Biochemical Society Transactions (2020) **48** 257–269
https://doi.org/10.1042/BST20190662

Review Article

Detection of CRISPR adaptation

Anna Shiriaeva[1,2], Ivan Fedorov[1,3], Danylo Vyhovskyi[1] and Konstantin Severinov[1,2,4]

[1]Center of Life Sciences, Skolkovo Institute of Science and Technology, Moscow 121205, Russia; [2]Waksman Institute, Rutgers, the State University of New Jersey, Piscataway, NJ 08854, U.S.A.; [3]Institute of Gene Biology, Russian Academy of Sciences, Moscow 119334, Russia; [4]Center for Precision Genome Editing and Genetic Technologies for Biomedicine, Institute of Gene Biology, Russian Academy of Sciences, Moscow 119334, Russia

Correspondence: Konstantin Severinov (severik@waksman.rutgers.edu)

Prokaryotic adaptive immunity is built when short DNA fragments called spacers are acquired into CRISPR (clustered regularly interspaced short palindromic repeats) arrays. CRISPR adaptation is a multistep process which comprises selection, generation, and incorporation of prespacers into arrays. Once adapted, spacers provide immunity through the recognition of complementary nucleic acid sequences, channeling them for destruction. To prevent deleterious autoimmunity, CRISPR adaptation must therefore be a highly regulated and infrequent process, at least in the absence of genetic invaders. Over the years, ingenious methods to study CRISPR adaptation have been developed. In this paper, we discuss and compare methods that detect CRISPR adaptation and its intermediates *in vivo* and propose suppressing PCR as a simple modification of a popular assay to monitor spacer acquisition with increased sensitivity.

Introduction

CRISPR-Cas are diverse (two classes, six types [1–3]) prokaryotic adaptive immunity systems that protect cells from phages and other mobile genetic elements (MGEs) [4,5]. They consist of CRISPR arrays and CRISPR-associated *cas* genes [6,7]. CRISPR arrays are composed of identical or highly similar repeats separated by unique DNA sequences called spacers [6,7]. The total number of spacers in array varies from one to several hundreds [6,8]. The source of the vast majority (~93%) of spacers remains unknown, they constitute the 'dark matter' of CRISPR [9]. Most of the remaining spacers map to MGEs and can be regarded as memories of prior encounters that cells store in CRISPR arrays [9]. Upstream of the CRISPR loci, there is an AT-rich sequence called the 'leader' [7]. CRISPR arrays are transcribed from a promoter located in the leader and the primary transcript is processed into CRISPR RNAs (crRNAs) containing a single spacer and flanking sequences derived from repeats [10–17]. Cas proteins together with crRNAs form effector complexes (Cascade complex in the type I-E system of *Escherichia coli*) that recognize 'protospacers' — DNA or, sometimes, RNA sequences, complementary to a crRNA spacer [13,18–20]. Recognition of protospacers in MGEs leads to their destruction [18–20].

CRISPR immunity is built during CRISPR adaptation, a process which entails incorporation of new spacers in the array [4]. New spacers are typically incorporated at the boundary between the leader and the first repeat and, therefore, the chronological order of spacer acquisition matches the inverse order of spacers in the array [4,21,22]. For every acquired spacer, a new copy of repeat is generated [4,21,22]. Two most conserved Cas proteins, Cas1 and Cas2, common to almost all CRISPR-Cas systems, catalyze integration of spacer precursors (prespacers) into arrays [23–25]. Generally, the acquisition of spacers is not specifically targeted to MGEs and thus spacers from cell's own genome can also be acquired [23,26]. This can result in auto-immune response inhibiting cell growth [27–29]. Not surprisingly, CRISPR adaptation is a tightly controlled process that normally proceeds with very low efficiency and can be difficult to detect both in natural settings and in laboratory experiments. Several methods for the detection of CRISPR adaptation have been developed and helped to shed light on molecular mechanisms governing spacer choice. These methods and their limitations are discussed below.

Received: 15 November 2019
Revised: 26 December 2019
Accepted: 3 January 2020

Version of Record published:
3 February 2020

Biochemical Society Transactions (2020) **48** 257–269
https://doi.org/10.1042/BST20190662

Selection-based methods of detection of CRISPR adaptation in individual cells or clones

An obvious way to detect the acquisition of a new spacer is to amplify the leader-proximal end of CRISPR array with a pair of primers: one matching the leader, and another matching an internal, pre-existing spacer [23,30,31]. Since new spacers are usually incorporated in front of the first, leader-proximal repeat, and result in repeat duplication [4,21–23], detection of PCR-products extended by integral number of spacer-repeat units reveals CRISPR adaptation events. However, since spacer acquisition can be very infrequent, specific selection of adapted cells is required (Figure 1). Examples of such selections include obtaining colonies of BIMs (bacteriophage insensitive mutants) (Figure 1A) [4,21,22,31,32] or PIMs (plasmid interfering mutants) (Figure 1B) [30,31,33]. These methods are cheap and do not require genetic manipulation of cells under study but they are biased towards interference-proficient spacers acquired from MGEs and thus cannot be used to detect spacers that do not lead to interference against MGEs or lead to self-interference due to acquisition of a spacer from cell's own genome (depending on the CRISPR-Cas subtype, when interference is inactivated, such spacers can constitute from 2 to 99% of acquired spacers [23,26,34,35]).

Two powerful experimental systems have been developed to overcome these limitations and increase the sensitivity of the detection of spacer acquisition events [36,37] (Figure 1C,D). Both systems are based on a reporter gene introduced upstream of the leader sequence of a specifically designed miniaturized *E. coli* CRISPR array. The reporter is transcribed from a promoter located downstream of the array in a direction opposite to the direction of leader-initiated CRISPR array transcription. The resulting mRNA includes a start codon followed by the leader-CRISPR array segment (cloned in reverse orientation) and the sequence of the reporter which does not have a translational start of its own. In cells with unexpanded CRISPR arrays translation of the reporter ORF is prevented due to an in-frame stop codon within the leader. Insertion of an additional 61-bp long unit (33-bp spacer/28-bp repeat) changes the reading frame and allows the synthesis of the reporter leading to either chloramphenicol resistance [36] (Figure 1C) or fluorescence [37] (Figure 1D) of cells that acquired a spacer. Rare chloramphenicol-resistant colonies can be directly screened for CRISPR array expansion by PCR. With the fluorescent protein-based system, live fluorescent microscopy is used to observe and quantify cells that acquired spacers [37]. Though this has not been implemented yet, the use of FACS (fluorescence-activated cell sorting) should allow one to enrich the population of cells with expanded arrays for downstream analysis. With both systems, the acquisition of spacers that carry stop codons located in the reading frame of the reporter remains undetected. Likewise, incorporation of more than one spacer-repeat unit or incorporation of a single non-standard spacer that fails to restore the reporter reading frame will be undetected. Finally, CRISPR-Cas systems where incorporation of a standard spacer-repeat unit does not shift the reading frame (i.e. introduces an insertion whose length is $n \times 3$ bp, where n is an integral number of nucleotides) cannot be studied.

