proximal to a priming protospacer preferentially donate new spacers during primed adaptation (18, 25–27), there was no gradient in prespacer usage with any of the plasmids (Fig. 1A), likely due to their small size.

For each plasmid, sequences of unique spacers derived from the nontarget strand and associated with AAG PAM were sorted according to spacer frequency in the data

set. The resulting frequency distributions for each plasmid are shown in Fig. 1A. As can be seen, the distributions are highly unequal, with some spacers being used much more frequently than others. We consider the 25% of most frequently used spacers as "hot." Conversely, 50% of spacers at the opposite end of the distribution are considered "cold." Together, sequences from the hot spacer group account for ~70% of all plasmid-borne spacers, while cold spacer group sequences account for ~10% of spacers. For subsequent analysis, unique hot and cold group spacers from each data set were combined and treated together.

No difference in nucleotide composition of "cold" and "hot" spacers was revealed. Dinucleotide frequency analysis was likewise uninformative (data not shown). Strikingly, analysis of trinucleotide frequencies showed that the AAG triplet was strongly underrepresented in the hot group (Fig. 1B) ($P = 7.4 \times 10^{-8}$).

We also considered whether sequences flanking plasmid prespacers have an effect on prespacer acquisition frequency during primed adaptation. Spacer-sized 33-bp regions upstream of AAG PAMs or downstream of "hot" and "cold" prespacers were also analyzed, but no strong bias was detected in either base composition or di/trinucleotide frequencies (see Fig. 1B for trinucleotide frequency).

The presence of the AAG trinucleotide within a prespacer controls the efficiency of its use as a donor of spacers during primed adaptation. To experimentally measure the contribution of nucleotide sequence to spacer acquisition efficiency, we studied the effects of sequence alterations in HS1 (hot spot 1), one of the most commonly used hot prespacers from the pG8mut-Km plasmid (Fig. 1A). The acquisition of this prespacer was analyzed previously, and it was shown that its usage depends on the AAG PAM (22). Six pG8mut-Km plasmid libraries containing randomized trinucleotides at HS1 positions 2 to 4, 5 to 7, 14 to 16, 20 to 22, 28 to 30, and 31 to 33 were prepared. Each library was transformed in uninduced KD263 cells, and pooled transformants were subjected to PCR with a pair of primers annealing upstream and downstream of plasmid region spanning the HS1 prespacer (Fig. 2A). Analysis of Illumina reads from obtained amplicons revealed that for each library, all 64 expected variants were present.

For each library, several thousand transformants were pooled and grown in the presence of inducers of *cas* gene expression in the absence of antibiotic. These conditions stimulate primed adaptation from the plasmid without selecting against cells that acquired interference-proficient spacers targeting the plasmid. Amplified DNA fragments corresponding to the expanded CRISPR array in cultures harboring each plasmid library were subjected to Illumina sequencing, and acquired spacers were analyzed. The overall pattern of plasmid-derived new spacers was the same in each library and matched the one observed for unmodified pG8mut-Km (Fig. 2B). The only exception were spacers corresponding to HS1, whose cumulative efficiency of adaptation decreased in the libraries compared to unmodified pG8mut-Km. Sequences of acquired spacers matching HS1 and its variants were extracted, and odds ratios between frequency of spacer variants and prespacer variants in corresponding libraries were determined. As can be seen from results presented in Fig. 2C, HS1 spacer variants with the AAG trinucleotide in the seed region (positions 2 to 4 and 5 to 7) were strongly underrepresented. The effect was much weaker at positions 14 to 16, 20 to 22, 28 to 30, and 31 to 33. We conclude that the library approach supports the bioinformatics analysis that shows that the presence of internal AAG inhibits prespacer usage during primed adaptation. The results also show that the effect is position specific and is most evident when the AAG trinucleotide is located in the seed of the future spacer.

Given the observed position specificity of library data, we reanalyzed hot spacers from the combined plasmid set (Fig. 1B) using a 6-base sliding window and concentrating on comparison of the 10% "hottest" and "coldest" spacers. The results, presented in Fig. 2D, confirmed the avoidance of AAG in the seed region of these spacers. The remaining positions exhibited a bias of marginal statistical significance, while no bias was observed in spacer-sized flanking sequences upstream or downstream of hot prespacers. The positional bias in AAG occurrence was also revealed using an inde-

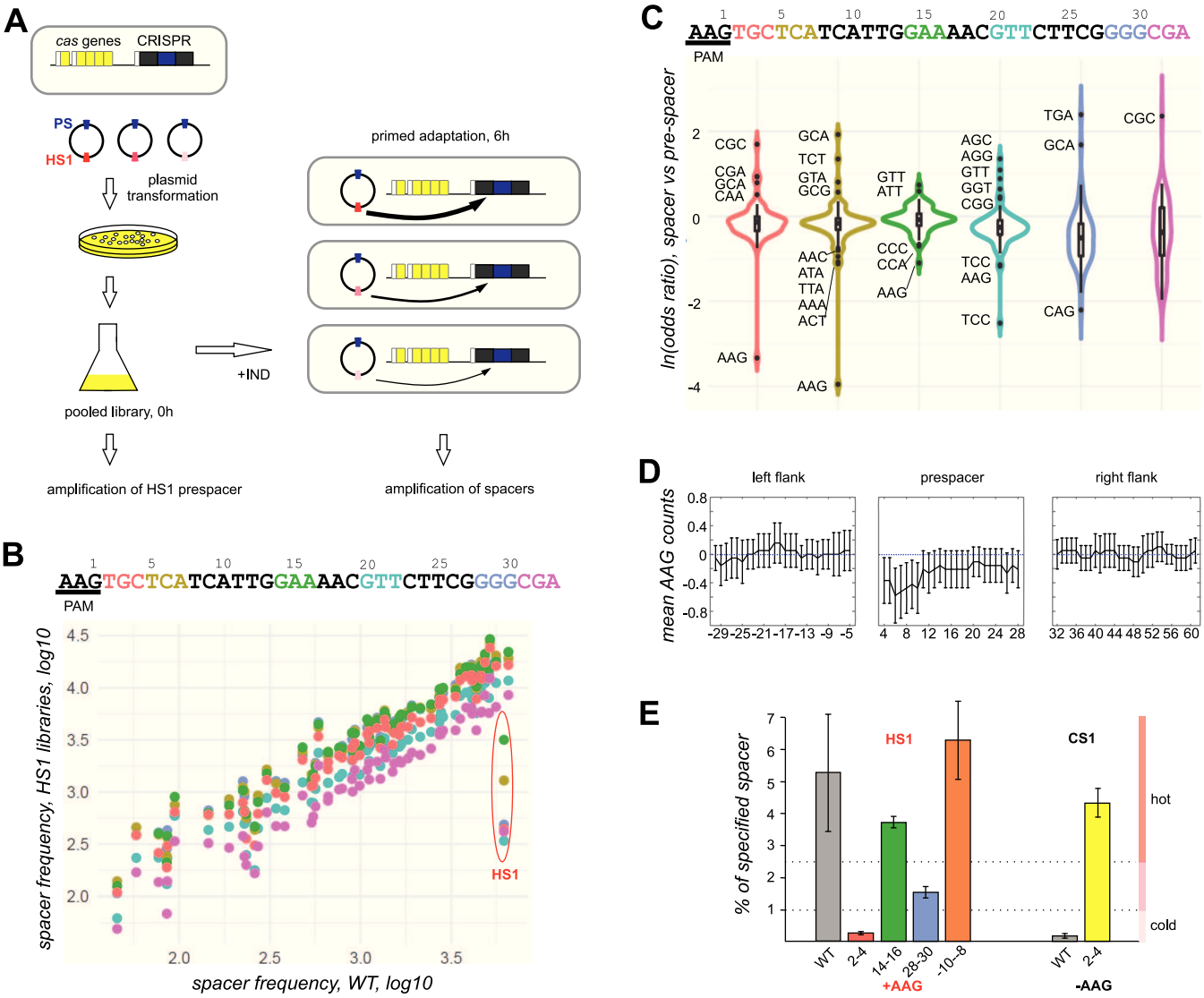


FIG 2 Experimental demonstration of position-specific AAG avoidance in hot prespacers during primed adaptation. (A) A workflow of the library-based approach to determine the effect of prespacer sequence on acquisition efficiency is presented. Engineered *E. coli* KD263 cells with inducible expression of *cas* genes and a CRISPR array with a single G8 spacer are transformed with a library of plasmids containing the G8 priming protospacer (blue) and randomized trinucleotides in the HS1 prespacer (shown by different hues of red); white rectangles represent promoter regions of *cas* genes and the CRISPR array. Transformants grown on selective medium are pooled and placed in a medium without antibiotic required for plasmid maintenance. The cultures are induced and grown for 6 h to allow primed adaptation to occur. In the pooled culture before induction, the HS1-containing region is amplified and subjected to Illumina sequencing. In the induced culture, the CRISPR array is amplified, and amplicon corresponding to expanded array is subjected to Illumina sequencing. (B) At the top, the sequence of the HS1 prespacer and its PAM is shown. Trinucleotides subjected to randomization in six different libraries are indicated by colors. Below, the frequency of spacers acquired by cells carrying each library is compared to the frequency of spacer acquisition in the initial plasmid (WT). Each dot represents a spacer, and the color of the dot corresponds to the color of the randomized trinucleotide. Dots corresponding to HS spacer and its variants are indicated. (C) Violin plots showing odds ratio of trinucleotides in HS1-derived spacers compared to prespacers in each library. (D) The left, middle, and right plots correspond, respectively, to 33 bp of upstream prespacer flank, the prespacer sequence, and the downstream prespacer flank. Coordinates on the *x* axis correspond to the center of the 6-bp sliding window, where +1 corresponds to G in AAG PAM. The difference between mean AAG counts in hot and cold prespacer categories is shown in the *y* axis. The error bars correspond to 95% confidence intervals. (E) Acquisition of HS1 and CS1 spacer variants from individual plasmids carrying trinucleotide substitutions. The bars show the percentage of HS1 and its variants and CS1 and its variant to overall plasmid-derived spacers acquired by cells carrying wild-type pG8mut_Km (WT) or derivatives carrying AAG trinucleotides at specified positions of HS1 or carrying an AAC trinucleotide instead of AAG at positions 2 to 4 of the CS1 prespacer. Mean values obtained from two independent experiments and standard deviations are given.

pendent approach, by analyzing the entire spacer set and correlating AAG counts in different prespacer regions and the corresponding spacer frequencies (see Fig. S1 in the supplemental material).

To directly demonstrate that the presence of AAG trinucleotide affects prespacer acquisition, individual plasmids containing AAG at HS1 positions 2 to 4, 14 to 16, and 28 to 30 were constructed and used in a primed adaptation experiment. Analysis of

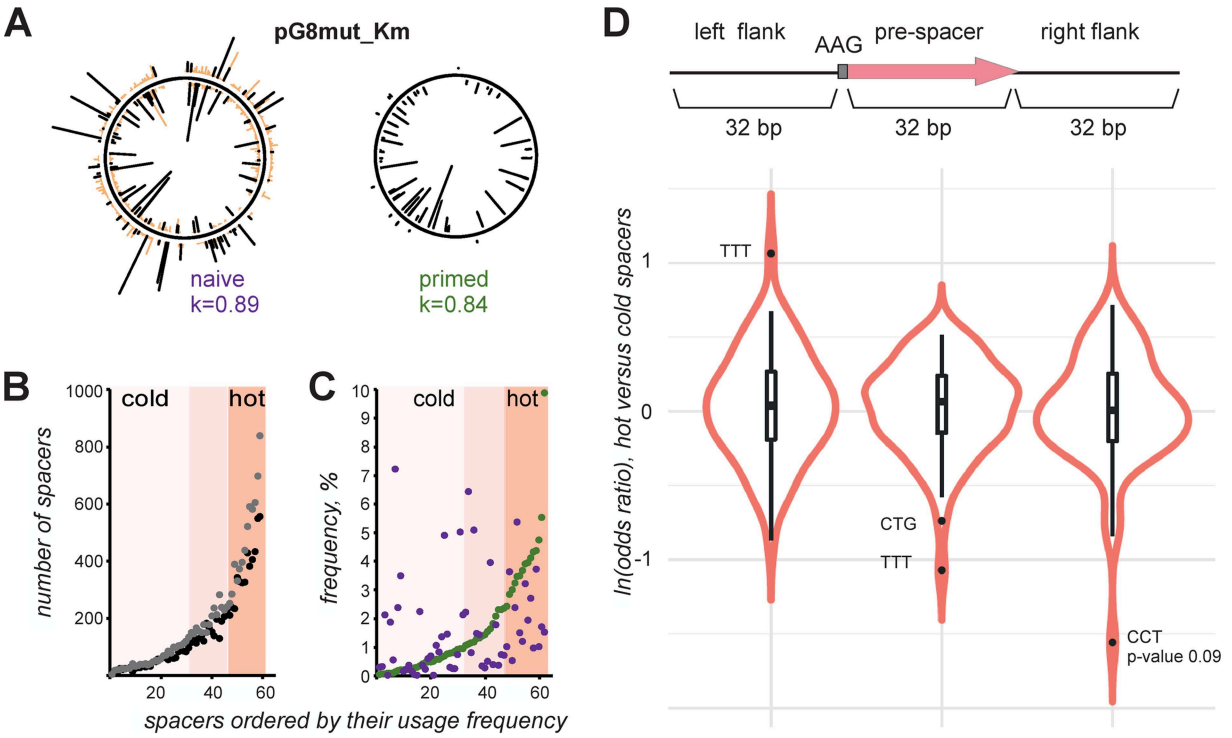


FIG 3 Comparison of prespacers acquired during naive and primed adaptation. (A) At the top, a graphical representation of spacers acquired in the course of naive (left) and primed (right) adaptation from the pG8mut_Km plasmid is presented. See the legend to Fig. 1A for details. For naive adaptation, spacers mapping to prespacers with the AAG PAM are shown by black lines. Spacers mapping to prespacers with non-AAG PAMs are marked in orange. (B) Spacers acquired during naive adaptation (A) that mapped to prespacers with the AAG PAM and the "inner" strand of plasmid DNA were ranked according to their occurrence in Illumina reads. Each dot represents one spacer (which corresponds to lines protruding from the plasmid map in panel A, left). Dots colored black and gray represent results from two independent experiments. Spacers in the lower half of the distribution were considered cold. The top 25% of most common spacers were considered hot. (C) Spacers acquired from pG8mut_Km in the course of primed adaptation were ranked as in Fig. 1A: each spacer is represented by a green dot. The frequency of corresponding spacers acquired in the course of naive adaptation is represented by dark violet dots. (D) Violin plots showing odds ratio of trinucleotides in hot versus cold prespacers and their flanking sequences from the naive adaptation experiment.

spacers acquired by cells carrying these plasmids revealed that compared to pG8mut-Km, the presence of AAG at positions 2 to 4 decreased the number of HS1-derived spacers more than 10 times (Fig. 2E). Introduction of AAG at positions 14 to 16 and 28 to 30 had a milder, 2- to 3-fold effect. When an AAG trinucleotide was introduced 5 nucleotides upstream of HS1 PAM, no effect on HS1 spacer acquisition efficiency was detected.

We also determined whether removal of an AAG trinucleotide increases the usage of a cold prespacer. The pG8mut-Km prespacer CS1 (cold spot 1) contains an AAG at positions 2 to 4. When substituted for AAC, the use of this prespacer increased ~16-fold, placing it in a hot spacer group.

The presence of AAG trinucleotide has no effect on prespacer usage during naive adaptation. We were interested in comparing prespacer choice preferences during primed and naive adaptation. The "naive" spacer set was obtained by transforming the pG8mut-Km plasmid in *E. coli* BL21(DE3) cells carrying a compatible plasmid coexpressing the Cas1 and Cas2 proteins. BL21(DE3) lacks its own *cas* operon, and in the presence of pCas1 + 2 is only capable of naive adaptation (15). Mapping of spacers acquired in the BL21(DE3) CRISPR array from pG8mut-Km is shown in Fig. 3A (left-hand side). As expected, there was no strand bias and many spacers originated from prespacers without AAG PAM (see Table S2 in the supplemental material). The pattern of spacers acquired during naive adaptation (Fig. 3A, left-hand side) is highly reproducible (Pearson coefficient of 0.89) and distinct from the pattern of spacers acquired from pG8mut-Km during primed adaptation (shown on the right-hand side of Fig. 3A). To compare prespacer preferences during two modes of adaptation, we

concentrated on prespacers with an AAG PAM mapping to the "inner" strand of pG8mut-Km, as shown in Fig. 3A. The efficiency of usage of such prespacers during naive adaption (ranked according to increasing occurrence of spacers as in Fig. 1B) is shown in Fig. 3B for two independent experiments. On Fig. 3C, frequencies of spacers from the naive set are plotted alongside the ranked set of spacers acquired during primed adaptation. Visual inspection of data and statistical analysis show that there is no correlation between the two sets (Pearson correlation of 0.19; *P* value of 0.14). In other words, a spacer that scores as cold (or hot) during primed adaptation can be either cold or hot or have intermediate frequency during naive adaptation.

Since the sets of hot and cold spacers in naive and primed adaptation are distinct, we wondered if any specific sequence signatures can be revealed in spacers that were acquired during naive adaptation with different efficiencies. For this analysis, unique spacers acquired from pG8mut-Km and the pCas1 + 2 plasmid coexpressing *cas1* and *cas2* were combined into a single set and analyzed jointly. However, no specific signal for single nucleotides, dinucleotides, and trinucleotides was observed. Consistent with results shown in Fig. 3C, the frequency of spacers acquired from prespacers associated with AAG PAM during naive adaptation was not affected by the presence of the internal AAG trinucleotide (Fig. 3D). Similar to observations with primed adaptation, upstream and downstream flanking sequences contained no specific features.