Detection of CRISPR adaptation in cell populations

In early studies of adaptation, the sequences of newly acquired spacers were determined for individual colonies by Sanger sequencing [4,21–23,30–33]. To analyze millions of CRISPR arrays in a single experiment, high-throughput sequencing (HTS) is usually used [38–40]. This allows one to study biases in spacer length, the distribution of corresponding protospacers along different DNA sources and their nucleotide composition [34,35,38–52]. In principle, with sufficient sequencing depth, HTS of total genomic DNA purified from a culture should reveal reads corresponding to expanded arrays [53]. In a model system of *E. coli* cultures overproducing the Cas1–Cas2 adaptation protein complex and transformed with spacer-sized oligonucleotides, ~350× genomic coverage allowed to confidently detect CRISPR array expansion that occurred in ~10% of cells [54]. Moreover, rarer off-target integration events elsewhere in the genome were also detected [54].

While clearly powerful and unbiased, the shotgun sequencing approach requires high sequencing coverage and provides very low (dozens) numbers of reads corresponding to expanded arrays making it unsuitable for studies aimed at qualitative understanding of spacer selection preferences [54]. Therefore, the common strategy is to prepare PCR amplicons of arrays from cultures undergoing CRISPR adaptation and then subject them for HTS [38–40]. Gel-electrophoresis is used to separate amplicons of initial, unexpanded CRISPR arrays (+0) from those that acquired one (+1), two (+2) or more spacer-repeat units (Figure 2A).

The main problem with PCR-based in-culture methods of detection of CRISPR adaptation is their low sensitivity due to more efficient amplification of shorter (and, in most interesting cases, much more abundant)

Biochemical Society Transactions (2020) **48** 257–269
https://doi.org/10.1042/BST20190662

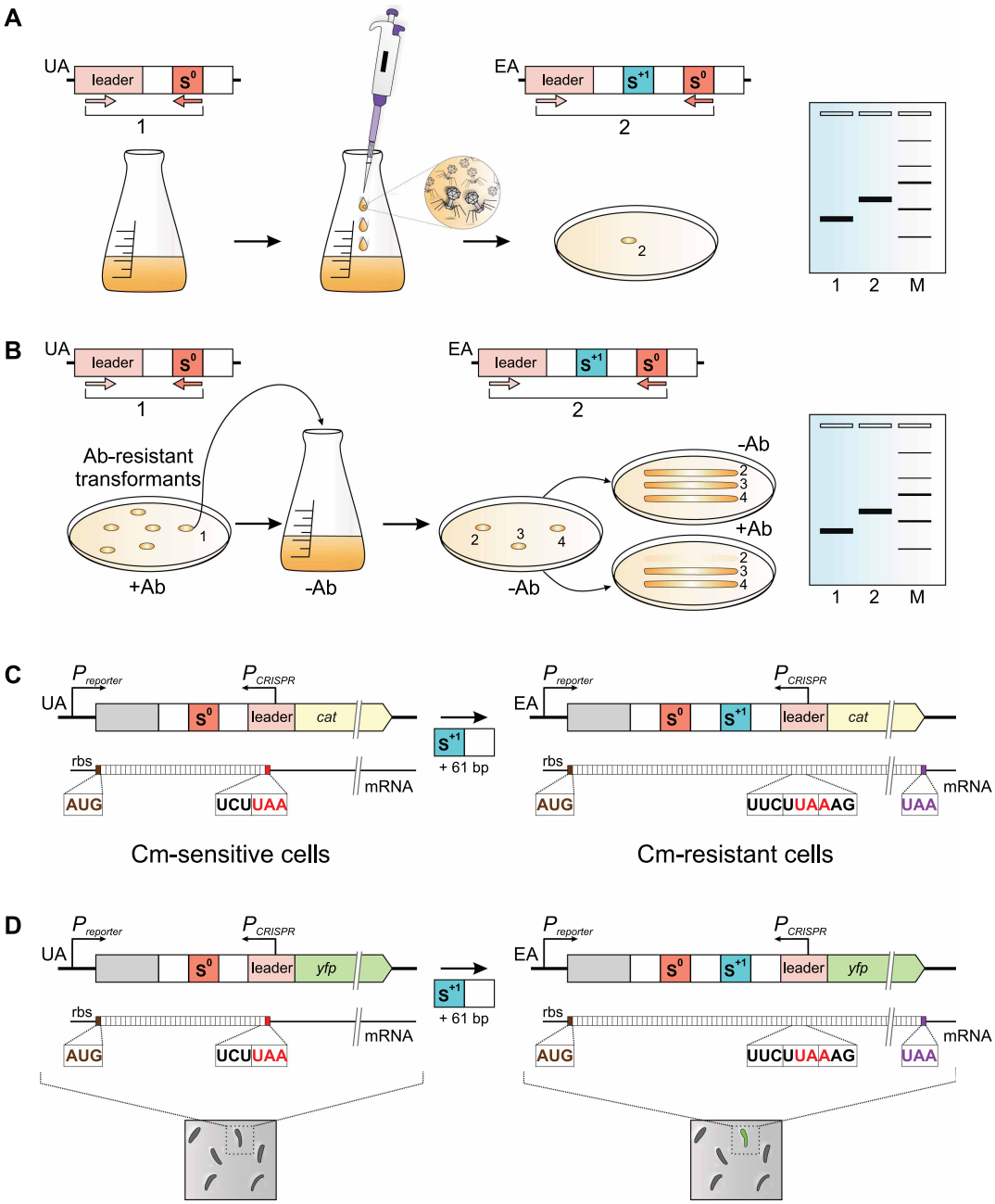


Figure 1. Selection-based methods to detect CRISPR adaptation. Part 1 of 2

Detection of acquisition of interference-proficient spacers (S^{+1}) in bacteriophage insensitive mutants (**A**) or in cells that lost plasmids (**B**). The structures of unexpanded (UA) CRISPR arrays in cells from an initial culture or of expanded (EA) arrays in cells that either survived phage infection or lost the plasmid are shown at the top. The leader (light peach rectangle), CRISPR repeats (white rectangles), pre-existing spacer S^0 (dark peach rectangle), and newly acquired spacer S^{+1} (turquoise rectangle) are shown. Acquisition of spacers is tested by PCR with primers matching the leader (light peach arrow), and the S^0 spacer (dark peach arrow). Amplicons from expanded and unexpanded arrays are shown as brackets below the primers. In **A**, a colony formed after phage infection is directly tested by PCR and results of agarose gel electrophoresis of amplicons obtained with starting cells (1) and the phage-resistant colony (2) are schematically shown on the right. In **B**, a liquid culture is inoculated with cells from an antibiotic-resistant (Ab-resistant) plasmid-bearing colony (1) with an unexpanded CRISPR array. After growth and plating on non-selective (–Ab) medium, the presence of the plasmid is tested by streaking colonies (2, 3, 4) on plates with and without Ab. Cells from the antibiotic-sensitive colony (2) are further tested by PCR to reveal a spacer acquisition event. M, molecular-weight size marker. Detection of CRISPR adaptation using *cat* (**C**) or *yfp* (**D**) reporter systems. Insertion of an

Biochemical Society Transactions (2020) **48** 257–269
https://doi.org/10.1042/BST20190662

Figure 1. Selection-based methods to detect CRISPR adaptation. Part 2 of 2

inverted leader/single-spacer CRISPR array disrupts the reading frame of *cat* (**C**) or *yfp* (**D**) reporter genes. The elements of CRISPR arrays are shown as in **A** and **B**. $P_{reporter}$ — a promoter directing the synthesis of reporter mRNA; P_{CRISPR} — a promoter directing the synthesis of pre-crRNA. The composition of reporter mRNAs is shown below. Each rectangle corresponds to one codon; rbs — a ribosome binding site; AUG — a start codon, UAA — a stop codon located in-frame with the AUG. Translation of mRNAs transcribed from unexpanded CRISPR reporters stops at the UAA codon in the leader sequence resulting in (**C**) chloramphenicol-sensitivity (Cm-sensitivity) or (**D**) absence of fluorescence. Insertion of a 61-bp (33-bp spacer/28-bp repeat) unit into the CRISPR array shifts the in-leader UAA stop codon out of the frame and restores the reporter gene reading frame resulting in chloramphenicol-resistance (**C**) or fluorescence (**D**) of cells with expanded CRISPR arrays.

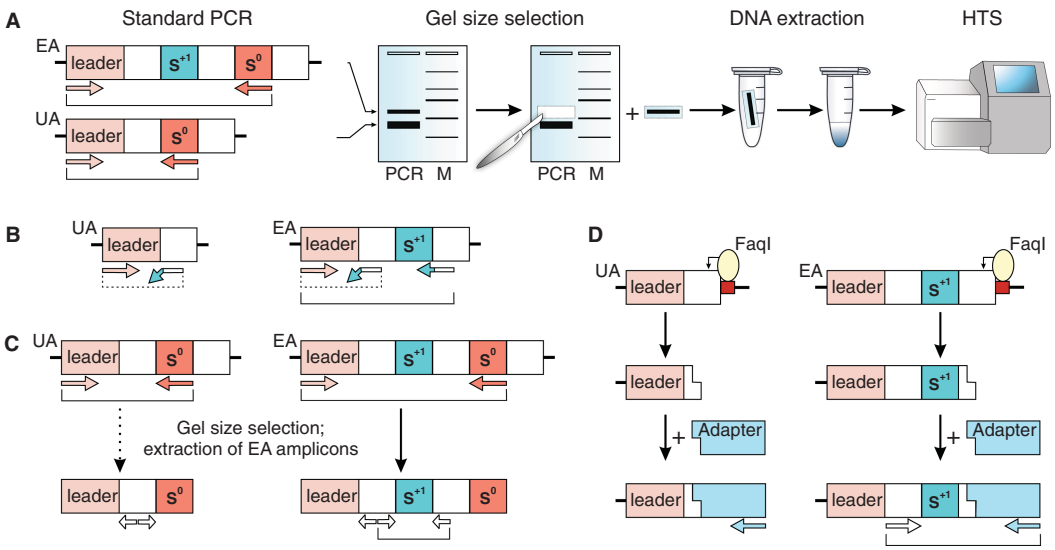


Figure 2. PCR-based methods of studying CRISPR adaptation in bacterial cultures.

In all panels, unexpanded CRISPR arrays are labeled UA; expanded arrays — EA, the leader is shown as a light peach rectangle; repeats — as white rectangles; the pre-existing spacer S^0 — as a dark peach rectangle; the newly acquired spacer S^{+1} — as a turquoise rectangle. Primers are shown as arrows below CRISPR arrays. A flipped end of a primer arrow indicates the absence of complementarity between the primer's end and the template. The brackets below primer pairs indicate PCR products (dashed lines correspond to low-efficiency PCR with degenerate primers whose 3′ ends are not complementary to the template). (**A**) Standard PCR-based spacer acquisition assay. The leader-proximal part of the CRISPR array is amplified with a leader-specific primer (light peach arrow) and a primer annealing to the S^0 spacer (dark peach arrow). The products of PCR are resolved on an agarose gel. M, molecular-weight size marker. The band corresponding to expanded CRISPR arrays (EA) is purified from the gel and subjected to HTS (high-throughput sequencing). (**B**) Selective amplification of expanded arrays by PCR with a leader-specific primer and a degenerate repeat-specific primer (white arrows with turquoise arrowheads). All nucleotides of the degenerate primer, except for the last one, are annealed to the repeat placing the 3′-end nucleotide opposite the last nucleotide of the leader or the last nucleotide of the acquired spacer. The last position of the degenerate primer includes one of three nucleotides that are not complementary to the leader's last nucleotide enabling efficient amplification only when the last nucleotide of the acquired spacer is different from the last nucleotide of the leader. (**C**) Selective amplification of expanded arrays using CAPTURE. Top, CRISPR arrays are amplified and the band corresponding to expanded CRISPR arrays is purified after electrophoresis as in **A**. Bottom, to further increase the percentage of expanded arrays in sequencing libraries, the products of the first PCR are reamplified with primers (white arrows) annealing to repeats. The expanded arrays are selectively amplified since the products of the first-stage PCR of unexpanded and expanded arrays contain one and two repeats, correspondingly. (**D**) Selective amplification of expanded arrays using SENECA. An FaqI endonuclease recognition site (maroon rectangle) is introduced immediately following the first CRISPR repeat. FaqI cleaves DNA upstream of its recognition site creating sticky ends in the repeat sequence. A matching adapter is ligated to this sticky end and PCR amplification with adapter-specific (light blue) and repeat-matching (white) primers is performed.

unexpanded CRISPR arrays [55]. In the case of *E. coli* type I-E system, the standard method allows one to reliably detect expanded arrays amplicons only in cultures which contain, in our experience, at least 5% of adapted cells. Several modifications aimed to increase the sensitivity have been developed. The simplest one relies on amplification with a leader-specific primer and a primer matching a newly acquired spacer whose sequence is known [40]. After calibration to the amount of PCR product amplified from a region outside of CRISPR array and reflecting the total number of DNA molecules in the sample, this method can be used to determine the efficiency of adaptation by qPCR [56]. The obvious drawback of this method is that it requires prior knowledge about acquired spacer(s) and thus cannot be applied to study spacer acquisition in systems with unknown adaptation preferences. However, it is very powerful when studying acquisition from spacer-sized oligonucleotides transformed into cells [51,54,57].

Another modification uses a leader-specific primer and 'degenerate' primers matching the repeat sequence and containing one extra 3′-end nucleotide (Figure 2B) [42]. The additional position contains, in equal proportions, three nucleotides except for the one complementary to the last nucleotide of the leader. While amplification products are only expected if (i) a spacer has been acquired and (ii) its last nucleotide is different from the last nucleotide of the leader, in practice amplicons from unexpanded arrays are still observed [42]. The method was reported to detect as little as 0.01% of cells with expanded arrays [42]. However, by design, up to ∼25% of acquired spacers remain undetected. Moreover, the method is effective only when applied to engineered miniaturized CRISPR arrays reduced to just one repeat, since multiple amplification products from unexpanded arrays with multiple repeats are produced which can't be distinguished from amplicons from expanded arrays.