DISCUSSION

Spacer diversity in CRISPR arrays from native bacterial strains is very high (28). Spacer selection is nonrandom, and strong and reproducible biases in acquired spacer repertoires were described for both naive and primed adaptation in laboratory experiments (16, 21, 29–32). While such biases can be produced by selection for spacers most efficient during CRISPR interference, preferences of the adaptation machinery must also play a role. Understanding the determinants of efficient spacer acquisition in the absence of selection may be useful for designing experiments in which adapted spacers are used to record cellular events in the absence of subsequent interference (29, 30). In this work, we compared the efficiency of prespacer selection by the *E. coli* type I-E CRISPR-Cas system during primed and naive adaptation in the absence of selection. Earlier analysis of efficiently acquired spacers during naive adaptation in this system revealed that actively used prespacers may contain motifs in their 3′ ends. However, these motifs appear to be mutually exclusive (AA at positions 32 and 33 according to Yosef et al. [21], compared to G at position 32 in the study by Shipman et al. [30]). In the case of primed adaptation by I-C and I-B CRISPR-Cas systems, it has been shown that nucleotide substitutions in the prespacer affect the efficiency of its use (31, 32). Overall, these earlier works show that prespacer sequence clearly contributes to its selection efficiency during adaptation. Our analysis failed to reveal determinants of prespacer naive adaptation efficiency. However, we observed very strong avoidance of AAG trinucleotide in spacers efficiently acquired during primed adaptation. The AAG trinucleotide is also the dominant (99.8%) PAM of prespacers that are acquired during primed adaptation. The complementary CTT trinucleotide is not avoided, which is consistent with a general view of primed adaptation that involves the recognition of the priming protospacer by the Cascade effector, followed by the recruitment of the Cas3 nuclease-helicase and its processive movement along the DNA away from the priming site in the 3′ to 5′ direction. Such directionality should allow discrimination between 5′-AAG-3′ and 5′-CTT-3′ sequences and will account for observed overall declining gradients of prespacer usage as the distance from the priming site increases.

A possible mechanistic basis of AAG avoidance in hot spacers is the competition between overlapping prespacers during spacer selection. We observed that for partially overlapping prespacers with AAG PAM, a prespacer located further away from the priming site has no effect on the use of prespacer located closer, while the reverse is not true (Fig. 4A). Such directionality is consistent with a view that the primed adaptation machinery slides in a 3′ direction from the priming site along the fully double-stranded DNA, occasionally recognizes an AAG trinucleotide, and then extracts

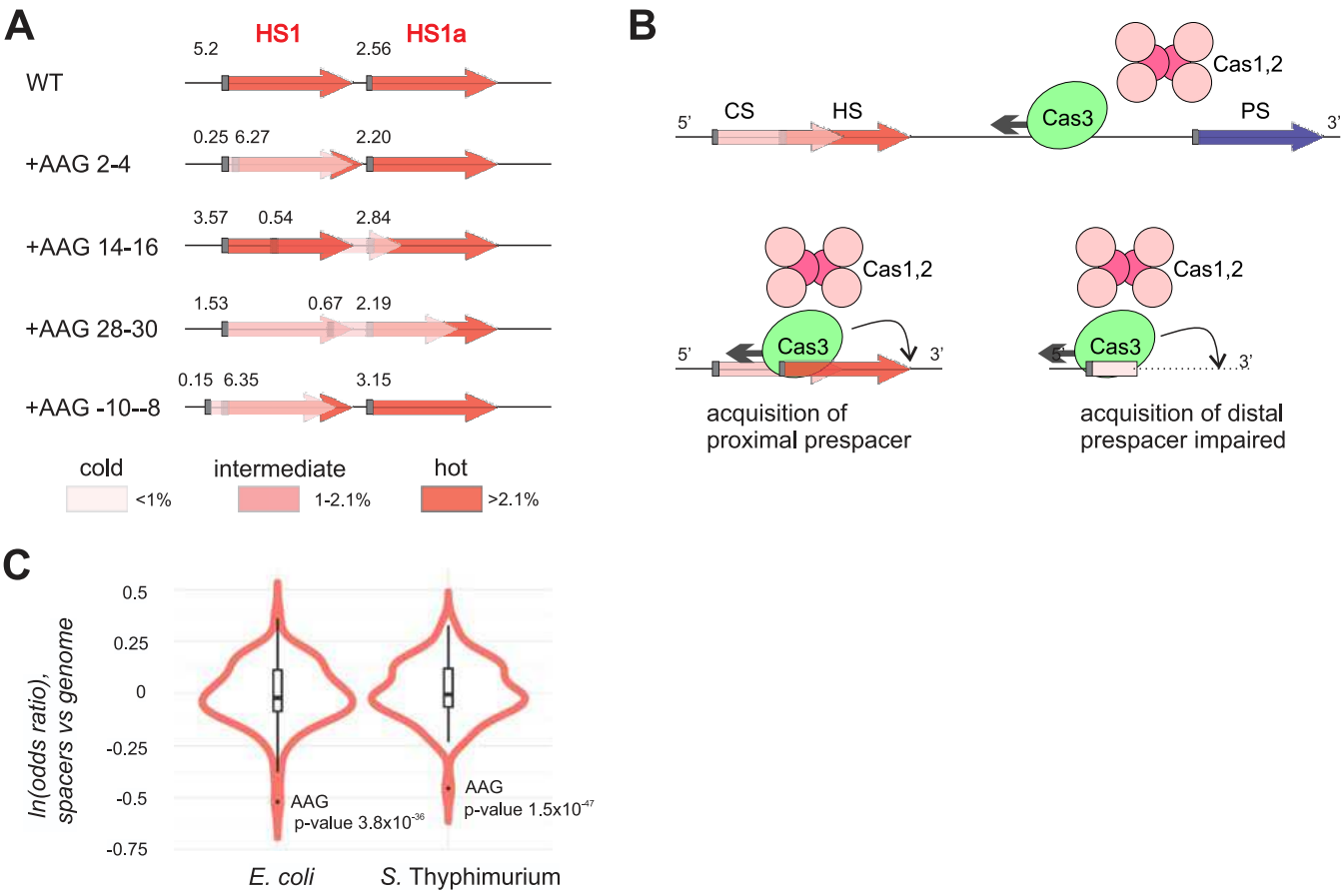


FIG 4 Interdependency of prespacer use during primed adaptation and a possible mechanism. (A) The scheme shows the relative percentages of spacers derived from HS1 and HS1a prespacers in experiments shown in Fig. 3E for cells transformed with plasmids carrying AAG trinucleotides at the indicated positions of HS1. Gray rectangles indicate AAG PAMs; numbers nearby depict the percentage of corresponding spacers (from averaging of two experimental replicas). The insertion of AAG into HS1 decreases its usage efficiency and gives rise to a new prespacer (Fig. 3E). The frequency of HS1a is unaffected by the introduction of the AAG PAM inside HS1 even if the new prespacer overlaps HS1a. The appearance of a new prespacer due to the introduction of a new AAG upstream of HS1 (+AAG −10 to −8) likewise has no effect on acquisition of HS1 spacers. (B) A model describing a mechanism that may account for observed interdependency of prespacer use is presented. Cas3 moves from the priming protospacer (PS) in a 3′ to 5′ direction. Upon encountering AAG trinucleotide, Cas1 and Cas2 use a ruler-like mechanism to extract a spacer in the backward direction. As a result, the efficiency of use of the overlapping prespacer located further downstream is decreased. (C) Violin plots showing the odds ratio of trinucleotides in spacers versus genome-wide frequency in fully sequenced *E. coli* and *S.* Typhimurium genomes.

a spacer-sized fragment located immediately upstream—i.e., opposite to the direction of lateral movement along the DNA (Fig. 4B). According to this model, one would expect that any internal AAG will have the same negative effect irrespective of its position within the prespacer. The unequal effects of AAG trinucleotides placed in the beginning, middle, and end regions of prespacers on adaptation efficiency revealed in our experiments, with much stronger inhibition produced by AAG located in PAM-proximal seed region, require a more sophisticated model and further experiments to explain.

Our results do not allow to distinguish whether interdependency of overlapping prespacer use is due to prespacer interaction with the adaptation machinery *sensu stricto* (i.e., the Cas1-Cas2 complex) or is determined at an earlier stage by Cas3, which may generate substrates for Cas1-Cas2 as it moves away from the priming site (22, 33). Data suggesting that Cas3 may specifically cleave at AAG PAMs have been presented. Also evidence for preferences for AAG PAMs by the Cas1-Cas2 complex both from structural data (34, 35) and analysis of spacers acquired during naive adaptation (15) is available. It is thus possible that Cas3 and Cas1-Cas2 cooperate with each other during primed adaptation, increasing the likelihood of selection of prespacers with AAG PAM, which should have the highest protective effect. The presence of Cas2-Cas3 fusions in

type I-F systems supports the idea of such synergy (36). For example, the observed negative effects of internal AAG sequences may be the consequence of Cas3 cleavage at these sites and hindering Cas1-Cas2 access to downstream DNA to begin spacer capture.

The absence or presence of internal AAG cannot be the only determinant of prespacer usage. The sampling frequencies of spacers in our set, which correspond to the same AAG counts in prespacers, differ by about 3 orders of magnitude (see Fig. S2 in the supplemental material). The coefficient of determination from the data presented in Fig. S2 shows that only ~25% of variability of spacer frequencies acquired during primed adaptation can be explained by the presence of internal AAGs. The rest of the variation must be determined by additional sequence or context-specific effects whose nature is currently unknown.

We used the avoidance of internal PAM signal to assess whether priming may have contributed to acquisition of spacers in natural isolates of *E. coli* and *Salmonella enterica* serovar Typhimurium. These two microorganisms contain a virtually identical type I-E CRISPR-Cas system with the same PAM and repeats, but share few common spacers. As can be seen from Fig. 4C, compared to overall genomic frequency, AAG is underrepresented in spacers from CRISPR arrays of fully sequenced *E. coli* and *S.* Typhimurium isolates, suggesting that priming occurs in natural settings in these bacteria.

MATERIALS AND METHODS

Strains and plasmids. The *E. coli* DH5α strain was used for cloning. The *E. coli* strain KD263 (K-12 F$^+$ *lacUV5-cas3 araBp8-cse1* CRISPR I repeat-spacer G8-repeat CRISPR II deleted) (37) and BL21(DE3) were used in primed and naive adaptation experiments, correspondingly.

In order to create the pG8mut_Km plasmid, a fragment of the pRSF1b plasmid (Novagen) containing a kanamycin resistance gene was amplified with primers kan-fragment forward and kan-fragment rev (see Table S3 in the supplemental material). The amplicon was purified, treated with the EcoRI and BamHI, and cloned into the pG8mut plasmid (23).

Library and individual mutant construction. Plasmid libraries with randomized trinucleotide in HS1 prespacer were obtained by a two-step PCR-based mutagenesis using iProof high-fidelity DNA polymerase (Bio-Rad). In the first step, pG8mut_Km was amplified with forward primer HSRun_for containing three randomized nucleotides inside the HS1 region and reverse primer HSRun_rev complementary to the constant region of HSRun_for. (The list of primers used in this work is presented in Table S3.) Twenty cycles of amplification were performed to generate linearized pG8mut_Km with randomized trinucleotides and short inverted repeats containing sequences of primer complementarity. Completed PCRs were treated with DpnI to eliminate the pG8mut_Km template, and reaction products were purified by the GeneJet PCR purification kit. At the second step, the products of the first amplification reactions were further amplified with primers HSRun_rev and HSRun_add, which contained regions complementary to inverted repeats introduced during the first stage. Five amplification cycles were performed. The products of amplification were purified as described above. Finally, a Gibson assembly cloning kit (New England Biolabs) was next used to generate circular plasmids through recombination between the inverted repeats following the manufacturer's recommendation. Using the procedure outlined above, six different libraries with randomized nucleotides at positions 2 to 4, 5 to 7, 13 to 15, 19 to 21, 28 to 30, and 31 to 33 of HS1 were generated. The results of Gibson assembly were transformed into DH5α cells by electroporation. At least 2,000 kanamycin-resistant colonies for each library were scraped off the plates and used for plasmid purification by GeneJet plasmid miniprep kit (Thermo Scientific).

Individual AAG trinucleotides were introduced in pG8mut_Km by a standard PCR-based site-specific mutagenesis protocol with primer pairs listed in Table S3.

CRISPR adaptation and plasmid prespacer and acquired spacer amplification. For primed adaptation, pG8mut_Km, its derivatives containing individual mutations, or plasmid libraries were electroporated into KD263. For library experiments, at least 2,000 kanamycin-resistant colonies were scrapped off plates for each library and pooled. The resulting cell suspension was diluted with LB to an optical density at 600 nm (OD_{600}) of 0.1 and allowed to grow at 37°C in the absence of antibiotic. In experiments with individual plasmids, a single colony was used to inoculate 5 ml LB supplemented with 50 μg/ml kanamycin. After overnight growth at 37°C, an aliquot of culture was diluted 100× with LB without antibiotic, and growth was continued. When cultures reached OD_{600}, they were induced by 1 mM arabinose and 1 mM IPTG (isopropyl-β-D-thiogalactopyranoside) at an OD of 0.4. The growth was continued for 6 h.

For naive adaptation, BL21(DE3) cells were electroporated with plasmids pCas1 + 2 (15) and pG8mut_Km. Individual colonies were grown overnight in liquid LB containing 50 μg/ml kanamycin and 50 μg/ml streptomycin. After overnight growth at 37°C, an aliquot of culture was diluted 100× with LB containing 50 μg/ml streptomycin and 0.1 mM IPTG. The growth was continued for 6 h.

Aliquots of cultures were withdrawn immediately before or 6 h postinduction, and total DNA was purified by a Thermo Scientific genomic DNA purification kit. To assess the diversity of HS1 prespacer

libraries, the corresponding plasmid region was amplified from 0-h total DNA samples using primers HS1long_for and HS1long_rev. To monitor CRISPR adaptation, CRISPR arrays were amplified from 6-h samples with primers Ec_LDR_F and M13_G8 for DNA from KD263 cultures and moj3-moj4 for BL21(DE3) cultures. Amplicons containing plasmid prespacers and extended CRISPR arrays were gel purified and used to create Illumina sequencing libraries with an NEBNext Ultra II DNA library preparation kit with U5 barcoding. High-throughput sequencing of amplicons was conducted on MiniSeq or HiSeq Illumina machines using the 2×150 paired-end mode.

Bioinformatics analysis. R script and Bioconductor packages ShortRead (38) and BioStrings (39) were utilized for Illumina reads preprocessing, prespacer and spacer extraction, mapping, and statistical analysis. R package ggplot2 (40) was used for plotting. The following parameters were used: FREDscore for read quality of $\geq$20, up to 2 mismatches for identification of CRISPR repeats or prespacer flanking regions, and 0 mismatches for mapping. Only uniquely mapped 33-bp-long spacers were taken for further analysis. Circular visualization of plasmid mapping results was done with EasyVisio tool developed by Ekaterina Rubtsova. Odds ratios for each mono-, di-, and trinucleotide were calculated based on Fisher's test. The odds ratios were calculated for prespacer libraries and acquired spacers or for hot and cold prespacers and/or their flanking sequences.

Spacers acquired during primed adaptation were mapped to the nontarget strand, and log values of their observed sampling frequencies (just sampling frequencies below) were used in the analysis. To decrease noise, the sampling frequencies of reads from different experiment replicas corresponding to same plasmids, which were mapped to same plasmid positions, were averaged. Sampling frequencies corresponding to different plasmids were then normalized to the same mean.

A window of 6 bp in length was slid across 33-bp prespacer sequences and the upstream and downstream prespacer flanking regions of the same length. For each window position, AAGs in the frame were counted, and their means for hot and cold categories (ν_h and ν_c, respectively) were subtracted. To estimate confidence bounds, it was assumed that the number of counts follows a Poisson distribution, so the standard deviation for the subtracted means was estimated to be $\sqrt{\nu_h+\nu_c}$.

To additionally assess significance of the AAG position within the prespacer, prespacers were divided into 3 nonoverlapping 11-bp-long regions (upstream, middle, and downstream). For each of these regions, Pearson's correlation coefficient (R) between the number of AAG counts and the corresponding spacer frequencies was calculated. Confidence bounds and P values for the obtained correlation coefficients were estimated through Fisher's z transformation.

To assess what fraction of variability in the spacer frequencies can be explained by AAG presence/absence, R between the number of AAG counts in the entire spacer and the corresponding spacer frequencies was calculated, from which the corresponding coefficient of determination (R^2) was obtained.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at https://doi.org/10.1128/mBio.02169-18.

FIG S1, TIF file, 14.5 MB.

FIG S2, TIF file, 1.8 MB.

TABLE S1, DOCX file, 0.1 MB.

TABLE S2, DOCX file, 0.1 MB.

TABLE S3, DOCX file, 0.1 MB.

ACKNOWLEDGMENTS

This study was supported by Russian Science Foundation grant 14-14-00988 and National Institutes of Health grant GM10407 to K.S. and Russian Foundation for Basic Research grant 16-04-00767 to E.S. O.M. was supported by Russian Foundation for Basic Research grant 18-34-00048. M.D. and J.G. are supported by the Ministry of Education and Science of the Republic of Serbia project ON173052.

REFERENCES

1. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P. 2007. CRISPR provides acquired resistance against viruses in prokaryotes. Science 315:1709–1712. https://doi.org/10.1126/science.1138140.
2. Brouns SJJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJH, Snijders APL, Dickman MJ, Makarova KS, Koonin EV, van der Oost J. 2008. Small CRISPR RNAs guide antiviral defense in prokaryotes. Science 321:960–964. https://doi.org/10.1126/science.1159689.
3. Marraffini LA, Sontheimer EJ. 2008. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. Science 322:1843–1845. https://doi.org/10.1126/science.1165771.
4. van der Oost J, Jore MM, Westra ER, Lundgren M, Brouns SJJ. 2009. CRISPR-based adaptive and heritable immunity in prokaryotes. Trends Biochem Sci 34:401–407. https://doi.org/10.1016/j.tibs.2009.05.002.
5. Marraffini LA, Sontheimer EJ. 2010. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. Nat Rev Genet 11:181–190. https://doi.org/10.1038/nrg2749.
6. Jore MM, Lundgren M, van Duijn E, Bultema JB, Westra ER, Waghmare SP, Wiedenheft B, Pul U, Wurm R, Wagner R, Beijer MR, Barendregt A, Zhou K, Snijders APL, Dickman MJ, Doudna JA, Boekema EJ, Heck AJR, van der Oost J, Brouns SJJ. 2011. Structural basis for CRISPR RNA-guided DNA recognition by Cascade. Nat Struct Mol Biol 18:529–536. https://doi.org/10.1038/nsmb.2019.
7. Szczelkun MD, Tikhomirova MS, Sinkunas T, Gasiunas G, Karvelis T,

Pschera P, Siksnys V, Seidel R. 2014. Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes. Proc Natl Acad Sci U S A 111:9798–9803. https://doi.org/10.1073/pnas.1402597111.
8. Wiedenheft B, Lander GC, Zhou K, Jore MM, Brouns SJJ, Van Der Oost J, Doudna JA, Nogales E. 2011. Structures of the RNA-guided surveillance complex from a bacterial immune system. Nature 477:486–489. https://doi.org/10.1038/nature10402.
9. Hale CR, Zhao P, Olson S, Duff MO, Graveley BR, Wells L, Terns RM, Terns MP. 2009. RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. Cell 139:945–956. https://doi.org/10.1016/j.cell.2009.07.040.
10. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. 2012. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science 337:816–822. https://doi.org/10.1126/science.1225829.
11. Mojica FJM, Díez-Villaseñor C, García-Martínez J, Almendros C. 2009. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. Microbiology 155:733–740.
12. Deveau H, Barrangou R, Garneau JE, Labonté J, Fremaux C, Boyaval P, Romero DA, Horvath P, Moineau S. 2008. Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. J Bacteriol 190:1390–1400. https://doi.org/10.1128/JB.01412-07.
13. Datsenko KA, Pougach K, Tikhonov A, Wanner BL, Severinov K, Semenova E. 2012. Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. Nat Commun 3:945. https://doi.org/10.1038/ncomms1937.
14. Swarts DC, Mosterd C, van Passel MWJ, Brouns SJJ. 2012. CRISPR interference directs strand specific spacer acquisition. PLoS One 7:e35888. https://doi.org/10.1371/journal.pone.0035888.
15. Yosef I, Goren MG, Qimron U. 2012. Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. Nucleic Acids Res 40:5569–5576. https://doi.org/10.1093/nar/gks216.
16. Savitskaya E, Semenova E, Dedkov V, Metlitskaya A, Severinov K. 2013. High-throughput analysis of type I-E CRISPR/Cas spacer acquisition in *E. coli*. RNA Biol 10:716–725. https://doi.org/10.4161/rna.24325.
17. Semenova E, Jore MM, Datsenko KA, Semenova A, Westra ER, Wanner B, van der Oost J, Brouns SJJ, Severinov K. 2011. Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. Proc Natl Acad Sci U S A 108:10098–10103. https://doi.org/10.1073/pnas.1104144108.
18. Strotskaya A, Savitskaya E, Metlitskaya A, Morozova N, Datsenko KA, Semenova E, Severinov K. 2017. The action of *Escherichia coli* CRISPR-Cas system on lytic bacteriophages with different lifestyles and development strategies. Nucleic Acids Res 45:1946–1957. https://doi.org/10.1093/nar/gkx042.
19. Levy A, Goren MG, Yosef I, Auster O, Manor M, Amitai G, Edgar R, Qimron U, Sorek R. 2015. CRISPR adaptation biases explain preference for acquisition of foreign DNA. Nature 520:505–510. https://doi.org/10.1038/nature14302.
20. Ivančić-Bace I, Cass SD, Wearne SJ, Bolt EL. 2015. Different genome stability proteins underpin primed and naïve adaptation in *E. coli* CRISPR-Cas immunity. Nucleic Acids Res 43:10821–10830. https://doi.org/10.1093/nar/gkv1213.
21. Yosef I, Shitrit D, Goren MG, Burstein D, Pupko T, Qimron U. 2013. DNA motifs determining the efficiency of adaptation into the *Escherichia coli* CRISPR array. Proc Natl Acad Sci U S A 110:14396–14401. https://doi.org/10.1073/pnas.1300108110.
22. Musharova O, Klimuk E, Datsenko KA, Metlitskaya A, Logacheva M, Semenova E, Severinov K, Savitskaya E. 2017. Spacer-length DNA intermediates are associated with Cas1 in cells undergoing primed CRISPR adaptation. Nucleic Acids Res 45:3297–3307. https://doi.org/10.1093/nar/gkx097.
23. Semenova E, Savitskaya E, Musharova O, Strotskaya A, Vorontsova D, Datsenko KA, Logacheva MD, Severinov K. 2016. Highly efficient primed spacer acquisition from targets destroyed by the *Escherichia coli* type I-E CRISPR-Cas interfering complex. Proc Natl Acad Sci U S A 113:7626–7631. https://doi.org/10.1073/pnas.1602639113.
24. Krivoy A, Rutkauskas M, Kuznedelov K, Musharova O, Rouillon C, Severinov K, Seidel R. 2018. Primed CRISPR adaptation in *Escherichia coli* cells does not depend on conformational changes in the Cascade effector complex detected *in vitro*. Nucleic Acids Res 46:4087–4098. https://doi.org/10.1093/nar/gky219.
25. Li M, Wang R, Zhao D, Xiang H. 2014. Adaptation of the *Haloarcula hispanica* CRISPR-Cas system to a purified virus strictly requires a priming process. Nucleic Acids Res 42:2483–2492. https://doi.org/10.1093/nar/gkt1154.
26. Vorontsova D, Datsenko KA, Medvedeva S, Bondy-Denomy J, Savitskaya EE, Pougach K, Logacheva M, Wiedenheft B, Davidson AR, Severinov K, Semenova E. 2015. Foreign DNA acquisition by the I-F CRISPR-Cas system requires all components of the interference machinery. Nucleic Acids Res 43:10848–10860. https://doi.org/10.1093/nar/gkv1261.
27. Staals RHJ, Jackson SA, Biswas A, Brouns SJJ, Brown CM, Fineran PC. 2016. Interference-driven spacer acquisition is dominant over naive and primed adaptation in a native CRISPR-Cas system. Nat Commun 7:12853. https://doi.org/10.1038/ncomms12853.
28. Shmakov SA, Sitnik V, Makarova KS, Wolf YI, Severinov KV, Koonin EV. 2017. The CRISPR spacer space is dominated by sequences from species-specific mobilomes. mBio 8:e01397-17.
29. Shipman SL, Nivala J, Macklis JD, Church GM. 2016. Molecular recordings by directed CRISPR spacer acquisition. Science 353:aaf1175. https://doi.org/10.1126/science.aaf1175.
30. Shipman SL, Nivala J, Macklis JD, Church GM. 2017. CRISPR-Cas encoding of a digital movie into the genomes of a population of living bacteria. Nature 547:345–349. https://doi.org/10.1038/nature23017.
31. Rao C, Chin D, Ensminger AW. 2017. Priming in a permissive type I-C CRISPR-Cas system reveals distinct dynamics of spacer acquisition and loss. RNA 23:1525–1538. https://doi.org/10.1261/rna.062083.117.
32. Li M, Gong L, Zhao D, Zhou J, Xiang H. 2017. The spacer size of I-B CRISPR is modulated by the terminal sequence of the protospacer. Nucleic Acids Res 45:4642–4654. https://doi.org/10.1093/nar/gkx229.
33. Künne T, Kieper SN, Bannenberg JW, Vogel AIM, Miellet WR, Klein M, Depken M, Suarez-Diez M, Brouns SJJ. 2016. Cas3-derived target DNA degradation fragments fuel primed CRISPR adaptation. Mol Cell 63:852–864. https://doi.org/10.1016/j.molcel.2016.07.011.
34. Wang J, Li J, Zhao H, Sheng G, Wang M, Yin M, Wang Y. 2015. Structural and mechanistic basis of PAM-dependent spacer acquisition in CRISPR-Cas systems. Cell 163:840–853. https://doi.org/10.1016/j.cell.2015.10.008.
35. Nuñez JK, Lee ASY, Engelman A, Doudna JA. 2015. Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity. Nature 519:193–198. https://doi.org/10.1038/nature14237.
36. Richter C, Gristwood T, Clulow JS, Fineran PC. 2012. *In vivo* protein interactions and complex formation in the *Pectobacterium atrosepticum* subtype I-F CRISPR/Cas system. PLoS One 7:e49549. https://doi.org/10.1371/journal.pone.0049549.
37. Shmakov S, Savitskaya E, Semenova E, Logacheva MD, Datsenko KA, Severinov K. 2014. Pervasive generation of oppositely oriented spacers during CRISPR adaptation. Nucleic Acids Res 42:5907–5916. https://doi.org/10.1093/nar/gku226.
38. Morgan M, Anders S, Lawrence M, Aboyoun P, Pagès H, Gentleman R. 2009. ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. Bioinformatics 25:2607–2608. https://doi.org/10.1093/bioinformatics/btp450.
39. Pages H, Gentleman R, Aboyoun P, et al. 2008. Biostrings: string objects representing biological sequences, and matching algorithms. R Package version 2:2008. https://bioconductor.org.
40. Wilkinson L. 2011. ggplot2: elegant graphics for data analysis by WICKHAM, H. Biometrics 67:678–679. https://doi.org/10.1111/j.1541-0420.2011.01616.x.

CONCLUSIONS AND FUTURE PERSPECTIVES

CONCLUSIONS AND FUTURE PERSPECTIVES

In this work, we analyzed natural diversity of CRISPR spacers in environmental prokaryotic communities and in publicly available sequenced genomes. The comparison of obtained environmental spacer sets with each other and with spacers from databases as well as with sequences of viruses allowed us to reach several conclusions:

- Exploration of natural CRISPR spacer diversity — the CRISPRome — greatly surpasses the diversity from genomes of cultivated strains and proves to be a valid approach for studying virus-host interactions.
- Several contemporary E. coli CRISPR arrays remain unchanged over 40 thousand years, consistent with the inactivity of the adaptation module of type I-E CRISPR-Cas systems in this organism.
- Thermus, Sulfolobus and Flavobacteria communities adapt to local viruses, with different CRISPR-Cas systems targeting different viruses.
- Flavobacterial and Sulfolobales, but not Thermus, spacer sets display a biogeographical pattern.
- Sulfolobus viruses SPV1 and SPV2 carry mini-CRISPR arrays with 1-2 spacers against each other. Due to high abundance, spacers from mini-arrays are major contributors to the total population immunity. CRISPR spacer targeting promotes genome microevolution of viral genomes, whereas avoidance of self-targeting by mini-CRISPR arrays likely promotes virus speciation.
- Similar to the CRISPR-mediated interplay between SPV1 and SPV2, several mobile genetic elements integrated in the genomes of thaumarchaea include long CRISPR arrays with spacers against other thaumarchaeal mobile elements.

Similarities and differences between the studied systems will be discussed below.

1. Natural CRISPR spacer diversity greatly surpasses the diversity from fully sequenced genomes

We used metagenome sequencing to assess the natural diversity of CRISPR spacers, the CRISPRome, in diverse uncultivated prokaryotic communities including sterile mammoth intestine (Chapter I), fish pathogen community from surface snow in Antarctic (Chapter II), or Sulfolobales population from a thermal field in Beppu, Japan (Chapter IV). Due to constant encounter between viruses and cells, CRISPR loci are among the fastest evolving regions in

microbial genomes. Thus, environmental populations of bacteria and archaea, where each species encounters multiple mobile genetic elements, is expected to encompass considerable spacer diversity. Indeed, the diversity of spacers in the CRISPRome collections from each sampling site greatly exceeded the diversity of spacers in cultivated, fully sequenced strains isolated from different geographical locations. The amount of obtained data allowed us to use CRISPRome for identification of PAM sequences and novel variants of CRISPR repeats (Chapters II, III, IV), to detect integrated elements in the host genomes (Chapter IV), reconstruct contigs of new viruses (Chapter IV) and detect A/T (G/C) biases in nucleotide composition of spacer sequences (Annex).

The diversity of CRISPR spacers in a CRISPR array is known to increase towards the leader-proximal end of the array, where newly acquired spacers are located. Theoretical modelling of host populations cocultivated with several viruses predicts that only the newest 5 spacers grant immunity to contemporary, circulating viruses, whereas the rest of spacers are “outdated”, as viruses matching old spacers have either disappeared from the culture or have escaped CRISPR targeting by mutations in the protospacer regions. Concurrently, spacers at the leader distal end of the array have to be removed to minimize the potential burden associated with replication of constantly increasing CRISPR arrays. Thus, to maintain the immune function, CRISPR arrays should be constantly renewed with addition of new spacers and purging of old spacers. Another model demonstrated that viruses targeted by multiple spacers present in multiple strains are less likely to evade the CRISPR immunity, which can explain why CRISPR spacer diversity is preserved in the population for a long period of time.

Spacer diversity in the community can be represented as (i) a collection of CRISPR arrays; (ii) a set of alleles, combining several CRISPR arrays; or (iii) a set of spacers. To analyze the evolution of spacer diversity in the studied populations, we attempted reconstruction of spacer arrays, which provide a temporal dimension to the analysis. Instead of linear CRISPR arrays, the reconstructions resulted in complex assemblies, best represent in the framework of networks connected through spacers from ancestrally shared CRISPR arrays. The recent acquisition of spacers by distinct individuals can be seen in the networks as branching towards the leader end. At the same time, independent deletions of old spacers occurred at the leader distal end of arrays (Chapter III, Supplementary Figures 2 and 6; Chapter IV, Supplementary Figure 7). Thus, the network representation revealed both facets of CRISPR arrays, the active turnover of terminal spacers (acquisition and deletion) as well as stable spacer diversity in Thermus and Sulfolobus populations. More careful analysis of the network structure could be applied to identify spacers

under selection pressure, possible determinants of CRISPR array recombination or requirements for deletion of old spacers.

2. Several contemporary E. coli CRISPR arrays of I-E type remain unchanged over 40 thousand years

A long-term dynamics of E. coli I-E CRISPR spacers was studied by comparing spacer diversity in contemporary E. coli isolates with spacers amplified from mammoth intestinal content (Chapter I). A final set of 1883 unique spacer sequences from the mammoth intersected with the contemporary E. coli spacer set, which at the time consisted of 1599 unique sequences. This comparison revealed 425 common spacers. Moreover, fragments of contemporary CRISPR arrays were found in the mammoth sequencing data as pairs and triplets of neighboring spacers, allowing reconstruction of long CRISPR arrays from the paleo-samples. The lack of spacer turnover and stability of the spacer content was found for the 425 spacers shared between ancient and present-day CRISPR arrays. Accordingly, these spacers could be reconstructed in the form of linear CRISPR arrays, rather than networks described above. The majority of ancient spacers, however, was not found in the database of contemporary E. coli CRISPR spacers. Both limited diversity of E. coli strains in the CRISPR database and extinction of mammoth-associated E. coli strains could explain this result. Additional experiments, such as analysis of natural CRISPR spacer diversity associated with various animals (e.g., elephants) could shed light on the long-term CRISPR dynamics in E. coli. Our results suggest that the adaptation module of type I-E CRISPR-Cas system in E. coli has been inactive for at least 40 thousand years. The preservation of the inactive, but potentially dangerous immune system in E. coli genome suggests that it plays an alternative role(s) in the cell, such as response to stress induced by DNA damage (165).

3. Biogeographical patterns in the CRISPRome data

Flavobacterium and Sulfolobus CRISPRome data (Chapters II and IV) displayed biogeographical pattern, with spacer sets from geographically proximal sampling sites being more similar to each other compared to those from more remote locations. Spacer sets from three Antarctic sites differed significantly from each other, with only a very minor portion of spacers being common to all three sites. The larger amount of common spacers between Druzhnaja and Progress stations is consistent with their geographical proximity (Chapter II, Figure 5B). A similar overlap was observed for Sulfolobales spacer sets from Beppu thermal field and spacers from Japanese Sulfolobales isolates (Chapter IV, Figure 1A), which indicates that Beppu Sulfolobales population and Japanese isolates were infected with similar viruses. These results

are in line with the previous observations made using the comparison of CRISPR arrays from completely sequenced genomes of microbes isolated from geographically remote locations (138, 146, 178).

Surprisingly, however, in contrast to Flavobacterium and Sulfolobales natural communities, Thermus CRISPRome from enrichment cultures established from samples collected from sites separated by thousands of kilometers (Italy, Chile and Russia), showed no dependence on geographical distance (Chapter III, Figure 2B). At present, we are unable to explain this observation. It is possible that overnight cultivation conditions selected limited number of similar Thermus strains in different samples. Careful control of ecological parameters of habitat at the collection sites and extension of analysis presented here to other Thermus communities around the world may help to resolve this conundrum.