Reamplification of gel-purified amplicons of expanded arrays allows one to increase the sensitivity of detection of CRISPR adaptation in cell cultures [58]. Products of standard amplification are separated by gel-electrophoresis and purified. The reamplification step is repeated until a fragment of expected length becomes clearly visible on the gel (Figure 2C). The use of automated BluePippin system (agarose gel electrophoresis with automated elution for size selection) allows one to improve the quality of separation, reduce contamination from unexpanded arrays amplicons, and increase the reproducibility of analysis. Even when amplicons of expanded arrays are invisible after first electrophoretic separation, DNA extracted from the corresponding position of the gel can be used for reamplification. Depending on the set of primers designed for reamplification ('internal', 'degenerate', or 'repeat-specific'), this method is reported to detect, correspondingly, 1, 0.01, and 0.1% of cells with expanded arrays within *E. coli* cultures. 'Internal' primers can be either the same as the ones used during initial amplification, or a leader-specific 'nested' reamplification primer annealing closer to the array can be used to increase specificity and avoid amplification of non-CRISPR DNA. In either case, amplification of unexpanded arrays co-purified with expanded ones is not suppressed. 'Degenerate' primers selectively suppress unexpanded array reamplification as described above. Reamplification with 'repeat-specific' primers relies on the fact that amplicons corresponding to expanded arrays have two repeats after the first PCR step (Figure 2C). Thus, amplification with primers matching the halves of the repeat sequence yields PCR product only for expanded arrays.

The SENECA [59] pipeline selectively amplifies expanded CRISPR arrays. At the heart of the method is the construction of a plasmid-borne CRISPR array with an FaqI endonuclease recognition site immediately downstream of a miniaturized CRISPR 'array' consisting of a single repeat preceded by the leader sequence (Figure 2D). Unlike most Type II restriction endonucleases, FaqI, a Type II-S enzyme, cleaves DNA outside of its recognition site generating a sticky end. The CRISPR array used in SENECA is designed such that the recognition of the FaqI site leads to cleavage in the upstream repeat. An Illumina adapter with a sticky end complementary to that generated by FaqI is ligated and PCR with a pair of primers, one complementary to repeat and another — to adapter, selectively amplifies expanded arrays, since the initial repeat sequence is lost after the FaqI treatment and thus amplicons from unexpanded arrays are not amplified. While the published SENECA protocol is based on the use of FaqI, other Type II-S restriction endonucleases could conceivably be used in lieu of FaqI.

The initial SENECA protocol was applied for analysis of spacers acquired by the type III CRISPR-Cas system of *Fusicatenibacter saccharivorans* heterologously expressed in *E. coli* [59]. While ∼700-fold enrichment for expanded arrays in a sequencing library was reported, the sensitivity of SENECA in terms of detecting the percentage of cells with expanded arrays was not determined. To benchmark SENECA against other methods to detect CRISPR adaptation, we constructed an *E. coli* strain with a spacer containing FaqI recognition site, S^{FaqI}, incorporated into the genomic type I-E CRISPR array (Figure 3A and Supplementary Figure S1A). Except for S^{FaqI}, the strain is isogenic to KD263, a strain with *cas* genes under control of inducible promoters and a miniaturized CRISPR array with a single spacer (Supplementary Fig. S1A and Supplementary Table S1). KD263 has

Biochemical Society Transactions (2020) **48** 257–269
https://doi.org/10.1042/BST20190662

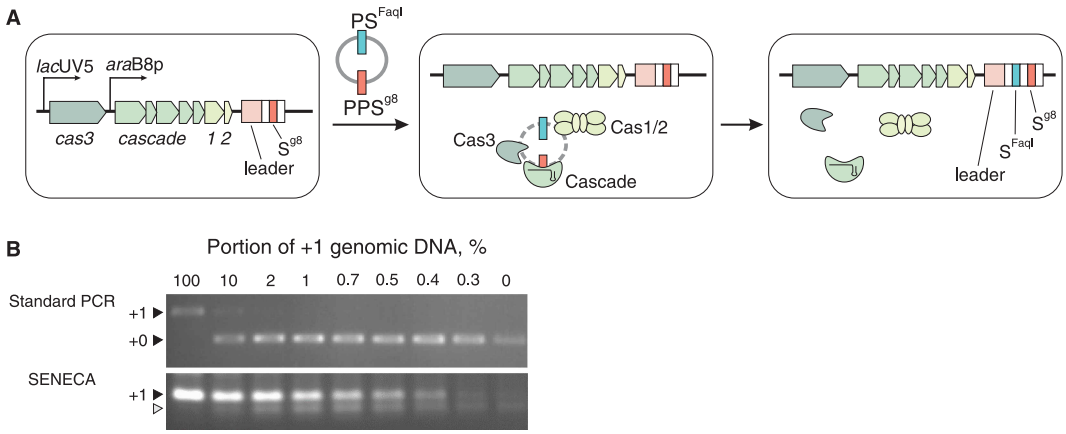


Figure 3. Assessing the sensitivity of SENECA to detect primed adaptation in the type I-E CRISPR-Cas system of *E. coli*.

(**A**) To create an *E. coli* strain suitable for detection of primed adaptation by SENECA, parental *E. coli* KD263 cells containing inducible *cas* genes and a CRISPR array with a single spacer S^{g8} are transformed with a plasmid carrying a priming protospacer PPS^{g8} matching S^{g8} and an FaqI site introduced in a previously characterized highly used 'hot' protospacer (PS^{FaqI}) shown as a turquoise rectangle. Upon the recognition of PPS^{g8} by the Cascade-crRNA effector complex, primed adaptation occurs and new spacers originating from the plasmid are integrated into the CRISPR array [31,41,65]. Among colonies that lost the plasmid and expanded their CRISPR array (see Figure 1B), a clone that acquired the S^{FaqI} spacer is selected. (**B**) Genomic DNA purified from parental cells carrying the S^{FaqI} spacer or a derivative with an expanded array (+1) was mixed at ratios indicated and the state of CRISPR arrays was assessed by standard PCR (see Figure 2A) or SENECA (see Figure 2D). The percentage of the +1 genomic DNA in the sample is indicated above the gel. At the left-hand side, amplicons corresponding to unexpanded and expanded arrays revealed by standard PCR are labeled +0 and +1, correspondingly. The expanded array amplicon generated by SENECA is also labeled +1. The lower minor band (gray triangle) corresponds to the results of linear amplification of the parental array [59].

been extensively used to study various aspects of CRISPR adaptation [41,49,56,60–64] and adaptation efficiencies of up to 50–80% have been reported at conditions of priming [41,62]. To compare the sensitivity of standard PCR detection and SENECA, DNA prepared from cells with unexpanded CRISPR array containing the S^{FaqI} spacer was mixed in different proportions with DNA prepared from cells containing an additional spacer-repeat unit (Figure 3B). With the standard protocol, expanded arrays were clearly detected when 10% of DNA had an additional spacer. With SENECA, an expected amplification product was still seen when DNA with expanded array constituted 0.4% of the total, a ~25-fold increase in sensitivity (Figure 3B). To test if SENECA introduces biases in amplification of individual spacers we analyzed spacers acquired during type I-E primed adaptation by standard protocol or SENECA (Supplementary Fig. S1). No changes in spacer selection patterns were detected between the two methods (Supplementary Fig. S1C).

Suppressing PCR

As yet another alternative, we sought to increase the efficiency of rare adaptation events detection by selective amplification of longer fragments with suppressing PCR (supPCR) that relies on amplification of fragments flanked by ~30-nt inverted terminal repeats (ITR) [66,67]. DNA fragments containing ITR form pan-like structures due to intra-molecular hybridization (Figure 4A). Since shorter fragments are more likely to form pan-like structures, annealing of a primer identical with ITR and used at the second, reamplification stage is inhibited [68]. To determine whether supPCR can increase the sensitivity of spacer acquisition detection, we used standard leader and first-spacer specific primers extended with ITR. As can be seen from Figure 4B and Supplementary Figure S2, preferential amplification of expanded CRISPR arrays at the second PCR stage with an ITR primer increased the sensitivity of detection of adaptation events by ~10-fold.

The ratio of DNA template to the ITR primer used during the second stage is a critical parameter for successful supPCR, as suppression relies on competition between intra-molecular hybridization of ITRs and

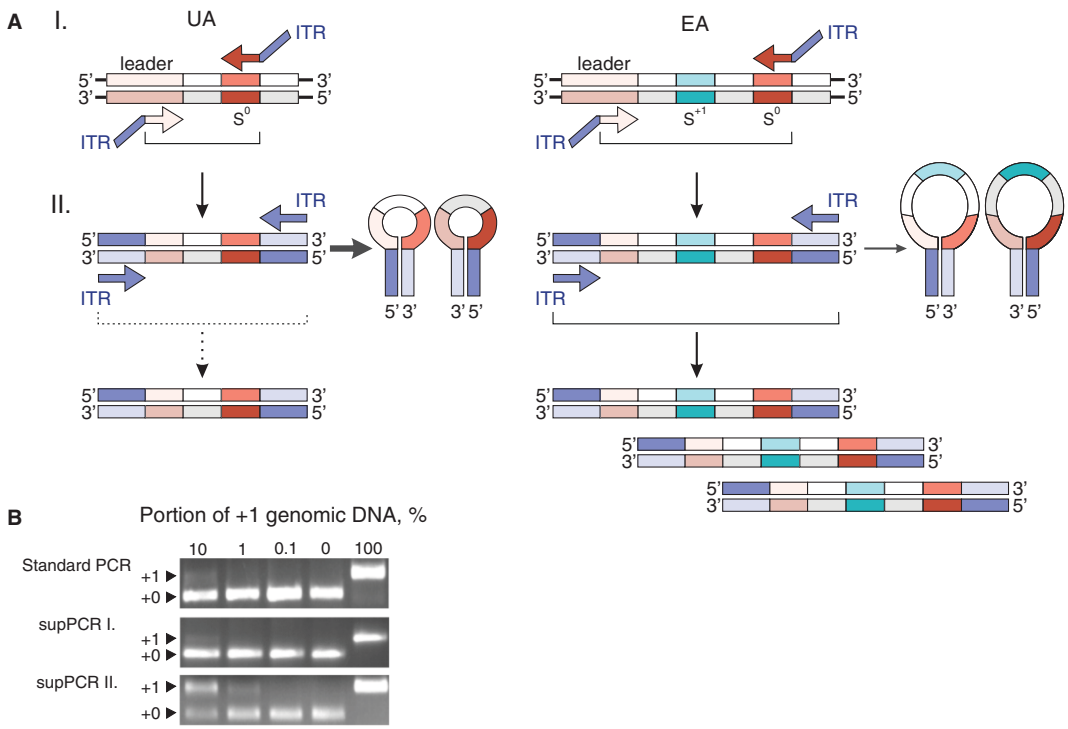


Figure 4. SupPCR increases the sensitivity of detection of CRISPR adaptation.

Schematic representation of the supPCR assay for detection of CRISPR adaptation is presented in **A**. Both strands of expanded (EA) and unexpanded (UA) CRISPR arrays are shown. The first-stage PCR (I) employs one primer matching the CRISPR array leader (beige arrow), and another one matching the leader-proximal pre-existing spacer S^0 (maroon arrow). Both primers carry at their 5′ ends identical 30-nt sequences (shown in lilac) that become ITRs after amplification. The second PCR stage (II) is carried using a primer whose sequence matches that of the ITR. Amplicons corresponding to expanded arrays are less likely to undergo intra-molecular hybridization, and, therefore, more effectively amplified with the ITR primer. Primers are shown as arrows below or above CRISPR arrays. The brackets below CRISPR arrays indicate regions amplified by PCR (dashed line corresponds to low-efficiency PCR on the UA template). (**B**) Genomic DNA purified from an *E. coli* strain with one additional spacer (+1) was mixed in indicated proportions with DNA from a strain with an unexpanded (+0) array. The mixtures were used as templates for supPCR. Products of 'standard' amplification (Figure 2A), as well as products of the first (I) and second (II) stages of supPCR separated by agarose gel electrophoresis are shown.

annealing of the ITR primer for target product amplification (Supplementary Fig. S2A) [68]. This depends on the length of a spacer-repeat unit incorporated in the extended array and has to be determined experimentally. Because the same ITR can be used for amplification of CRISPR arrays from different systems, other parameters, such as annealing temperature and the length of suppression primer remain the same for different amplification reactions. We applied supPCR to analyze spacers acquired during primed adaptation by the *E. coli* type I-E system and detected similar patterns of acquired spacers for the standard protocol and supPCR (Supplementary Fig. S3). In addition to the *E. coli* type I-E system, we have successfully used supPCR to reveal spacer acquisition in the *Thermus thermophilus* Type III CRISPR arrays (our unpublished observations).