4. Thermus, Sulfolobus and Flavobacteria communities adapt to local viruses, with different CRISPR-Cas systems targeting different viruses

CRISPRome spacers of Thermus, Sulfolobus and Flavobacteria natural communities preferably target viruses, isolated/sequenced from the same source (Thermus: Chapter III, Table 2; Flavobacteria: Chapter II, Figure 5B; Sulfolobus: Chapter IV, Figure 1D). This result is consistent with local spacer targeting reported for many other environments and seems to be a general phenomenon (143, 148, 178, 179). Notably, SSV1 virus isolated in Beppu, Japan more than 30 years ago (180), is one of the most targeted viruses by present-day CRISPRome spacers from Japan (Chapter IV), emphasizing the longevity of the CRISPR-Cas "immunological memory". Despite isolation or virome sequencing of new viruses from the same source, the origin of the vast majority of spacers remains unclear. Given that number of spacers against the virus should negatively correlate with abundance of the virus in the population, the majority of spacers should target the "rare" viruses, i.e., minor components of the corresponding viromes. Indeed, CRISPR spacers derived from a metagenome of hypersaline environment mostly target low-abundance viruses in the virome (143). The most targeted virus in the Sulfolobales population from Beppu – SBRV1 – contributes less than 1% of all virome reads (Chapter IV). For the largest available dataset (Sulfolobales CRISPRome, 40705 unique spacers), we could reconstruct contigs of low-abundance virus genomes by tiling spacer sequences (Chapter IV, Supplementary Figure 3C). The reconstruction of the viral contigs from the CRISPRome data is conceptually similar to the reconstruction of plant virus genomes from small interfering RNA sequences (181). It is conceivable that complete viral genomes could be assembled using this approach, provided sufficient depth of CRISPRome sequencing and abundant CRISPR targeting.

Spacers associated with particular CRISPR-Cas systems specifically targeted distinct viruses (Chapters III and IV), which can be explained by a narrow host range or anti-CRISPR proteins, encoded by the virus. Moreover, different regions of the virus genomes are targeted with different frequencies. For example, almost all protospacers found in the genome of Thermus phage phiFa were located in the early genes. The infection of E. coli with phage T5 resulted in similar pattern of acquired spacers: all spacers were concentrated in the narrow genomic region of pre-early genes, which is injected in the cell before the rest of the genome (182). Thus, uneven distribution of spacers could reflect specific aspects of phage lifestyle. Another interesting example of uneven distribution of spacer hits along the genome is found in archaeal virus SBFV3 where the majority of protospacers are localized in the genomic termini, which are known to be the most variable in the SBFV3 genome. Spacer targeting of auxiliary genes located in the SBFV3 termini, including one anti-CRISPR protein coding gene, might increase the efficiency of antivirus response. Alternatively, in the case of filamentous viruses with linear genomes, such as SBFV3, either of the two termini is the first to penetrate into the cell interior and, thus, might be detected by the CRISPR-Cas system sooner than the central genomic region.

5. CRISPR-mediated interviral conflicts

Although the primary role of CRISPR-Cas systems is to defend bacteria and archaea against invading mobile genetic elements, the system has been hijacked by MGE on multiple independent occasions for various purposes. For instance, Vibrio phage ICP1 contains the complete CRISPR-Cas system to counter the PLE – a mobile genetic element, induced upon virus infection (183). Multiple CRISPR arrays were found in prophages of Clostridium difficile, with spacers matching sequences of other prophages (184). Tn7-like transposons encode minimalistic CRISPR-Cas systems with small CRISPR arrays. Spacers from transposon-encoded arrays match sequences of plasmids and phages, possibly facilitating the CRISPR-mediated transposition into the corresponding mobile elements and subsequent horizontal transmission between the hosts (185). Together these results support the "guns for hire" concept (186), whereby the CRISPR-Cas machinery of the host is adapted by mobile genetic elements for internal conflicts.

Analysis of mobile genetic elements integrated in the genomes of archaea from the phylum Thaumarchaeota has revealed several elements carrying long CRISPR arrays which, in some of the elements were associated with the cas genes (Chapter V). Interestingly, two orthologous elements found in the genomes of Ca. Nitrosocaldus isolates differed in the leader-proximal regions of the corresponding CRISPR arrays, suggesting active adaptation in the "mobile"

CRISPR arrays. Furthermore, we also obtained evidence for CRISPR-mediated conflicts between an integrative conjugative element and a provirus carried by soil thaumarchaea. Notably, the provirus was targeted by spacers from both the mobile element and a chromosomal CRISPR array. It is conceivable that such CRISPR-carrying mobile elements provide advantage to their host cells and the interaction between them might be considered a form of symbiosis.

A perhaps more unexpected was the discovery of mini-CRISPR arrays in two portoglioviruses, SPV1 and SPV2, with spacers targeting each other. In comparison to other examples described above, SPV1 and SPV2 implemented the most minimalistic solution – the mini-arrays includes only 1-2 spacers and with the leader sequence occupy only ~150 bp of intergenic space. Remarkably, SPV1 and SPV2 genomes are 92% identical to each other and yet instead of cooperating they appear to compete with each other. Indeed, virome sequencing suggests that SPV1 and SPV2 restrict each other through a distinct CRISPR-mediated superinfection exclusion mechanism. This strategy might ensure that the virus which is the first to infect the cell secures the resources for propagation and its components (e.g., structural proteins) are not highjacked by the superinfecting virus. The identical genomic position of one of the mini-arrays in SPV1 and SPV2 genomes implies that the two viruses have diverged after the acquisition of the first mini-array.

Another interesting consequence of CRISPR targeting between closely related viruses is that this process is likely to drive virus speciation. Indeed, changes in the SPV1 and SPV2 genomes are significantly correlated with the CRISPR targeting. It would be interesting to sequence other variants of SPV viruses detected in the virome and CRISPRome data to gain further insight into CRISPR-driven virus speciation. Notably, it has been recently suggested that CRISPR spacers acquired during inter-species mating of halophilic archaea also influence speciation (187).

Several additional mini-array candidates were found in the CRISPRome data, suggesting diverse population of SPVs present in different samples. Although most spacers from mini-arrays target SPV viruses, several spacers matched other mobile genetic elements, including unrelated viruses and cryptic integrated plasmids. This finding demonstrates that interactions mediated by mini-CRISPR arrays are not limited to inter-SPV conflicts. Unlike for spacers from long CRISPR arrays, for majority of spacers from the mini-arrays protospacers can be found, suggesting a fast spacer turnover in mini-arrays. Notably, SPV1 was shown to be a non-lytic virus, which stably propagates within the host cell without killing it or visibly affecting its growth under laboratory conditions (188). Thus, similar to thaumarchaeal CRISPR-carrying mobile elements, the symbiotic association between the SPVs and host cells might be beneficial to both parties.

Many aspects of the proposed CRISPR-mediated superinfection exclusion mechanism remain unclear and will require additional experiments. How did SPVs acquire mini-arrays? Mini-arrays could originate from a leader-repeat unit, which was acquired from the host through illegitimate recombination and subsequently expanded with new spacers by the host adaptation machinery. Based on different locations of mini-arrays and different leader sequences we can assume three independent events of mini-array acquisitions from two different host CRISPR arrays. Why do CRISPR arrays of SPVs remain miniature (1-2 spacers) instead of expanding and cataloging spacers like cellular CRISPR arrays do? It is possible that the length of SPV genome is limited by the volume of the icosahedral capsid, so that mini-arrays of SPVs cannot reach the size of the host CRISPR arrays. Moreover, the acquisition of new spacers in mini-arrays seems to be a rare event, as no spacers where added to SPVs during 20 days cultivation. How do SPVs evade CRISPR-Cas immunity of the host? According to the analysis of spacers from long CRISPR arrays at least some of the SPV1, SPV2 hosts have spacers matching SPV1 and SPV2 genomes. While an anti-CRISPR protein was reported for Sulfolobales viruses (154, 162), the proposed superinfection exclusion mechanism of SPV1 and SPV2 relies on non-inhibited CRISPR-Cas machinery of the host and could not be combined with interference-blocking anti-CRISPR proteins in SPVs. One possibility might be that SPVs block efficient expression and processing of the host-encoded CRISPR arrays, with mini-arrays being more efficiently produced. How do SPVs distinguish between self and non-self DNA during spacer acquisition? Viruses which acquired spacer from self DNA, without CRISPR-blocking mechanism will not be able to replicate and will be eliminated from the population. Answering these and other questions is a promising line of future research which should shed light on the molecular mechanisms of inter-viral conflicts and reveal additional facet of CRISPR-Cas systems.

ANNEX

ANNEX

1. Spacer diversity in Thermus and Sulfolobales genomes

Degenerate primer sequences allowed us to amplify spacers associated with wide range of CRISPR repeats. Some sequencing reads contain more than one spacer. In such cases, the original sequence of the CRISPR repeat is interspersed between the two spacers and can be analyzed. In both, Sulfolobus and Thermus datasets, new variants of CRISPR repeats, which were not present in any of the sequenced genomes, were found (Tables 1 and 2, respectively), suggesting large unexplored diversity of Thermus and Sulfolobus populations. Variations in CRISPR repeats mostly appeared outside 5' 8 nt tag and stem-forming regions, preserving predicted secondary structure.

Table 1. New variants of Thermus CRISPR repeats.

CRISPR repeat sequence	type	percentage from all CRISPR repeats of this type
GTTTCAAACCCTCATAGGTACGGTTCAAAG	I-A	84%
GTTGCACCGGCCCGAAAGGGCCGGGGAGGATTGAAAC	I-C	1%
GTTGCATCCAAGCTTCATGGCTTGGCTACGTTGCAGG	I-U	28%
GTCGCATCCAAGCTTCACAGCTTGGCTACGTTGCAGG	I-U	5%
GTTGCAAAAGTTGCTTCCCCGTCAGGGGATTGCGAC	III	34%
GTTGCAAAAGTGGCTTCCCCGTCAGGGGATTGCGAC	III	24%

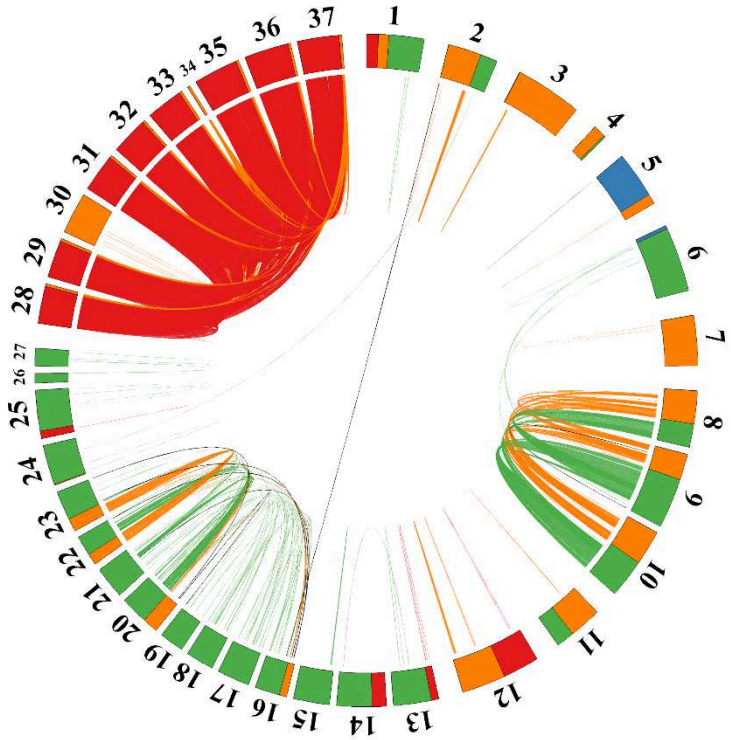
Table 2. New variants of Sulfolobales CRISPR repeats.

CRISPR repeat sequence	type	percentage from all CRISPR repeats of this type
GTTAATCTTCTATAGAATTGAAAG	A	7%
GTAAAAACATAAAGAAACTAAAAC	B	60%
GATTAATCCTAGAAGGAATTGAAAG	D	10%
GATGTATCCCAAAAGGAATTGAAAG	D	2%

Variations in CRISPR repeat sequences are shown with red color. Stem-forming nucleotides are shown with grey background. CRISPR repeat type, number of occurrences in HTS reads and frequency of new repeat sequences among CRISPR repeats of the same type are specified.

An alternative approach to access spacer diversity in natural prokaryotic community is analysis of spacers in fully sequenced isolates or amplification and sequencing of leader-proximal and leader-distant ends of array with specific primers (138, 146, 189). We analyzed the diversity of CRISPR spacers in sequenced isolates of Thermus and Sulfolobus (Chapter III, Figure 1 and

Figure 11, respectively). Both organisms, despite belonging to different domains of life, typically carry more than one CRISPR array in the genome. Sulfolobales CRISPR arrays are considerably longer than CRISPR arrays of Thermus (60 vs 14 spacers, respectively). The majority of Thermus spacers was strain-specific, with only 34 spacers (2.0%) being found in more than one genome. Even for very closely related T. thermophilus strains isolated in Japan (labeled as 22, 23, and 24 in Chapter III, Figure 1) only 6.7% (18 out of 269) spacers were shared. In contrast, for Sulfolobales CRISPR spacers a substantial fraction (26%) were shared between two or more strains of the same species, but only two spacers were common for different species or genera (Figure 10). S. acidocaldarius have the most conserved set of spacers: up to 98% of spacers are identical between two members of this species isolated from distant places (Japan, USA and Mexico). On the contrary, spacer sets of S. islandicus strains isolated from the same hot spring in Kamchatka intersected only by 34%. These patterns can be correlated with the richness of the corresponding mobilomes and overall genome conservation in S. acidocaldarius and S. islandicus. No viruses (except for one provirus) or plasmids have been described for S.



acidocaladarius, while a great diversity of viruses and plasmids are associated with S. islandicus (190).

Figure 11. A circular diagram of 9044 spacers from 37 fully sequenced genomes of Sulfolobales. Isolates used for analysis are numbered outside of spacer diagram. Spacers belonging to arrays of same CRISPR-Cas system

types are indicated by identical colors. Spacers that differ from each other by less than 2 nucleotides are connected by lines whose colors correspond to colors indicating CRISPR-Cas type. Black lines connect spacers shared by arrays of different types.

The difference between Thermus and Sulfolobales can be explained by a less sampled biogeography of Thermus isolates (Thermus biogeography: Chapter III, Figure 1; Sulfolobales biogeography: Chapter IV, Figure 1A), and faster spacer turnover in relatively short CRISPR arrays of Thermus.

2. Dinucleotide composition in Sulfolobales CRISPR spacers

It has been recently suggested that %GC content and oligonucleotide composition of spacers have a strong correlation with the composition of the source genome (191). However, the CRISPR interference mechanism implies a specific context of protospacer to distinguish between self and non-self DNA: 1) PAM sequence upstream of the protospacer is required for type I and type II interference; 2) a sequence similar to CRISPR repeat blocks DNA interference in type III systems. To investigate possible biases in specific sequences between spacers and source genomes connected with CRISPR interference mechanism, we analyzed dinucleotide compositions in S. islandicus, S. acidocaldarius, M. hakonensis and A. manzaensis genomes (Figure 12). Each selected genome has a dominant CRISPR repeat type - A, B, C, or D, correspondingly. The %GC content of selected genomes varies from 44% in Metallosphaera to 30% in Acidianus. Dinucleotide compositions of spacers correlate with the source genomes for all types of CRISPR repeats (see Figure 12). Significant differences, however, were found for frequencies of complementary dinucleotides in spacers for type A and type D repeats: GA, AG, and AA dinucleotides were underrepresented in comparison to their complementary sequences TC, CT, and TT. Underrepresented dinucleotides GA, AG, and AA constitute the end of A and D CRISPR repeat sequences “GAAAG”, which inhibit interference by type III complexes. Another explanation for the asymmetry in dinucleotide frequencies is purine over pyrimidine biases (G>C and A>T) found in coding vs. noncoding strands (192). During DNA interference by type III complexes, crRNA recognizes protospacer sequences in mRNA, so protective spacers must originate from the noncoding strand with C>G and T>A excesses.