Detection of spacer precursors *in vivo*

The methods described above detect the final result of CRISPR adaptation, spacers integrated into the CRISPR array. New spacers are derived from 'prespacers', short pieces of DNA that are generated inside the cell by various processes and then channeled into CRISPR arrays by Cas1 and Cas2 proteins [24,25,34,50–52,69–75]. The knowledge about prespacer structure is essential for the understanding of CRISPR adaptation preferences and the ability to control it. *In vitro*, the Cas1–Cas2 complex preferentially uses double-stranded DNA fragments with blunt ends or 3′-end overhangs as substrates for integration [24,70,76–78]. Shipman et al. were the

Biochemical Society Transactions (2020) **48** 257–269
https://doi.org/10.1042/BST20190662

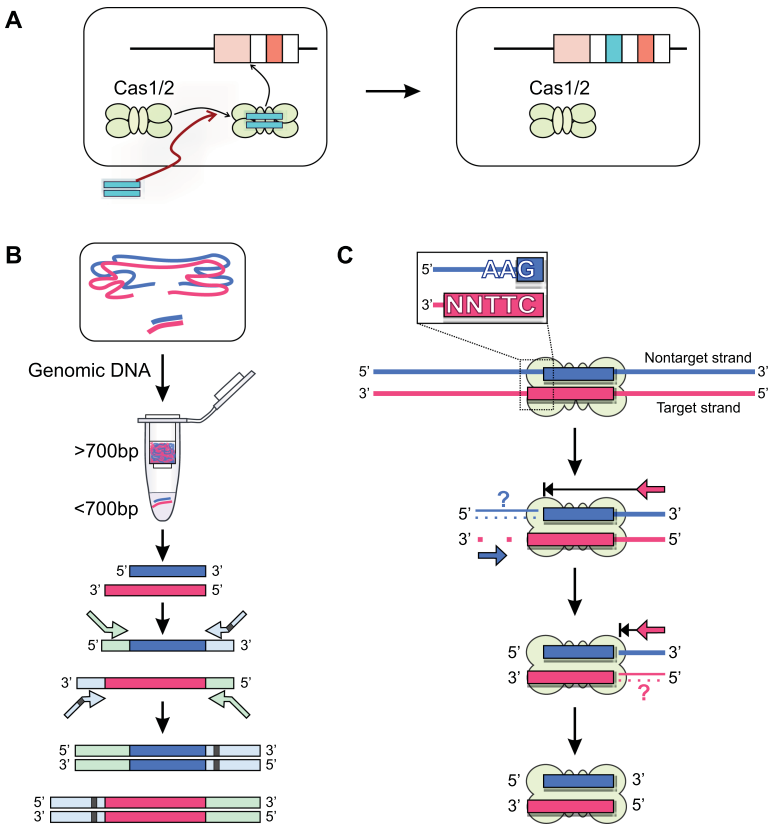


Figure 5. Methods of analysis of prespacer generation, structure, and acquisition efficiency.
(**A**) Electroporation of double-stranded oligonucleotides (turquoise) into cells with a CRISPR array (the CRISPR leader - a light peach rectangle; repeats - white rectangles; the pre-existing spacer - a dark peach rectangle). Cas1 and Cas2 (light green ovals) form a complex that binds the oligonucleotides that can function as prespacers and integrates them into the CRISPR array as new spacers (turquoise rectangle). (**B**) The FragSeq pipeline. High-molecular mass DNA is purified from cells together with short fragments (the strands of double-stranded DNA are shown in blue and red). Short DNA fragments are recovered, denatured and ligated to single-stranded adapters (light blue and light green rectangles). The resulting molecules are amplified with barcoded primers (light blue and light green arrows). (**C**) Prespacer generation during type I-E primed adaptation in *E. coli* as revealed by FragSeq [79] and primer extension [63]. A prespacer (blue and red rectangles) in DNA (blue and red lines) is bound by the Cas1–Cas2 complex (light green). The inset shows nucleotide composition at the PAM-derived end of the prespacer as revealed by FragSeq. The red strand upstream of the PAM is degraded (dotted line) preventing annealing of a primer used for primer extension (blue arrow). The blue strand is incised between AA and G of the AAG PAM. Primer extension reaction with a primer annealing to the blue strand (red arrow) downstream of the prespacer generates two extension products (shown as thin black arrows), which implies that the PAM-derived 5′ end is produced first, and that the PAM-distal 3′ end is generated via an endonucleolytic cleavage event. The mechanism of generation of prespacer 5′ ends is not known (double solid/dashed lines with a question mark).

first to test prespacer requirements for integration *in vivo* [57]. They demonstrated that double-stranded synthetic oligonucleotides electroporated into *E. coli* cells expressing type I-E *cas1* and *cas2* can be trimmed to the size of a canonical *E. coli* spacer (33 bp) and integrated into the CRISPR array (Figure 5A) [57]. The presence of consensus PAM sequence (AAG/CTT) increased the acquisition efficiency of prespacers ∼5-fold and ensured specific orientation of integration with the PAM-derived G/C immediately following the first repeat [57].

We developed a method for strand-specific HTS of short DNA fragments generated *in vivo* that we called FragSeq and that allowed us to detect prespacers in *E. coli* cells undergoing primed CRISPR adaptation (Figure 5B) [79]. The method uses protocols that avoid the loss of short fragments during purification of DNA from cell lysates (phenol–chloroform DNA purification) and size selection of spacer-sized fragments prior to

library preparation. Our experience shows that even if as little as 0.1% of genomic DNA is randomly fragmented during DNA purification and size selection, it creates background noise that complicates detection of specific signal from short DNA fragments present in the cell. To avoid this complication the FragSeq experiment should be designed in such a way that a specific signal from a particular region of the genome/plasmid from which prespacers are expected is compared with a control region non-specifically fragmented during downstream handling. In our case, we achieve this using self-targeting *E. coli* cells, in which a crRNA directs the interference machinery to recognize a protospacer positioned within bacterial genome. Upon induction of self-targeting cells stop dividing but remain alive for at least several hours. To ensure that fragments with different end structures are detected, single-strand specific sequencing is required to correctly determine the length of each prespacer strand, meaning that adapters need to be ligated to both single- and double-stranded ends of prespacer fragments at the library construction stage. We used a combination of thermostable 5′ App DNA/RNA ligase (NEB) which is a K97A variant of RNA ligase from *Methanobacterium thermoautotrophicum* [80] and T4 RNA ligase 1 (NEB). However, we observed a strong bias by the 5′ App DNA/RNA ligase to ligate adapters to DNA fragments carrying a CGN motif on their 3′ ends, which skews the libraries. It is thus possible that the less biased T4 RNA ligase 1 alone may be better suited for library construction. Alternatively, commercial kits for single-strand specific sequencing such as Accel-NGS 1S Plus DNA Library Kit (Swift Biosciences) can be used. But it should be considered that these kits include a step of addition of a low-complexity tail onto 3′ ends, which can complicate the mapping of prespacer ends.

In our work, we used FragSeq to analyze prespacers formed during primed adaptation [79]. The hallmark of primed adaptation by the type I-E CRISPR-Cas system is the acquisition of spacers with the sequence of the nontranscribed strand originating from the nontarget strand of the DNA degraded by CRISPR interference machinery [30,31]. In line with this, we detected 32–34-nt fragments (starting with G or AG on their 5′ ends) that originated from the nontarget strand upstream of the priming protospacer (PPS) due to incision within the AAG motif [79]. Longer 37–38-nt fragments originated from the target strand and had a CTTNN motif on their 3′ ends [79]. Using the approach of Shipman et al. we tested complementary pairs of the most abundant fragments and concluded that prespacers with a blunt PAM-distal end, a 33–34-bp double-stranded region and a 4–3-nt 3′-end overhang on the PAM-derived end are integrated into the CRISPR array with the highest efficiency [79]. It thus appears that the asymmetry in type I-E prespacer structure contributes to specific orientation of prespacer integration since the PAM-distal 3′ end is processed (becomes blunt) earlier than the PAM-derived end and, therefore, has higher chances to be engaged into integration at the leader-repeat boundary.

Earlier, Musharova et al. [63] performed primer extension assay on total genomic DNA purified from *E. coli* cells undergoing primed adaptation to test if there are any specific cleavage sites near two protospacers corresponding to frequently acquired spacers. The reaction with a primer annealed to the nontarget strand downstream of the protospacer sequence revealed two products, one corresponding to DNA cleaved within the AAG PAM and another — at the PAM-distal boundary (Figure 5C) [63]. This observation is fully consistent with the detection of 32–34-nt fragments by FragSeq. Though 37–38-nt CTT-containing fragments were revealed by FragSeq, no primer extension products were detected for the target strand. The possible explanation for this discrepancy could be that while at least the PAM-distal 3′ end of a prespacer is produced via endonucleolytic cleavage of the nontarget strand, the PAM-derived 3′ end is produced due to exonucleolytic cleavage of the target strand upstream of the protospacer resulting in loss of the sequence complementary to the primer used in Musharova et al. (Figure 5C).

Perspectives

- **Importance of the field.** CRISPR adaptation is a complex multistep process that remains the least understood stage of adaptive prokaryotic immunity. While most practical applications of CRISPR are built around the interference stage, understanding of adaptation, in addition to providing fundamental insights into the process, will allow one to efficiently create cells with specific resistance profiles and record time-resolved ‘memories’ of specific events. These developments will not be possible without highly sensitive methods to reveal rare spacer acquisition events.

Biochemical Society Transactions (2020) **48** 257–269
https://doi.org/10.1042/BST20190662

- **A summary of the current thinking.** Despite considerable effort, in the absence of selection, the best available methods can detect adaptation when cells with expanded arrays constitute at least 0.01% of the population. These levels are likely several orders of magnitude above those observed in natural settings, where high levels of CRISPR adaptation, at least in the absence of viral infection, should be highly unfavorable because of subsequent deleterious autoimmunity. Indeed, most of the currently used methods are used to study artificial systems where adaptation is activated by overexpression of *cas* genes responsible.

- **Future directions.** Since some of the available methods, such as SENECA, the use of degenerate primers, and supPCR are orthogonal to each other, combining them may help to increase sensitivity somewhat. In the absence of further developments, very deep sequencing will probably become the method of choice to study adaptation, especially in organisms for which no methods of genetic manipulation are available. Development and application of methods like FragSeq to study short nucleic acids inside the cells should allow complementary studies of prespacers. Of most interest will be comparative analyses of prespacers generated by different CRISPR-Cas systems types and establishment of relationship between prespacer generation and DNA maintenance processes in the cell. Selection of prespacers is virtually unstudied. New sensitive and high-throughput methods will have to be developed to uncover the details of this earliest stage of generation of adaptive immunity.

Abbreviations

Ab, antibiotic; BIMs, bacteriophage insensitive mutants; CAPTURE, CRISPR adaptation PCR technique using reamplification and electrophoresis; Cm, chloramphenicol; CRISPR, clustered regularly interspaced short palindromic repeats; EA, expanded CRISPR arrays; FACS, fluorescence-activated cell sorting; HTS, high-throughput sequencing; ITR, inverted terminal repeats; MGEs, mobile genetic elements; PAM, protospacer adjacent motif; PIMs, plasmid interfering mutants; PPS, priming protospacer; PS, protospacer; rbs, ribosome binding site; S, spacer; SENECA, selective amplification of expanded CRISPR arrays; UA, unexpanded CRISPR arrays.

Author Contributions

K.S. conceived the article. A.S., K.S., I.F., and D.V wrote the manuscript. A.S., I.F., and D.V prepared the figures. D.V. performed SENECA experiments and data analysis. I.F. performed supPCR experiments and data analysis.

Funding

This work was supported by grant 075-15-2019-1661 from the Ministry of Science and Higher Education of the Russian Federation and NIH RO1 grant GM104071 to KS.

Acknowledgements

We dedicate this paper to the memory of Dr. Ekaterina Savitskaya who played a central role in CRISPR research in our laboratory.

Competing Interests

The authors declare that there are no competing interests associated with the manuscript.

References

1 Makarova, K.S., Wolf, Y.I., Alkhnbashi, O.S., Costa, F., Shah, S.A., Saunders, S.J. et al. (2015) An updated evolutionary classification of CRISPR-Cas systems. *Nat. Rev. Microbiol.* **13**, 722–736 https://doi.org/10.1038/nrmicro3569

2 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 https://doi.org/10.1016/j.molcel.2015.10.008

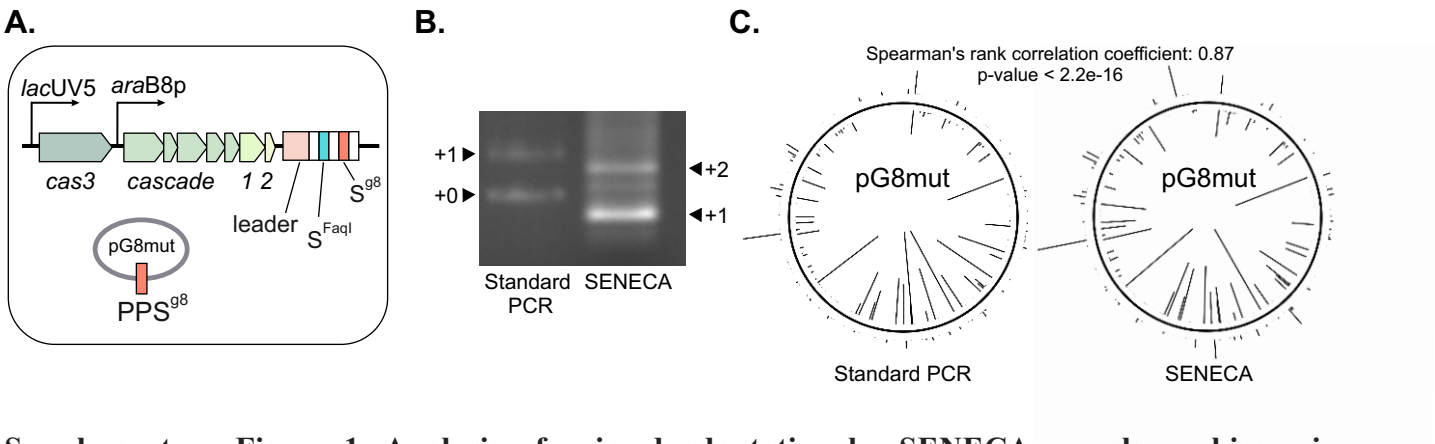
3 Makarova, K.S., Wolf, Y.I. and Koonin, E.V. (2018) Classification and nomenclature of CRISPR-Cas systems: where from here? *CRISPR J.* **1**, 325–336 https://doi.org/10.1089/crispr.2018.0033