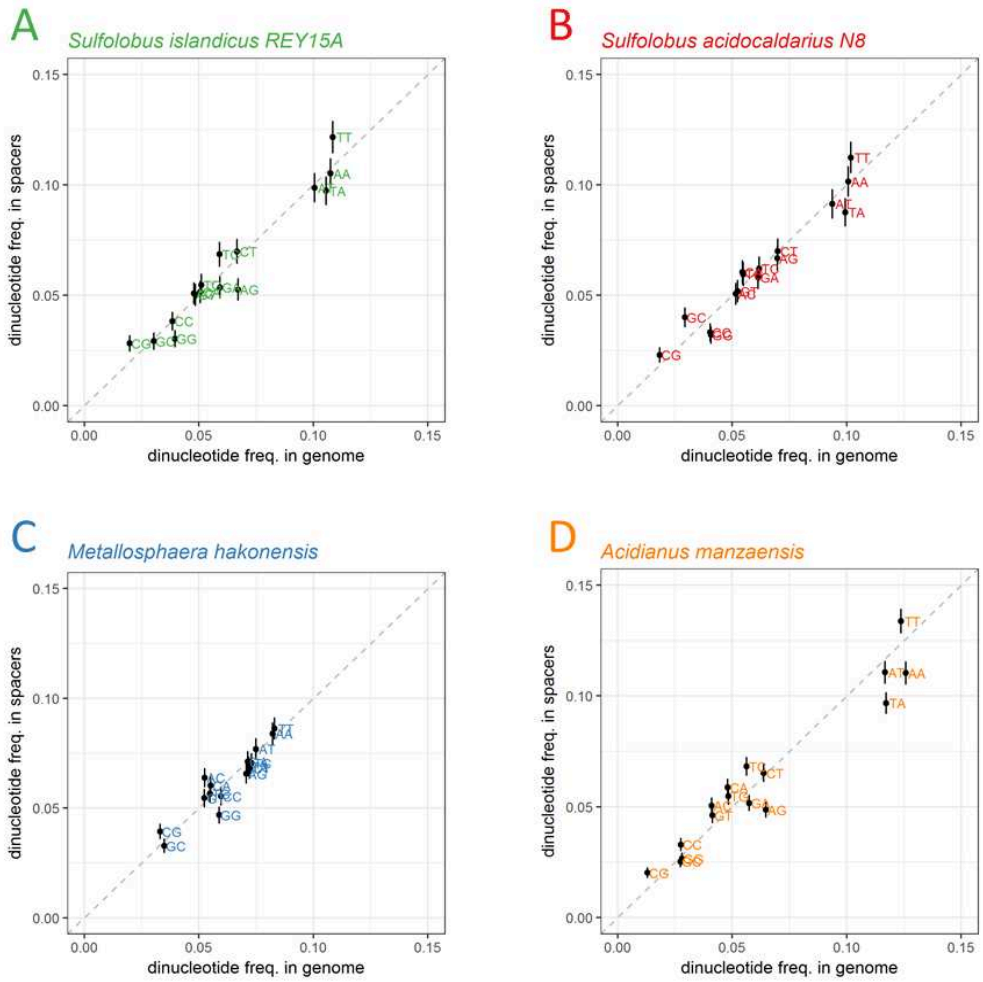


Figure 12. Dinucleotide frequencies in CRISPR spacers and corresponding host genomes. A comparison of dinucleotide frequencies for 4 selected genomes from Sulfolobales order and dinucleotide frequencies of CRISPR spacers from these genomes. Colors represent different types of CRISPR repeat (A, B, C, or D), the color labeling scheme is the same as in Figure 11. The diagonal is shown by dashed line. Error bars show confidence intervals for the proportion.

3. Short-term dynamics of CRISPR spacers: predation of Sulfolobales strains during cocultivation with viruses?

Sulfolobales and their viruses are known to coexist in enrichment cultures during 30-40 days-long cultivation. Presence of newly acquired spacers in CRISPR arrays of S. islandicus was observed ~10-30 days after infection (118). The temporal dynamics of spacer content was studied in two parallel series of enrichment cultures. The original environmental sample was the most diverse, whereas 10- and 20-days enrichment samples retained ~20% and ~15% of the diversity of the initial sample (Chapter IV, Supplementary Figure 2B). The substantial loss of spacers during the first 10 days of cultivation is likely to be caused by suboptimal cultivation conditions for some of the strains. Strains which survived after 10 days were considered as

cultivable. Thus, the difference in strain abundance between 10 days and 20 days can caused by competition between strains for limited resources, virus predation or some other factors.

We analyzed the dynamics of spacer groups, which are associated with different viruses present in the same culture (spacers that match virus genomes with >85% identity). Three scenarios were envisioned. (i) The significant increase in total abundance of spacers against a certain virus, which could be interpreted as an advantage of strains carrying protective spacers against this virus. (ii) A significant decrease in the total abundance of spacers against a certain virus, which would correspond to inefficient CRISPR-Cas protection and subsequent virus predation of the host. (iii) Finally, if no significant difference is found, the dynamics of strain abundance in the enrichment cultures is not connected with the CRISPR-Cas immunity, but depends on other factors. To estimate the significance of change in the total abundance for a group of spacers we randomly sampled 1 000 000 groups of spacers of the same size and calculated the distribution of log2ratio between 10 and 20 days samples (Figure 13).

The results for SPV1 and SPV2 viruses were biased by super-abundant spacers from mini-arrays (see Chapter IV); thus, we focused on three other viruses, targeted by the large number of spacers (Chapter IV, Figure 1D): SBRV1, SBV1, and SBFV3. The significant result was obtained for SBFV3 targeting spacers (Figure 13), which were decreased in abundance between 10 and 20 days as dramatically as 1% of the most decreased groups of spacers in sample J14. In sample J15, where SBFV3 was not present, SBFV3 targeting spacers showed the same behavior as random group of spacers. We concluded that this decrease is caused by SBFV3 predation of its host.

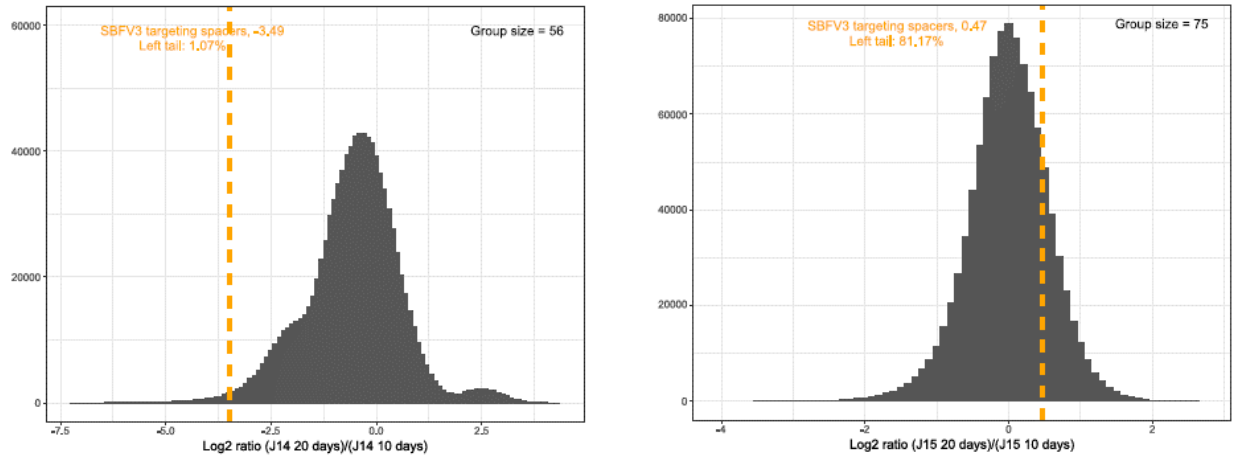


Figure 13. Log2ratio of spacer abundances between 10 and 20 days of cultivation. Left panel – J14 sample, right panel – J15 sample. Orange line shows log2ratio for spacers targeting SBFV3 virus.

4. PCR amplification of CRISPR spacers

A method for PCR amplification of CRISPR spacers with degenerate primers complementary to the CRISPR repeat sequence was optimized and validated using E. coli strain K12 as an experimental model (Chapter I, Figure 1B). The genome of E. coli strain K12 contains two CRISPR arrays with 12 and 6 spacers. The product of PCR amplification was sequenced and all 18 spacers were found in the sequencing data, validating the approach. However, the frequency of HTS reads varied for spacers from the same CRISPR array. Namely, the leader-proximal spacer of the first CRISPR array was ~10 times less abundant in HTS reads than the rest of the spacers of the same array. We observed a similar (~5 fold) difference between the most abundant and the least abundant spacer in the reconstructed CRISPR array of Sulfolbus (Chapter IV, Supplemntary Figure 8).

Abundance of spacer sequences in HTS reads can be described as a function of two parameters: (i) frequency of the host strain in population and (ii) PCR efficiency. As PCR efficiency depends on sequence of spacer and surrounding CRISPR repeats, the latter being almost constant for each spacer of the same CRISPR array, the abundance of spacers in enrichment cultures should change proportionally to the frequency of the host strain. This hypothesis is supported by the presence of groups of spacers with linearly correlated frequencies in two independent enrichment cultures (Chapter IV, Supplementary figure 8). We were unable to identify sequence features of spacers and surrounding repeats that determine the PCR efficiency. Presence of different motifs, G/C content, predicted secondary structure of ssDNA or mutations in CRISPR repeat sequences did not correlate with the abundance of spacers in HTS reads. More careful modelling of PCR amplification procedure might help to resolve this problem.

5. Cross-type CRISPR spacers

When spacer sequences associated with different CRISPR repeat types were compared to each other with 85% identity threshold, examples of spacers shared between two CRISPR repeats were found (see Figure 14), suggesting convergent and independent sampling of the same viral locus by different CRISPR-Cas systems. Approximately 5% of the analyzed Sulfolobales spacers associated with one type of CRISPR repeat match spacers from another CRISPR repeat. For the Thermus dataset, this value varied from 0% for the Etna spacer set to 6% for the Vesuvius spacers (see Figure 14A). The same phenomenon was observed for database spacers of Thermus (Chapter I, Figure 1) and S. islandicus strains isolated from Kamchatka (numbers 16-23 in Figure 1). For S. islandicus strains, 14 out of 552 (2.5%) spacers intersect between A and D

CRISPR repeats. Thus, given that all CRISPR repeat types are associated with different adaptation modules, intersecting spacers must have been independently acquired from the same locus of the same virus by two adaptation modules. Shared spacers are not equally distributed between CRISPR repeat types (Figure 14B), which can imply specificity of different CRISPR adaptation modules to different viruses. The specificity of different adaptation modules to different viruses can be a consequence of a narrow host range of a virus or virus-encoded anti-CRISPR proteins.

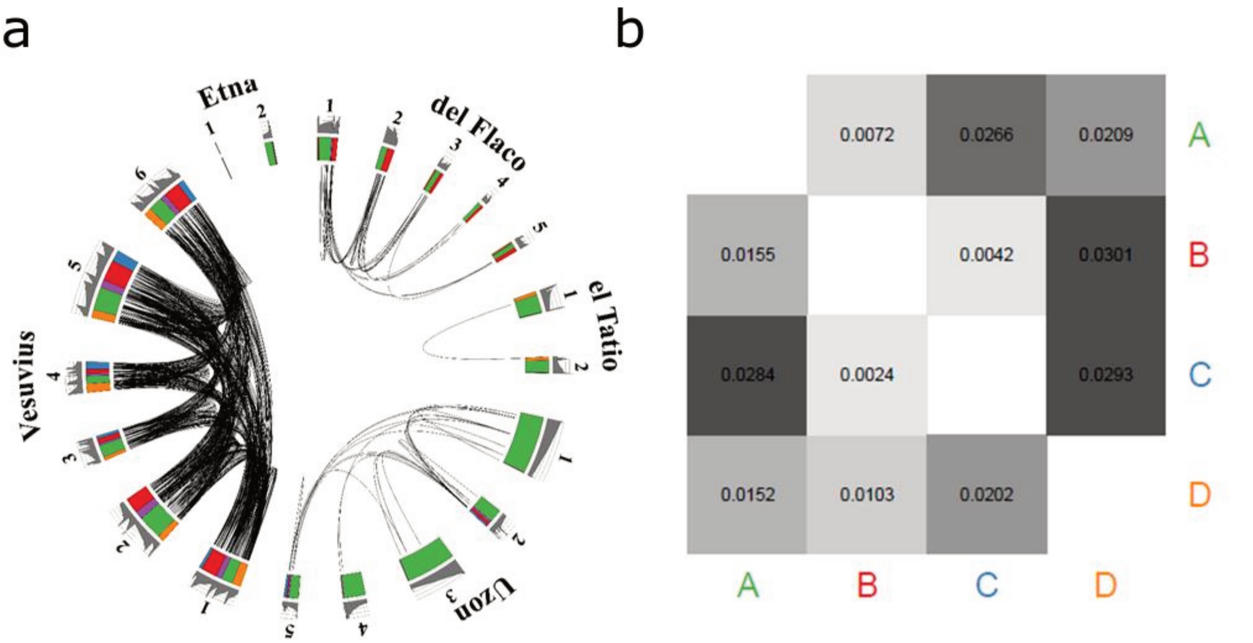


Figure 14. Intersection between spacer sets of different CRISPR repeats. A. For the Thermus dataset, spacers associated with different CRISPR repeat types that differ from each other by less than 2 nucleotides are connected by black lines**. B.** For Sulfolobales dataset, the fraction of spacers from the CRISPR repeat type in a row matching with CRISPR repeat type in the column is indicated.

REFERENCES

REFERENCES

1. Hegge JW, Swarts DC, van der Oost J. Prokaryotic Argonaute proteins: novel genome-editing tools? Nat Rev Microbiol. 2018;16(1):5-11.
2. Makarova KS, Wolf YI, Snir S, Koonin EV. Defense islands in bacterial and archaeal genomes and prediction of novel defense systems. J Bacteriol. 2011;193(21):6039-56.
3. Roberts RJ, Belfort M, Bestor T, Bhagwat AS, Bickle TA, Bitinaite J, et al. A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. Nucleic Acids Res. 2003;31(7):1805-12.
4. Tock MR, Dryden DT. The biology of restriction and anti-restriction. Curr Opin Microbiol. 2005;8(4):466-72.
5. Goldfarb T, Sberro H, Weinstock E, Cohen O, Doron S, Charpak-Amikam Y, et al. BREX is a novel phage resistance system widespread in microbial genomes. EMBO J. 2015;34(2):169-83.
6. Gordeeva J, Morozova N, Sierro N, Isaev A, Sinkunas T, Tsvetkova K, et al. BREX system of Escherichia coli distinguishes self from non-self by methylation of a specific DNA site. Nucleic Acids Res. 2019;47(1):253-65.
7. Swarts DC, Makarova K, Wang Y, Nakanishi K, Ketting RF, Koonin EV, et al. The evolutionary journey of Argonaute proteins. Nat Struct Mol Biol. 2014;21(9):743-53.
8. Azlan A, Dzaki N, Azzam G. Argonaute: The executor of small RNA function. J Genet Genomics. 2016;43(8):481-94.
9. Swarts DC, Szczepaniak M, Sheng G, Chandradoss SD, Zhu Y, Timmers EM, et al. Autonomous Generation and Loading of DNA Guides by Bacterial Argonaute. Mol Cell. 2017;65(6):985-98 e6.
10. Swarts DC, Koehorst JJ, Westra ER, Schaap PJ, van der Oost J. Effects of Argonaute on Gene Expression in Thermus thermophilus. PLoS One. 2015;10(4):e0124880.
11. Olovnikov I, Chan K, Sachidanandam R, Newman DK, Aravin AA. Bacterial argonaute samples the transcriptome to identify foreign DNA. Mol Cell. 2013;51(5):594-605.
12. Yuan YR, Pei Y, Ma JB, Kuryavyi V, Zhadina M, Meister G, et al. Crystal structure of A. aeolicus argonaute, a site-specific DNA-guided endoribonuclease, provides insights into RISC-mediated mRNA cleavage. Mol Cell. 2005;19(3):405-19.
13. Kaya E, Doxzen KW, Knoll KR, Wilson RC, Strutt SC, Kranzusch PJ, et al. A bacterial Argonaute with noncanonical guide RNA specificity. Proc Natl Acad Sci U S A. 2016;113(15):4057-62.

14. Makarova KS, Wolf YI, van der Oost J, Koonin EV. Prokaryotic homologs of Argonaute proteins are predicted to function as key components of a novel system of defense against mobile genetic elements. Biol Direct. 2009;4:29.
15. van Houte S, Buckling A, Westra ER. Evolutionary Ecology of Prokaryotic Immune Mechanisms. Microbiol Mol Biol Rev. 2016;80(3):745-63.
16. Parma DH, Snyder M, Sobolevski S, Nawroz M, Brody E, Gold L. The Rex system of bacteriophage lambda: tolerance and altruistic cell death. Genes Dev. 1992;6(3):497-510.
17. Chopin MC, Chopin A, Bidnenko E. Phage abortive infection in lactococci: variations on a theme. Curr Opin Microbiol. 2005;8(4):473-9.
18. Goeders N, Van Melderen L. Toxin-antitoxin systems as multilevel interaction systems. Toxins (Basel). 2014;6(1):304-24.
19. Brantl S. Bacterial type I toxin-antitoxin systems. RNA Biol. 2012;9(12):1488-90.
20. Smith JA, Magnuson RD. Modular organization of the Phd repressor/antitoxin protein. J Bacteriol. 2004;186(9):2692-8.
21. Blower TR, Pei XY, Short FL, Fineran PC, Humphreys DP, Luisi BF, et al. A processed noncoding RNA regulates an altruistic bacterial antiviral system. Nat Struct Mol Biol. 2011;18(2):185-90.
22. Tan Q, Awano N, Inouye M. YeeV is an Escherichia coli toxin that inhibits cell division by targeting the cytoskeleton proteins, FtsZ and MreB. Mol Microbiol. 2011;79(1):109-18.
23. Wang X, Lord DM, Cheng HY, Osbourne DO, Hong SH, Sanchez-Torres V, et al. A new type V toxin-antitoxin system where mRNA for toxin GhoT is cleaved by antitoxin GhoS. Nat Chem Biol. 2012;8(10):855-61.
24. Pecota DC, Wood TK. Exclusion of T4 phage by the hok/sok killer locus from plasmid R1. J Bacteriol. 1996;178(7):2044-50.
25. Koga M, Otsuka Y, Lemire S, Yonesaki T. Escherichia coli rnlA and rnlB compose a novel toxin-antitoxin system. Genetics. 2011;187(1):123-30.
26. Fineran PC, Blower TR, Foulds IJ, Humphreys DP, Lilley KS, Salmond GP. The phage abortive infection system, ToxIN, functions as a protein-RNA toxin-antitoxin pair. Proc Natl Acad Sci U S A. 2009;106(3):894-9.
27. Mojica FJ, Diez-Villasenor C, Garcia-Martinez J, Soria E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. J Mol Evol. 2005;60(2):174-82.
28. Makarova KS, Koonin EV. Annotation and Classification of CRISPR-Cas Systems. Methods Mol Biol. 2015;1311:47-75.

29. Koonin EV, Makarova KS, Zhang F. Diversity, classification and evolution of CRISPR-Cas systems. Curr Opin Microbiol. 2017;37:67-78.
30. Koonin EV, Makarova KS. Origins and evolution of CRISPR-Cas systems. Philos Trans R Soc Lond B Biol Sci. 2019;374(1772):20180087.
31. Kunin V, Sorek R, Hugenholtz P. Evolutionary conservation of sequence and secondary structures in CRISPR repeats. Genome Biol. 2007;8(4):R61.
32. Sefcikova J, Roth M, Yu G, Li H. Cas6 processes tight and relaxed repeat RNA via multiple mechanisms: A hypothesis. Bioessays. 2017;39(6).
33. Ivanova N, Daum C, Lang E, Abt B, Kopitz M, Saunders E, et al. Complete genome sequence of Haliangium ochraceum type strain (SMP-2). Stand Genomic Sci. 2010;2(1):96-106.
34. Godde JS, Bickerton A. The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes. J Mol Evol. 2006;62(6):718-29.
35. Peng X, Brugger K, Shen B, Chen L, She Q, Garrett RA. Genus-specific protein binding to the large clusters of DNA repeats (short regularly spaced repeats) present in Sulfolobus genomes. J Bacteriol. 2003;185(8):2410-7.
36. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, et al. CRISPR provides acquired resistance against viruses in prokaryotes. Science. 2007;315(5819):1709-12.
37. Yosef I, Goren MG, Qimron U. Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. Nucleic Acids Res. 2012;40(12):5569-76.
38. Alkhnbashi OS, Shah SA, Garrett RA, Saunders SJ, Costa F, Backofen R. Characterizing leader sequences of CRISPR loci. Bioinformatics. 2016;32(17):i576-i85.
39. Li H. Structural Principles of CRISPR RNA Processing. Structure. 2015;23(1):13-20.
40. Niewoehner O, Jinek M, Doudna JA. Evolution of CRISPR RNA recognition and processing by Cas6 endonucleases. Nucleic Acids Res. 2014;42(2):1341-53.
41. Deltcheva E, Chylinski K, Sharma CM, Gonzales K, Chao Y, Pirzada ZA, et al. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. Nature. 2011;471(7340):602-7.
42. Sternberg SH, Richter H, Charpentier E, Qimron U. Adaptation in CRISPR-Cas Systems. Mol Cell. 2016;61(6):797-808.
43. Tyson GW, Banfield JF. Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. Environ Microbiol. 2008;10(1):200-7.
44. Bernick DL, Cox CL, Dennis PP, Lowe TM. Comparative genomic and transcriptional analyses of CRISPR systems across the genus Pyrobaculum. Front Microbiol. 2012;3:251.

45. Martynov A, Severinov K, Ispolatov I. Optimal number of spacers in CRISPR arrays. PLoS Comput Biol. 2017;13(12):e1005891.
46. Weinberger AD, Sun CL, Plucinski MM, Denef VJ, Thomas BC, Horvath P, et al. Persisting viral sequences shape microbial CRISPR-based immunity. PLoS Comput Biol. 2012;8(4):e1002475.
47. Xiao Y, Ng S, Nam KH, Ke A. How type II CRISPR-Cas establish immunity through Cas1-Cas2-mediated spacer integration. Nature. 2017;550(7674):137-41.
48. Nunez JK, Lee AS, Engelman A, Doudna JA. Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity. Nature. 2015;519(7542):193-8.
49. Yoganand KN, Sivathanu R, Nimkar S, Anand B. Asymmetric positioning of Cas1-2 complex and Integration Host Factor induced DNA bending guide the unidirectional homing of protospacer in CRISPR-Cas type I-E system. Nucleic Acids Res. 2017;45(1):367-81.
50. Nivala J, Shipman SL, Church GM. Spontaneous CRISPR loci generation in vivo by non-canonical spacer integration. Nat Microbiol. 2018;3(3):310-8.
51. Nunez JK, Bai L, Harrington LB, Hinder TL, Doudna JA. CRISPR Immunological Memory Requires a Host Factor for Specificity. Mol Cell. 2016;62(6):824-33.
52. Westra ER, Pul U, Heidrich N, Jore MM, Lundgren M, Stratmann T, et al. H-NS-mediated repression of CRISPR-based immunity in Escherichia coli K12 can be relieved by the transcription activator LeuO. Mol Microbiol. 2010;77(6):1380-93.
53. Wang J, Li J, Zhao H, Sheng G, Wang M, Yin M, et al. Structural and Mechanistic Basis of PAM-Dependent Spacer Acquisition in CRISPR-Cas Systems. Cell. 2015;163(4):840-53.
54. Shah SA, Erdmann S, Mojica FJ, Garrett RA. Protospacer recognition motifs: mixed identities and functional diversity. RNA Biol. 2013;10(5):891-9.
55. Datsenko KA, Pougach K, Tikhonov A, Wanner BL, Severinov K, Semenova E. Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. Nat Commun. 2012;3:945.
56. Heler R, Samai P, Modell JW, Weiner C, Goldberg GW, Bikard D, et al. Cas9 specifies functional viral targets during CRISPR-Cas adaptation. Nature. 2015;519(7542):199-202.
57. Silas S, Mohr G, Sidote DJ, Markham LM, Sanchez-Amat A, Bhaya D, et al. Direct CRISPR spacer acquisition from RNA by a natural reverse transcriptase-Cas1 fusion protein. Science. 2016;351(6276):aad4234.
58. Levy A, Goren MG, Yosef I, Auster O, Manor M, Amitai G, et al. CRISPR adaptation biases explain preference for acquisition of foreign DNA. Nature. 2015;520(7548):505-10.
59. Savitskaya E, Semenova E, Dedkov V, Metlitskaya A, Severinov K. High-throughput analysis of type I-E CRISPR/Cas spacer acquisition in E. coli. RNA Biol. 2013;10(5):716-25.

60. Lemak S, Nocek B, Beloglazova N, Skarina T, Flick R, Brown G, et al. The CRISPR-associated Cas4 protein Pcal_0546 from Pyrobaculum calidifontis contains a [2Fe-2S] cluster: crystal structure and nuclease activity. Nucleic Acids Res. 2014;42(17):11144-55.
61. Hudaiberdiev S, Shmakov S, Wolf YI, Terns MP, Makarova KS, Koonin EV. Phylogenomics of Cas4 family nucleases. BMC Evol Biol. 2017;17(1):232.
62. Lee H, Zhou Y, Taylor DW, Sashital DG. Cas4-Dependent Prespacer Processing Ensures High-Fidelity Programming of CRISPR Arrays. Mol Cell. 2018;70(1):48-59 e5.
63. Kieper SN, Almendros C, Behler J, McKenzie RE, Nobrega FL, Haagsma AC, et al. Cas4 Facilitates PAM-Compatible Spacer Selection during CRISPR Adaptation. Cell Rep. 2018;22(13):3377-84.
64. Shiimori M, Garrett SC, Graveley BR, Terns MP. Cas4 Nucleases Define the PAM, Length, and Orientation of DNA Fragments Integrated at CRISPR Loci. Mol Cell. 2018;70(5):814-24 e6.
65. Krupovic M, Beguin P, Koonin EV. Casposons: mobile genetic elements that gave rise to the CRISPR-Cas adaptation machinery. Curr Opin Microbiol. 2017;38:36-43.
66. Koonin EV, Makarova KS. Mobile Genetic Elements and Evolution of CRISPR-Cas Systems: All the Way There and Back. Genome Biol Evol. 2017;9(10):2812-25.
67. Beguin P, Charpin N, Koonin EV, Forterre P, Krupovic M. Casposon integration shows strong target site preference and recapitulates protospacer integration by CRISPR-Cas systems. Nucleic Acids Res. 2016;44(21):10367-76.
68. Krupovic M, Shmakov S, Makarova KS, Forterre P, Koonin EV. Recent Mobility of Casposons, Self-Synthesizing Transposons at the Origin of the CRISPR-Cas Immunity. Genome Biol Evol. 2016;8(2):375-86.
69. Kwon AR, Kim JH, Park SJ, Lee KY, Min YH, Im H, et al. Structural and biochemical characterization of HP0315 from Helicobacter pylori as a VapD protein with an endoribonuclease activity. Nucleic Acids Res. 2012;40(9):4216-28.
70. Jackson RN, Golden SM, van Erp PB, Carter J, Westra ER, Brouns SJ, et al. Structural biology. Crystal structure of the CRISPR RNA-guided surveillance complex from Escherichia coli. Science. 2014;345(6203):1473-9.
71. Redding S, Sternberg SH, Marshall M, Gibb B, Bhat P, Guegler CK, et al. Surveillance and Processing of Foreign DNA by the Escherichia coli CRISPR-Cas System. Cell. 2015;163(4):854-65.
72. Zhao H, Sheng G, Wang J, Wang M, Bunkoczi G, Gong W, et al. Crystal structure of the RNA-guided immune surveillance Cascade complex in Escherichia coli. Nature. 2014;515(7525):147-50.

73. Hayes RP, Xiao Y, Ding F, van Erp PB, Rajashankar K, Bailey S, et al. Structural basis for promiscuous PAM recognition in type I-E Cascade from E. coli. Nature. 2016;530(7591):499-503.
74. van Erp PB, Jackson RN, Carter J, Golden SM, Bailey S, Wiedenheft B. Mechanism of CRISPR-RNA guided recognition of DNA targets in Escherichia coli. Nucleic Acids Res. 2015;43(17):8381-91.
75. Mulepati S, Heroux A, Bailey S. Structural biology. Crystal structure of a CRISPR RNA-guided surveillance complex bound to a ssDNA target. Science. 2014;345(6203):1479-84.
76. Blosser TR, Loeff L, Westra ER, Vlot M, Kunne T, Sobota M, et al. Two distinct DNA binding modes guide dual roles of a CRISPR-Cas protein complex. Mol Cell. 2015;58(1):60-70.
77. Xiao Y, Luo M, Hayes RP, Kim J, Ng S, Ding F, et al. Structure Basis for Directional R-loop Formation and Substrate Handover Mechanisms in Type I CRISPR-Cas System. Cell. 2017;170(1):48-60 e11.
78. Westra ER, van Erp PB, Kunne T, Wong SP, Staals RH, Seegers CL, et al. CRISPR immunity relies on the consecutive binding and degradation of negatively supercoiled invader DNA by Cascade and Cas3. Mol Cell. 2012;46(5):595-605.
79. Mulepati S, Bailey S. In vitro reconstitution of an Escherichia coli RNA-guided immune system reveals unidirectional, ATP-dependent degradation of DNA target. J Biol Chem. 2013;288(31):22184-92.
80. van Erp PBG, Patterson A, Kant R, Berry L, Golden SM, Forsman BL, et al. Conformational Dynamics of DNA Binding and Cas3 Recruitment by the CRISPR RNA-Guided Cascade Complex. ACS Chem Biol. 2018;13(2):481-90.
81. Kunne T, Kieper SN, Bannenberg JW, Vogel AI, Miellet WR, Klein M, et al. Cas3-Derived Target DNA Degradation Fragments Fuel Primed CRISPR Adaptation. Mol Cell. 2016;63(5):852-64.
82. Jinek M, Jiang F, Taylor DW, Sternberg SH, Kaya E, Ma E, et al. Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. Science. 2014;343(6176):1247997.
83. Sternberg SH, Redding S, Jinek M, Greene EC, Doudna JA. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. Nature. 2014;507(7490):62-7.
84. Jiang F, Taylor DW, Chen JS, Kornfeld JE, Zhou K, Thompson AJ, et al. Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. Science. 2016;351(6275):867-71.
85. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science. 2012;337(6096):816-21.

86. Gasiunas G, Barrangou R, Horvath P, Siksnys V. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. Proc Natl Acad Sci U S A. 2012;109(39):E2579-86.
87. Koonin EV, Makarova KS. Discovery of Oligonucleotide Signaling Mediated by CRISPR-Associated Polymerases Solves Two Puzzles but Leaves an Enigma. ACS Chem Biol. 2018;13(2):309-12.
88. Samai P, Pyenson N, Jiang W, Goldberg GW, Hatoum-Aslan A, Marraffini LA. Co-transcriptional DNA and RNA Cleavage during Type III CRISPR-Cas Immunity. Cell. 2015;161(5):1164-74.
89. Tamulaitis G, Venclovas C, Siksnys V. Type III CRISPR-Cas Immunity: Major Differences Brushed Aside. Trends Microbiol. 2017;25(1):49-61.
90. Kazlauskiene M, Tamulaitis G, Kostiuk G, Venclovas C, Siksnys V. Spatiotemporal Control of Type III-A CRISPR-Cas Immunity: Coupling DNA Degradation with the Target RNA Recognition. Mol Cell. 2016;62(2):295-306.
91. Tamulaitis G, Kazlauskiene M, Manakova E, Venclovas C, Nwokeoji AO, Dickman MJ, et al. Programmable RNA shredding by the type III-A CRISPR-Cas system of Streptococcus thermophilus. Mol Cell. 2014;56(4):506-17.
92. Estrella MA, Kuo FT, Bailey S. RNA-activated DNA cleavage by the Type III-B CRISPR-Cas effector complex. Genes Dev. 2016;30(4):460-70.
93. Kazlauskiene M, Kostiuk G, Venclovas C, Tamulaitis G, Siksnys V. A cyclic oligonucleotide signaling pathway in type III CRISPR-Cas systems. Science. 2017;357(6351):605-9.
94. Niewoehner O, Garcia-Doval C, Rostol JT, Berk C, Schwede F, Bigler L, et al. Type III CRISPR-Cas systems produce cyclic oligoadenylate second messengers. Nature. 2017;548(7669):543-8.
95. Makarova KS, Wolf YI, Alkhnbashi OS, Costa F, Shah SA, Saunders SJ, et al. An updated evolutionary classification of CRISPR-Cas systems. Nat Rev Microbiol. 2015;13(11):722-36.
96. Ozcan A, Pausch P, Linden A, Wulf A, Schuhle K, Heider J, et al. Type IV CRISPR RNA processing and effector complex formation in Aromatoleum aromaticum. Nat Microbiol. 2019;4(1):89-96.
97. Shmakov SA, Makarova KS, Wolf YI, Severinov KV, Koonin EV. Systematic prediction of genes functionally linked to CRISPR-Cas systems by gene neighborhood analysis. Proc Natl Acad Sci U S A. 2018;115(23):E5307-E16.

98. Zetsche B, Gootenberg JS, Abudayyeh OO, Slaymaker IM, Makarova KS, Essletzbichler P, et al. Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. Cell. 2015;163(3):759-71.
99. Shmakov S, Abudayyeh OO, Makarova KS, Wolf YI, Gootenberg JS, Semenova E, et al. Discovery and Functional Characterization of Diverse Class 2 CRISPR-Cas Systems. Mol Cell. 2015;60(3):385-97.
100. Swarts DC, van der Oost J, Jinek M. Structural Basis for Guide RNA Processing and Seed-Dependent DNA Targeting by CRISPR-Cas12a. Mol Cell. 2017;66(2):221-33 e4.
101. Fonfara I, Richter H, Bratovic M, Le Rhun A, Charpentier E. The CRISPR-associated DNA-cleaving enzyme Cpf1 also processes precursor CRISPR RNA. Nature. 2016;532(7600):517-21.
102. Yan WX, Hunnewell P, Alfonse LE, Carte JM, Keston-Smith E, Sothiselvam S, et al. Functionally diverse type V CRISPR-Cas systems. Science. 2019;363(6422):88-91.
103. Smargon AA, Cox DBT, Pyzocha NK, Zheng K, Slaymaker IM, Gootenberg JS, et al. Cas13b Is a Type VI-B CRISPR-Associated RNA-Guided RNase Differentially Regulated by Accessory Proteins Csx27 and Csx28. Mol Cell. 2017;65(4):618-30 e7.
104. Abudayyeh OO, Gootenberg JS, Konermann S, Joung J, Slaymaker IM, Cox DB, et al. C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. Science. 2016;353(6299):aaf5573.
105. Westra ER, van Houte S, Gandon S, Whitaker R. The ecology and evolution of microbial CRISPR-Cas adaptive immune systems. Philos Trans R Soc Lond B Biol Sci. 2019;374(1772):20190101.
106. Makarova KS, Wolf YI, Koonin EV. Comparative genomics of defense systems in archaea and bacteria. Nucleic Acids Res. 2013;41(8):4360-77.
107. Jorth P, Whiteley M. An evolutionary link between natural transformation and CRISPR adaptive immunity. MBio. 2012;3(5).
108. Touchon M, Charpentier S, Clermont O, Rocha EP, Denamur E, Branger C. CRISPR distribution within the Escherichia coli species is not suggestive of immunity-associated diversifying selection. J Bacteriol. 2011;193(10):2460-7.
109. Iranzo J, Lobkovsky AE, Wolf YI, Koonin EV. Evolutionary dynamics of the prokaryotic adaptive immunity system CRISPR-Cas in an explicit ecological context. J Bacteriol. 2013;195(17):3834-44.
110. Bernheim A, Bikard D, Touchon M, Rocha EPC. A matter of background: DNA repair pathways as a possible cause for the sparse distribution of CRISPR-Cas systems in bacteria. Philos Trans R Soc Lond B Biol Sci. 2019;374(1772):20180088.

111. Lillestol RK, Shah SA, Brugger K, Redder P, Phan H, Christiansen J, et al. CRISPR families of the crenarchaeal genus Sulfolobus: bidirectional transcription and dynamic properties. Mol Microbiol. 2009;72(1):259-72.
112. Gudbergsdottir S, Deng L, Chen Z, Jensen JV, Jensen LR, She Q, et al. Dynamic properties of the Sulfolobus CRISPR/Cas and CRISPR/Cmr systems when challenged with vector-borne viral and plasmid genes and protospacers. Mol Microbiol. 2011;79(1):35-49.
113. Lintner NG, Kerou M, Brumfield SK, Graham S, Liu H, Naismith JH, et al. Structural and functional characterization of an archaeal clustered regularly interspaced short palindromic repeat (CRISPR)-associated complex for antiviral defense (CASCADE). J Biol Chem. 2011;286(24):21643-56.
114. Rouillon C, Zhou M, Zhang J, Politis A, Beilsten-Edmands V, Cannone G, et al. Structure of the CRISPR interference complex CSM reveals key similarities with cascade. Mol Cell. 2013;52(1):124-34.
115. Zhang J, Rouillon C, Kerou M, Reeks J, Brugger K, Graham S, et al. Structure and mechanism of the CMR complex for CRISPR-mediated antiviral immunity. Mol Cell. 2012;45(3):303-13.
116. Sokolowski RD, Graham S, White MF. Cas6 specificity and CRISPR RNA loading in a complex CRISPR-Cas system. Nucleic Acids Res. 2014;42(10):6532-41.
117. Liu T, Liu Z, Ye Q, Pan S, Wang X, Li Y, et al. Coupling transcriptional activation of CRISPR-Cas system and DNA repair genes by Csa3a in Sulfolobus islandicus. Nucleic Acids Res. 2017;45(15):8978-92.
118. Erdmann S, Garrett RA. Selective and hyperactive uptake of foreign DNA by adaptive immune systems of an archaeon via two distinct mechanisms. Mol Microbiol. 2012;85(6):1044-56.
119. Rollie C, Graham S, Rouillon C, White MF. Prespacer processing and specific integration in a Type I-A CRISPR system. Nucleic Acids Res. 2018;46(3):1007-20.
120. Plagens A, Tripp V, Daume M, Sharma K, Klingl A, Hrle A, et al. In vitro assembly and activity of an archaeal CRISPR-Cas type I-A Cascade interference complex. Nucleic Acids Res. 2014;42(8):5125-38.
121. Manica A, Zebec Z, Teichmann D, Schleper C. In vivo activity of CRISPR-mediated virus defence in a hyperthermophilic archaeon. Mol Microbiol. 2011;80(2):481-91.
122. Mousaei M, Deng L, She Q, Garrett RA. Major and minor crRNA annealing sites facilitate low stringency DNA protospacer binding prior to Type I-A CRISPR-Cas interference in Sulfolobus. RNA Biol. 2016;13(11):1166-73.

123. Deng L, Garrett RA, Shah SA, Peng X, She Q. A novel interference mechanism by a type IIIB CRISPR-Cmr module in Sulfolobus. Mol Microbiol. 2013;87(5):1088-99.
124. Peng W, Feng M, Feng X, Liang YX, She Q. An archaeal CRISPR type III-B system exhibiting distinctive RNA targeting features and mediating dual RNA and DNA interference. Nucleic Acids Res. 2015;43(1):406-17.
125. Liu T, Li Y, Wang X, Ye Q, Li H, Liang Y, et al. Transcriptional regulator-mediated activation of adaptation genes triggers CRISPR de novo spacer acquisition. Nucleic Acids Res. 2015;43(2):1044-55.
126. He F, Vestergaard G, Peng W, She Q, Peng X. CRISPR-Cas type I-A Cascade complex couples viral infection surveillance to host transcriptional regulation in the dependence of Csa3b. Nucleic Acids Res. 2017;45(4):1902-13.
127. Quax TE, Voet M, Sismeiro O, Dillies MA, Jagla B, Coppee JY, et al. Massive activation of archaeal defense genes during viral infection. J Virol. 2013;87(15):8419-28.
128. Deng L, Kenchappa CS, Peng X, She Q, Garrett RA. Modulation of CRISPR locus transcription by the repeat-binding protein Cbp1 in Sulfolobus. Nucleic Acids Res. 2012;40(6):2470-80.
129. Groenen PM, Bunschoten AE, van Soolingen D, van Embden JD. Nature of DNA polymorphism in the direct repeat cluster of Mycobacterium tuberculosis; application for strain differentiation by a novel typing method. Mol Microbiol. 1993;10(5):1057-65.
130. Shariat N, Dudley EG. CRISPRs: molecular signatures used for pathogen subtyping. Appl Environ Microbiol. 2014;80(2):430-9.
131. Cowan LS, Diem L, Brake MC, Crawford JT. Transfer of a Mycobacterium tuberculosis genotyping method, Spoligotyping, from a reverse line-blot hybridization, membrane-based assay to the Luminex multianalyte profiling system. J Clin Microbiol. 2004;42(1):474-7.
132. Mokrousov I, Narvskaya O, Limeschenko E, Vyazovaya A. Efficient discrimination within a Corynebacterium diphtheriae epidemic clonal group by a novel macroarray-based method. J Clin Microbiol. 2005;43(4):1662-8.
133. Cui Y, Li Y, Gorge O, Platonov ME, Yan Y, Guo Z, et al. Insight into microevolution of Yersinia pestis by clustered regularly interspaced short palindromic repeats. PLoS One. 2008;3(7):e2652.
134. Fabre L, Zhang J, Guigon G, Le Hello S, Guibert V, Accou-Demartin M, et al. CRISPR typing and subtyping for improved laboratory surveillance of Salmonella infections. PLoS One. 2012;7(5):e36995.

135. Delannoy S, Beutin L, Burgos Y, Fach P. Specific detection of enteroaggregative hemorrhagic Escherichia coli O104:H4 strains by use of the CRISPR locus as a target for a diagnostic real-time PCR. J Clin Microbiol. 2012;50(11):3485-92.
136. Shariat N, Kirchner MK, Sandt CH, Trees E, Barrangou R, Dudley EG. Subtyping of Salmonella enterica serovar Newport outbreak isolates by CRISPR-MVLST and determination of the relationship between CRISPR-MVLST and PFGE results. J Clin Microbiol. 2013;51(7):2328-36.
137. Bochkareva OO, Dranenko NO, Ocheredko ES, Kanevsky GM, Lozinsky YN, Khalaycheva VA, et al. Genome rearrangements and phylogeny reconstruction in Yersinia pestis. PeerJ. 2018;6:e4545.
138. Held NL, Herrera A, Whitaker RJ. Reassortment of CRISPR repeat-spacer loci in Sulfolobus islandicus. Environ Microbiol. 2013;15(11):3065-76.
139. Skennerton CT, Imelfort M, Tyson GW. Crass: identification and reconstruction of CRISPR from unassembled metagenomic data. Nucleic Acids Res. 2013;41(10):e105.
140. Grissa I, Vergnaud G, Pourcel C. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res. 2007;35(Web Server issue):W52-7.
141. Edgar RC. PILER-CR: fast and accurate identification of CRISPR repeats. BMC Bioinformatics. 2007;8:18.
142. Moller AG, Liang C. MetaCRAST: reference-guided extraction of CRISPR spacers from unassembled metagenomes. PeerJ. 2017;5:e3788.
143. Emerson JB, Andrade K, Thomas BC, Norman A, Allen EE, Heidelberg KB, et al. Virus-host and CRISPR dynamics in Archaea-dominated hypersaline Lake Tyrrell, Victoria, Australia. Archaea. 2013;2013:370871.
144. Gudbergsdottir SR, Menzel P, Krogh A, Young M, Peng X. Novel viral genomes identified from six metagenomes reveal wide distribution of archaeal viruses and high viral diversity in terrestrial hot springs. Environ Microbiol. 2016;18(3):863-74.
145. Snyder JC, Bateson MM, Lavin M, Young MJ. Use of cellular CRISPR (clusters of regularly interspaced short palindromic repeats) spacer-based microarrays for detection of viruses in environmental samples. Appl Environ Microbiol. 2010;76(21):7251-8.
146. Held NL, Herrera A, Cadillo-Quiroz H, Whitaker RJ. CRISPR associated diversity within a population of Sulfolobus islandicus. PLoS One. 2010;5(9).
147. Robles-Sikisaka R, Naidu M, Ly M, Salzman J, Abeles SR, Boehm TK, et al. Conservation of streptococcal CRISPRs on human skin and saliva. BMC Microbiol. 2014;14:146.

148. Pride DT, Sun CL, Salzman J, Rao N, Loomer P, Armitage GC, et al. Analysis of streptococcal CRISPRs from human saliva reveals substantial sequence diversity within and between subjects over time. Genome Res. 2011;21(1):126-36.
149. Andersson AF, Banfield JF. Virus population dynamics and acquired virus resistance in natural microbial communities. Science. 2008;320(5879):1047-50.
150. Stern A, Mick E, Tirosh I, Sagy O, Sorek R. CRISPR targeting reveals a reservoir of common phages associated with the human gut microbiome. Genome Res. 2012;22(10):1985-94.
151. Ye Y, Zhang Q. Characterization of CRISPR RNA transcription by exploiting stranded metatranscriptomic data. RNA. 2016;22(7):945-56.
152. Borges AL, Davidson AR, Bondy-Denomy J. The Discovery, Mechanisms, and Evolutionary Impact of Anti-CRISPRs. Annu Rev Virol. 2017;4(1):37-59.
153. Pawluk A, Davidson AR, Maxwell KL. Anti-CRISPR: discovery, mechanism and function. Nat Rev Microbiol. 2018;16(1):12-7.
154. Guo T, Han W, She Q. Tolerance of Sulfolobus SMV1 virus to the immunity of I-A and III-B CRISPR-Cas systems in Sulfolobus islandicus. RNA Biol. 2018:1-8.
155. Hynes AP, Rousseau GM, Agudelo D, Goulet A, Amigues B, Loehr J, et al. Widespread anti-CRISPR proteins in virulent bacteriophages inhibit a range of Cas9 proteins. Nat Commun. 2018;9(1):2919.
156. Pawluk A, Staals RH, Taylor C, Watson BN, Saha S, Fineran PC, et al. Inactivation of CRISPR-Cas systems by anti-CRISPR proteins in diverse bacterial species. Nat Microbiol. 2016;1(8):16085.
157. Watters KE, Fellmann C, Bai HB, Ren SM, Doudna JA. Systematic discovery of natural CRISPR-Cas12a inhibitors. Science. 2018;362(6411):236-9.
158. Chowdhury S, Carter J, Rollins MF, Golden SM, Jackson RN, Hoffmann C, et al. Structure Reveals Mechanisms of Viral Suppressors that Intercept a CRISPR RNA-Guided Surveillance Complex. Cell. 2017;169(1):47-57 e11.
159. Bondy-Denomy J, Garcia B, Strum S, Du M, Rollins MF, Hidalgo-Reyes Y, et al. Multiple mechanisms for CRISPR-Cas inhibition by anti-CRISPR proteins. Nature. 2015;526(7571):136-9.
160. Yang H, Patel DJ. Inhibition Mechanism of an Anti-CRISPR Suppressor AcrIIA4 Targeting SpyCas9. Mol Cell. 2017;67(1):117-27 e5.
161. Harrington LB, Doxzen KW, Ma E, Liu JJ, Knott GJ, Edraki A, et al. A Broad-Spectrum Inhibitor of CRISPR-Cas9. Cell. 2017;170(6):1224-33 e15.

162. He F, Bhoobalan-Chitty Y, Van LB, Kjeldsen AL, Dedola M, Makarova KS, et al. Anti-CRISPR proteins encoded by archaeal lytic viruses inhibit subtype I-D immunity. Nat Microbiol. 2018;3(4):461-9.
163. Maxwell KL. The Anti-CRISPR Story: A Battle for Survival. Mol Cell. 2017;68(1):8-14.
164. Pougach K, Semenova E, Bogdanova E, Datsenko KA, Djordjevic M, Wanner BL, et al. Transcription, processing and function of CRISPR cassettes in Escherichia coli. Mol Microbiol. 2010;77(6):1367-79.
165. Bozic B, Repac J, Djordjevic M. Endogenous Gene Regulation as a Predicted Main Function of Type I-E CRISPR/Cas System in E. coli. Molecules. 2019;24(4).
166. Savitskaya E, Lopatina A, Medvedeva S, Kapustin M, Shmakov S, Tikhonov A, et al. Dynamics of Escherichia coli type I-E CRISPR spacers over 42 000 years. Mol Ecol. 2017;26(7):2019-26.
167. Stern A, Keren L, Wurtzel O, Amitai G, Sorek R. Self-targeting by CRISPR: gene regulation or autoimmunity? Trends Genet. 2010;26(8):335-40.
168. Babu M, Beloglazova N, Flick R, Graham C, Skarina T, Nocek B, et al. A dual function of the CRISPR-Cas system in bacterial antivirus immunity and DNA repair. Mol Microbiol. 2011;79(2):484-502.
169. Westra ER, Buckling A, Fineran PC. CRISPR-Cas systems: beyond adaptive immunity. Nat Rev Microbiol. 2014;12(5):317-26.
170. Thony-Meyer L, Kaiser D. devRS, an autoregulated and essential genetic locus for fruiting body development in Myxococcus xanthus. J Bacteriol. 1993;175(22):7450-62.
171. Boysen A, Ellehauge E, Julien B, Sogaard-Andersen L. The DevT protein stimulates synthesis of FruA, a signal transduction protein required for fruiting body morphogenesis in Myxococcus xanthus. J Bacteriol. 2002;184(6):1540-6.
172. Viswanathan P, Murphy K, Julien B, Garza AG, Kroos L. Regulation of dev, an operon that includes genes essential for Myxococcus xanthus development and CRISPR-associated genes and repeats. J Bacteriol. 2007;189(10):3738-50.
173. Zegans ME, Wagner JC, Cady KC, Murphy DM, Hammond JH, O'Toole GA. Interaction between bacteriophage DMS3 and host CRISPR region inhibits group behaviors of Pseudomonas aeruginosa. J Bacteriol. 2009;191(1):210-9.
174. Cady KC, O'Toole GA. Non-identity-mediated CRISPR-bacteriophage interaction mediated via the Csy and Cas3 proteins. J Bacteriol. 2011;193(14):3433-45.
175. Sampson TR, Saroj SD, Llewellyn AC, Tzeng YL, Weiss DS. A CRISPR/Cas system mediates bacterial innate immune evasion and virulence. Nature. 2013;497(7448):254-7.

176. Louwen R, Horst-Kreft D, de Boer AG, van der Graaf L, de Knegt G, Hamersma M, et al. A novel link between Campylobacter jejuni bacteriophage defence, virulence and Guillain-Barre syndrome. Eur J Clin Microbiol Infect Dis. 2013;32(2):207-26.
177. Gunderson FF, Cianciotto NP. The CRISPR-associated gene cas2 of Legionella pneumophila is required for intracellular infection of amoebae. MBio. 2013;4(2):e00074-13.
178. Held NL, Whitaker RJ. Viral biogeography revealed by signatures in Sulfolobus islandicus genomes. Environ Microbiol. 2009;11(2):457-66.
179. Sorokin VA, Gelfand MS, Artamonova, II. Evolutionary dynamics of clustered irregularly interspaced short palindromic repeat systems in the ocean metagenome. Appl Environ Microbiol. 2010;76(7):2136-44.
180. Martin A, Yeats S, Janekovic D, Reiter WD, Aicher W, Zillig W. SAV 1, a temperate u.v.-inducible DNA virus-like particle from the archaebacterium Sulfolobus acidocaldarius isolate B12. EMBO J. 1984;3(9):2165-8.
181. Pooggin MM. Small RNA-Omics for Plant Virus Identification, Virome Reconstruction, and Antiviral Defense Characterization. Front Microbiol. 2018;9:2779.
182. Strotskaya A, Savitskaya E, Metlitskaya A, Morozova N, Datsenko KA, Semenova E, et al. The action of Escherichia coli CRISPR-Cas system on lytic bacteriophages with different lifestyles and development strategies. Nucleic Acids Res. 2017;45(4):1946-57.
183. Seed KD, Lazinski DW, Calderwood SB, Camilli A. A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. Nature. 2013;494(7438):489-91.
184. Hargreaves KR, Flores CO, Lawley TD, Clokie MR. Abundant and diverse clustered regularly interspaced short palindromic repeat spacers in Clostridium difficile strains and prophages target multiple phage types within this pathogen. MBio. 2014;5(5):e01045-13.
185. Peters JE, Makarova KS, Shmakov S, Koonin EV. Recruitment of CRISPR-Cas systems by Tn7-like transposons. Proc Natl Acad Sci U S A. 2017;114(35):E7358-E66.
186. Koonin EV KM. A Movable Defense. Scientist. 2015(29(1)):46-53.
187. Turgeman-Grott I, Joseph S, Marton S, Eizenshtein K, Naor A, Soucy SM, et al. Pervasive acquisition of CRISPR memory driven by inter-species mating of archaea can limit gene transfer and influence speciation. Nat Microbiol. 2019;4(1):177-86.
188. Liu Y, Brandt D, Ishino S, Ishino Y, Koonin EV, Kalinowski J, et al. New archaeal viruses discovered by metagenomic analysis of viral communities in enrichment cultures. Environ Microbiol. 2018.
189. Diez-Villasenor C, Almendros C, Garcia-Martinez J, Mojica FJ. Diversity of CRISPR loci in Escherichia coli. Microbiology. 2010;156(5):1351-61.

190. Prangishvili D, Bamford DH, Forterre P, Iranzo J, Koonin EV, Krupovic M. The enigmatic archaeal virosphere. Nat Rev Microbiol. 2017;15(12):724-39.
191. Shmakov SA, Sitnik V, Makarova KS, Wolf YI, Severinov KV, Koonin EV. The CRISPR Spacer Space Is Dominated by Sequences from Species-Specific Mobilomes. MBio. 2017;8(5).
192. Lao PJ, Forsdyke DR. Thermophilic bacteria strictly obey Szybalski's transcription direction rule and politely purine-load RNAs with both adenine and guanine. Genome Res. 2000;10(2):228-36.

ACKNOWLEDGEMENTS

I would like to thank all the jury members Prof. Guennadi Sezonov, Prof. Mikhail Gelfand, Prof. Dmitri Pervouchine, Dr. Tamara Basta-Le Berre, Dr. David Bikard and Prof. Olga Soutourina for their time and consideration. I also want to acknowledge all the members of my Individual Thesis Committee in Russia and France for fruitful discussions during annual reviews.

I am thankful to my supervisors Prof. Konstantin Severinov and Dr. Mart Krupovic for sharing their experience and passion for science, thoughtful guidance, motivation and support, and for the enormous help with research projects, manuscripts and thesis.

I would like to acknowledge my colleagues and coauthors Dr. Anna Lopatina, Dr. Ekaterina Savitskaya, Dr. Ying Liu and Dr. Daria Artamonova for performing all the experiments and providing data for my bioinformatic work.

I wish to thank Prof. Patrick Forterre for providing the opportunity to work in his laboratory and Ana Cova Rodrigues for the help with organization of my joint thesis.

Special thanks to all the members of Moscow and Paris labs, who supported me on this long journey, without you it would not have been such a pleasant time.

RÉSUMÉ

Introduction

Le système CRISPR-Cas est un système immunitaire procaryote de type interférence ARN dirigé contre des éléments génétiques mobiles, tels que les virus et les plasmides[1]. Le système consiste en un ou plusieurs loci CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats ; courtes répétitions palindromiques groupées et régulièrement espacées) associés à des protéines Cas (CRISPR-associated proteins) dont ils sont séparés par une séquence dite leader. Toutes les protéines Cas peuvent être fonctionnellement attribuées à des modules d'adaptation, d'expression et d'interférence[2]. Les protéines Cas du module d'adaptation incorporent des fragments de l'ADN viral dans le locus CRISPR en tant que spacers (ou espaceurs) entre les répétitions. La transcription et la maturation du locus CRISPR donnent lieu à la production d'ARN de protection, l'ARN CRISPR (crRNA). Les protéines du module d'interférence, dirigées par les crRNA, reconnaissent et clivent des régions apparentées dans l'ADN ou l'ARN d'un élément génétique mobile. Sur la base de la composition des modules d'interférence et d'adaptation, les systèmes CRISPR-Cas sont classés en 2 classes, 6 types et environ 30 sous-types[3]. Les systèmes CRISPR-Cas sont présents dans 90% des archées, mais seulement dans 50% des bactéries[4]. Les organismes thermophiles sont particulièrement enrichis en systèmes CRISPR-Cas (et autres systèmes de défense) par rapport aux procaryotes mésophiles et psychrophiles[5]. Selon les simulations théoriques, les hôtes possédant un système CRISPR sont plus avantagés dans des conditions de faible diversité virale, comme c'est le cas dans les environnements géothermiques chauds par exemple[6].

L'analyse des spacers CRISPR est une précieuse source d'informations sur les interactions virus-hôte, puisqu'ils correspondent à de courts fragments d'ADN de virus précédemment rencontrés et « enregistrés » dans les loci CRISPR. De plus, les cellules portant des spacers protecteurs devraient acquérir un avantage et devenir plus nombreuses. Une telle analyse peut être particulièrement enrichissante lorsqu'elle est appliquée à des données métagénomiques. Outre l'extraction à partir de données métagénomiques ou de loci CRISPR[7,8], les spacers CRISPR peuvent être directement amplifiés et analysés à partir d'isolats bactériens individuels ou de communautés entières[9-11]. Ainsi, la comparaison des loci CRISPR de populations isolées de la même espèce a par exemple révélé une grande diversité de séquences spacers, bien supérieure à celle observée dans la séquence leader du locus CRISPR[8,12-14]. L'analyse de l'évolution du contenu en séquences spacer a également fourni des exemples d'acquisition de nouveaux spacers, de suppression d'anciens, et de recombinaison de loci CRISPR entre différentes souches[10,15-17]. Les spacers CRISPR peuvent aussi être utilisés pour identifier les

séquences virales dans des métagénomes et détecter les modifications dans les populations virales[7,18,19]. Des exemples de spacers ciblant de préférence des phages locaux du même site d'échantillonnage ont été rapportés[10,16,20,21]. La présence de multiples spacers contre un même génome viral dans les souches hôtes rend plus difficile la parade du virus par acquisition de mutations dans un des sites concernés, favorisant une grande diversité des spacers sur des échelles de temps plus longues[22].

Buts de la recherche

En utilisant une amplification PCR avec des amorces complémentaires des répétitions CRISPR suivie d'un séquençage de nouvelle génération (NGS), la diversité des spacers CRISPR dans différentes populations naturelles de procaryotes (le CRISPRome) a été analysée :

● Les spacers CRISPR du système I-E d'E. coli provenant de l'intestin d'un mammouth (chapitre I).

● Les spacers CRISPR du système II-C de Flavobacterium provenant de neige de surface autour de trois stations en Antarctique (chapitre II).

● Les spacers CRISPR des systèmes I-A, I-B, I-C, I-E, I-U et III-A/B de Thermus provenant de cinq sources thermales géographiquement distantes (Chapitre III).

● Les spacers CRISPR de Sulfolobus provenant des sources chaudes de Beppu au Japon (chapitre IV).

Les résultats de l'analyse du CRISPRome permettent de répondre à plusieurs questions :

● Dans quelle mesure la diversité des spacers CRISPR est-elle représentée dans les bases de données actuelles ? (Chapitres I, II, III, IV, V)

● Quelle est la dynamique à court et à long terme de la diversité des séquences spacers ? (Chapitres I, IV)

● Les populations procaryotes géographiquement proches/lointaines ont-elles une diversité de spacers similaires/différentes ? (Chapitres II, III, IV)

● Les populations procaryotes ont-elles une immunité CRISPR contre les virus locaux ? (Chapitres II, III, IV, V)

Résultats

La comparaison des spacers environnementaux les uns avec les autres et avec des spacers de bases de données ainsi que des séquences de virus nous a permis de tirer plusieurs conclusions :

• L’amplification par PCR des spacers, suivie du séquençage NGS, nous a permis d’obtenir une diversité de spacers issus de communautés procaryotes non-cultivées, provenant de l’intestin stérile d’un mammouth (Chapitre I), d’un pathogène de poissons provenant de neige de surface en Antarctique (Chapitre II), et de Sulfolobales des sources chaudes de Beppu au Japon (Chapitre IV). La diversité naturelle des spacers CRISPR (le CRISPRome) dépasse de beaucoup la diversité des génomes des souches cultivées, et son exploration s’avère être une approche valable pour l’étude des interactions virus-hôte. Par exemple, l’alignement des séquences spacer contre les chromosomes de l'hôte s'est révélée une approche efficace pour identifier les éléments génétiques mobiles intégrés (chapitre IV). Le jeu de données CRISPRome de Sulfolobus a été utilisé pour assembler plusieurs nouveaux contigs viraux en combinant les séquences spacer se chevauchant.

• Une dynamique à long terme des spacers CRISPR I-E de E. coli a été étudiée en comparant la diversité des spacers dans les génomes publiés d'E. coli avec des spacers amplifiés à partir du contenu intestinal de mammouth. Cette amplification a été réalisée avec des amorces complémentaires de la séquence répétée CRISPR I-E et a généré un total de 1883 séquences spacer uniques qui a ensuite été comparé à un ensemble de spacers E. coli actuels constitué de 1599 séquences uniques. Cette comparaison a révélé 425 spacers communs. Des loci contemporains complets ou presque complets ont pu être reconstruits en utilisant des paires de spacers voisins. Dans l'ensemble, plusieurs loci CRISPR d'E. coli contemporains sont restés inchangés au cours des 40 000 dernières années, confirmant l'inactivité du module d'adaptation des systèmes CRISPR-Cas de type I-E dans cet organisme.

• Les spacers du CRISPRome des communautés naturelles de Thermus, Sulfolobus et Flavobacteries ciblent de préférence des virus isolés de la même source, avec différents systèmes CRISPR-Cas ciblant différents virus (Thermus : Chapitre III, Tableau 2 ; Flavobacteria : Chapitre II, Figure 5B ; Sulfolobus : Chapitre IV, Figure 1D). Ce résultat est en accord avec le ciblage local de spacers déjà rapporté pour de nombreux autres environnements, et semble être un phénomène général. La spécificité de différents modules d'adaptation à différents virus peut être une conséquence d'une plage d'hôtes étroite pour un virus ou de protéines anti-CRISPR codées par un virus.

• Les données CRISPRome de Flavobacterium et Sulfolobus (Chapitres II et IV) montrent un schéma phylogéographique, avec les ensembles de spacers provenant de sites d'échantillonnage géographiquement proches plus similaires que ceux provenant d'emplacements plus éloignés.

Ainsi, les ensembles de spacers de trois sites en Antarctique diffèrent considérablement les uns des autres, avec seulement une infime fraction des éléments spacers commune aux trois sites. La grande proportion de spacers communs entre les sites Druzhnaja et Progress est cohérente avec la proximité géographique de ces stations (Chapitre II, Figure 5B). Aucun recoupement n'a été détecté avec des spacers issus de génomes séquencés de flavobactéries, et seuls quelques phages de flavobactéries connus sont ciblés par des spacers provenant de l'Antarctique, suggérant l'existence de communautés virales distinctes dans l'Antarctique. De même, des spacers ont été trouvés en commun entre les CRISPRomes Sulfolobales des sources chaudes de Beppu et ceux des isolats de Sulfolobales issus du Japon (chapitre IV, figure 1A). Cela indique que la population de Sulfolobales de Beppu et celle représentée dans les isolats japonais a été infectée par des virus similaires.

Contrairement aux communautés naturelles de Flavobactéries et de Sulfolobales, le CRISPRome issu de celle de Thermus provenant de cultures d'enrichissement n'a montré aucune corrélation vis-à-vis de la distance géographique entre les sites d'échantillonnage (chapitre III, figure 2B). Cette observation demeure sans explication pour le moment. Il est possible que certaines propriétés physicochimiques de l'eau non-enregistrées lors de la collecte des échantillons en soient responsables. Un contrôle minutieux des paramètres écologiques de l'habitat sur les sites de collecte et l'extension de l'analyse présentée ici à d'autres communautés Thermus du monde entier pourraient aider à résoudre ce problème.

• Les virus de Sulfolobus SPV1 et SPV2 portent des mini-loci CRISPR avec 1 à 2 spacers uniquement. Ces mini-loci sont précédés par des séquences leader, similaires à celle précédant les loci CRISPR présents dans les génomes Sulfolobus, mais ne sont pas associés aux gènes cas.

Les positions relatives des mini-loci CRISPR contenant 2 spacers dans les génomes SPV1 et SPV2 sont identiques, mais les spacers correspondants sont différents, ce qui suggère un renouvellement actif des spacers. Les spacers des mini-loci ciblent des virus étroitement apparentés présents au sein de la même population. Le ciblage par des spacers transmis par le mini-loci viral représente un mécanisme distinct d'exclusion de surinfection par ces virus apparentés et semble favoriser la spéciation des virus d'archées. Dans l'échantillon environnemental initial et dans les enrichissements sur 10 jours, les spacers des longs loci CRISPR de l'hôte étaient les principaux contributeurs à l'immunité totale contre les virus SPV1 et SPV2. Cependant, après 20 jours, l'abondance de spacers des mini-loci a augmenté de façon substantielle, tandis que le nombre de spacers des longs locus a diminué, probablement en raison de la prédation de l'hôte par SPV1 et SPV2. De plus, les spacers des loci de l'hôte ciblent indistinctement SPV1 et SPV2 (à en juger par l'identité entre les spacers et les proto-spacers),

alors que les spacers des mini-loci sont spécifiques à SPV1 ou à SPV2. Le ciblage des spacers CRISPR favorise la microévolution des génomes viraux, alors que l'évitement de l'auto-ciblage par les mini-loci CRISPR favorise probablement la spéciation du virus.

Références

1 Mojica, F. J., Diez-Villasenor, C., Garcia-Martinez, J. & Soria, E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. J Mol Evol **60**, 174-182, doi:10.1007/s00239-004-0046-3 (2005).

2 Makarova, K. S. & Koonin, E. V. Annotation and Classification of CRISPR-Cas Systems. Methods Mol Biol **1311**, 47-75, doi:10.1007/978-1-4939-2687-9_4 (2015).

3 Koonin, E. V., Makarova, K. S. & Zhang, F. Diversity, classification and evolution of CRISPR-Cas systems. Curr Opin Microbiol **37**, 67-78, doi:10.1016/j.mib.2017.05.008 (2017).

4 Westra, E. R., van Houte, S., Gandon, S. & Whitaker, R. The ecology and evolution of microbial CRISPR-Cas adaptive immune systems. Philos Trans R Soc Lond B Biol Sci **374**, 20190101, doi:10.1098/rstb.2019.0101 (2019).

5 Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Comparative genomics of defense systems in archaea and bacteria. Nucleic Acids Res **41**, 4360-4377, doi:10.1093/nar/gkt157 (2013).

6 Iranzo, J., Lobkovsky, A. E., Wolf, Y. I. & Koonin, E. V. Evolutionary dynamics of the prokaryotic adaptive immunity system CRISPR-Cas in an explicit ecological context. J Bacteriol **195**, 3834-3844, doi:10.1128/JB.00412-13 (2013).

7 Gudbergsdottir, S. R., Menzel, P., Krogh, A., Young, M. & Peng, X. Novel viral genomes identified from six metagenomes reveal wide distribution of archaeal viruses and high viral diversity in terrestrial hot springs. Environ Microbiol **18**, 863-874, doi:10.1111/1462-2920.13079 (2016).

8 Held, N. L., Herrera, A., Cadillo-Quiroz, H. & Whitaker, R. J. CRISPR associated diversity within a population of Sulfolobus islandicus. PLoS One **5**, doi:10.1371/journal.pone.0012988 (2010).

9 Robles-Sikisaka, R. et al. Conservation of streptococcal CRISPRs on human skin and saliva. BMC Microbiol **14**, 146, doi:10.1186/1471-2180-14-146 (2014).

10 Pride, D. T. et al. Analysis of streptococcal CRISPRs from human saliva reveals substantial sequence diversity within and between subjects over time. Genome Res **21**, 126-136, doi:10.1101/gr.111732.110 (2011).

11 Kuno, S., Yoshida, T., Kaneko, T. & Sako, Y. Intricate interactions between the bloom-forming cyanobacterium Microcystis aeruginosa and foreign genetic elements, revealed by diversified clustered regularly interspaced short palindromic repeat (CRISPR) signatures. Appl Environ Microbiol **78**, 5353-5360, doi:10.1128/AEM.00626-12 (2012).

12 Andersson, A. F. & Banfield, J. F. Virus population dynamics and acquired virus resistance in natural microbial communities. Science **320**, 1047-1050, doi:10.1126/science.1157358 (2008).

13 Kunin, V. et al. A bacterial metapopulation adapts locally to phage predation despite global dispersal. Genome Res **18**, 293-297, doi:10.1101/gr.6835308 (2008).

14 Tyson, G. W. & Banfield, J. F. Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. Environ Microbiol **10**, 200-207, doi:10.1111/j.1462-2920.2007.01444.x (2008).

15 Held, N. L., Herrera, A. & Whitaker, R. J. Reassortment of CRISPR repeat-spacer loci in Sulfolobus islandicus. Environ Microbiol **15**, 3065-3076, doi:10.1111/1462-2920.12146 (2013).

16 Held, N. L. & Whitaker, R. J. Viral biogeography revealed by signatures in Sulfolobus islandicus genomes. Environ Microbiol **11**, 457-466, doi:10.1111/j.1462-2920.2008.01784.x (2009).

17 Kimura, S. et al. Incomplete Selective Sweeps of Microcystis Population Detected by the Leader-End CRISPR Fragment Analysis in a Natural Pond. Front Microbiol **9**, 425, doi:10.3389/fmicb.2018.00425 (2018).

18 Snyder, J. C., Bateson, M. M., Lavin, M. & Young, M. J. Use of cellular CRISPR (clusters of regularly interspaced short palindromic repeats) spacer-based microarrays for detection of viruses in environmental samples. Appl Environ Microbiol **76**, 7251-7258, doi:10.1128/AEM.01109-10 (2010).

19 Stern, A., Mick, E., Tirosh, I., Sagy, O. & Sorek, R. CRISPR targeting reveals a reservoir of common phages associated with the human gut microbiome. Genome Res **22**, 1985-1994, doi:10.1101/gr.138297.112 (2012).

20 Emerson, J. B. et al. Virus-host and CRISPR dynamics in Archaea-dominated hypersaline Lake Tyrrell, Victoria, Australia. Archaea **2013**, 370871, doi:10.1155/2013/370871 (2013).

21 Sorokin, V. A., Gelfand, M. S. & Artamonova, II. Evolutionary dynamics of clustered irregularly interspaced short palindromic repeat systems in the ocean metagenome. Appl Environ Microbiol **76**, 2136-2144, doi:10.1128/AEM.01985-09 (2010).

22 Childs, L. M., England, W. E., Young, M. J., Weitz, J. S. & Whitaker, R. J. CRISPR-induced distributed immunity in microbial populations. PLoS One **9**, e101710, doi:10.1371/journal.pone.0101710 (2014).