4 Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S. et al. (2007) CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709–1712 https://doi.org/10.1126/science.1138140
5 Marraffini, L.A. and Sontheimer, E.J. (2008) CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science* **322**, 1843–1845 https://doi.org/10.1126/science.1165771
6 Mojica, F., Díez-Villaseñor, C., Soria, E. and Juez, G. (2000) Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. *Mol. Microbiol.* **36**, 244–246 https://doi.org/10.1046/j.1365-2958.2000.01838.x
7 Jansen, R., van Embden, J.D.A., Gaastra, W. and Schouls, L.M. (2002) Identification of genes that are associated with DNA repeats in prokaryotes. *Mol. Microbiol.* **43**, 1565–1575 https://doi.org/10.1046/j.1365-2958.2002.02839.x
8 Pourcel, C., Touchon, M., Villeriot, N., Vernadet, J.-P., Couvin, D., Toffano-Nioche, C. et al. (2019) CRISPRCasdb a successor of CRISPRdb containing CRISPR arrays and cas genes from complete genome sequences, and tools to download and query lists of repeats and spacers. *Nucleic Acids Res.* **48**, D535–D544 https://doi.org/10.1093/nar/gkz915
9 Shmakov, S.A., Sitnik, V., Makarova, K.S., Wolf, Y.I., Severinov, K.V. and Koonin, E.V. (2017) The CRISPR spacer space is dominated by sequences from species-specific mobilomes. *MBio* **8**, e01397-17 https://doi.org/10.1128/mBio.01397-17
10 Tang, T.-H., Bachellerie, J.-P., Rozhdestvensky, T., Bortolin, M.-L., Huber, H., Drungowski, M. et al. (2002) Identification of 86 candidates for small non-messenger RNAs from the archaeon *Archaeoglobus fulgidus*. *Proc. Natl Acad. Sci. U.S.A.* **99**, 7536–7541 https://doi.org/10.1073/pnas.112047299
11 Tang, T.-H., Polacek, N., Zywicki, M., Huber, H., Brugger, K., Garrett, R. et al. (2005) Identification of novel non-coding RNAs as potential antisense regulators in the archaeon *Sulfolobus solfataricus*. *Mol. Microbiol.* **55**, 469–481 https://doi.org/10.1111/j.1365-2958.2004.04428.x
12 Lillestøl, R.K., Redder, P., Garrett, R.A. and Brügger, K. (2006) A putative viral defence mechanism in archaeal cells. *Archaea* **2**, 59–72 https://doi.org/10.1155/2006/542818
13 Brouns, S.J.J., Jore, M.M., Lundgren, M., Westra, E.R., Slijkhuis, R.J.H., Snijders, A.P.L. et al. (2008) Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321**, 960–964 https://doi.org/10.1126/science.1159689
14 Hale, C., Kleppe, K., Terns, R.M. and Terns, M.P. (2008) Prokaryotic silencing (psi)RNAs in *Pyrococcus furiosus*. *RNA* **14**, 2572–2579 https://doi.org/10.1261/rna.1246808
15 Carte, J., Wang, R., Li, H., Terns, R.M. and Terns, M.P. (2008) Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. *Genes. Dev.* **22**, 3489–3496 https://doi.org/10.1101/gad.1742908
16 Lillestøl, R.K., Shah, S.A., Brügger, K., Redder, P., Phan, H., Christiansen, J. et al. (2009) CRISPR families of the crenarchaeal genus Sulfolobus: bidirectional transcription and dynamic properties. *Mol. Microbiol.* **72**, 259–272 https://doi.org/10.1111/j.1365-2958.2009.06641.x
17 Pul, U., Wurm, R., Arslan, Z., Geissen, R., Hofmann, N. and Wagner, R. (2010) Identification and characterization of *E. coli* CRISPR-Cas promoters and their silencing by H-NS. *Mol. Microbiol.* **75**, 1495–1512 https://doi.org/10.1111/j.1365-2958.2010.07073.x
18 Hale, C.R., Zhao, P., Olson, S., Duff, M.O., Graveley, B.R., Wells, L. et al. (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
19 Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A. and Charpentier, E. (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 https://doi.org/10.1126/science.1225829
20 Gasiunas, G., Barrangou, R., Horvath, P. and Siksnys, V. (2012) Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proc. Natl. Acad. Sci. U.S.A.* **109**, E2579–E2586 https://doi.org/10.1073/pnas.1208507109
21 Deveau, H., Barrangou, R., Garneau, J.E., Labonté, J., Fremaux, C., Boyaval, P. et al. (2008) Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. *J. Bacteriol.* **190**, 1390–1400 https://doi.org/10.1128/JB.01412-07
22 Horvath, P., Romero, D.A., Coûté-Monvoisin, A.-C., Richards, M., Deveau, H., Moineau, S. et al. (2008) Diversity, activity, and evolution of CRISPR loci in *Streptococcus thermophilus*. *J. Bacteriol.* **190**, 1401–1412 https://doi.org/10.1128/JB.01415-07
23 Yosef, I., Goren, M.G. and 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
24 Nuñez, J.K., Lee, A.S.Y., Engelman, A. and Doudna, J.A. (2015) Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity. *Nature* **519**, 193–198 https://doi.org/10.1038/nature14237
25 Xiao, Y., Ng, S., Nam, K.H. and Ke, A. (2017) How type II CRISPR-Cas establish immunity through Cas1-Cas2-mediated spacer integration. *Nature* **550**, 137–141 https://doi.org/10.1038/nature24020
26 Wei, Y., Terns, R.M. and Terns, M.P. (2015) Cas9 function and host genome sampling in Type II-A CRISPR-Cas adaptation. *Genes Dev.* **29**, 356–361 https://doi.org/10.1101/gad.257550.114
27 Stern, A., Keren, L., Wurtzel, O., Amitai, G. and Sorek, R. (2010) Self-targeting by CRISPR: gene regulation or autoimmunity? *Trends Genet.* **26**, 335–340 https://doi.org/10.1016/j.tig.2010.05.008
28 Edgar, R. and Qimron, U. (2010) The *Escherichia coli* CRISPR system protects from λ lysogenization, lysogens, and prophage induction. *J. Bacteriol.* **192**, 6291–6294 https://doi.org/10.1128/JB.00644-10
29 Vercoe, R.B., Chang, J.T., Dy, R.L., Taylor, C., Gristwood, T., Clulow, J.S. et al. (2013) Cytotoxic chromosomal targeting by CRISPR/Cas systems can reshape bacterial genomes and expel or remodel pathogenicity islands. *PLoS Genet.* **9**, e1003454 https://doi.org/10.1371/journal.pgen.1003454
30 Swarts, D.C., Mosterd, C., van Passel, M.W.J. and Brouns, S.J.J. (2012) CRISPR interference directs strand specific spacer acquisition. *PLoS ONE* **7**, e35888 https://doi.org/10.1371/journal.pone.0035888
31 Datsenko, K.A., Pougach, K., Tikhonov, A., Wanner, B.L., Severinov, K. and Semenova, E. (2012) Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. *Nat. Commun.* **3**, 945 https://doi.org/10.1038/ncomms1937
32 van der Ploeg, J.R. (2009) Analysis of CRISPR in *Streptococcus mutans* suggests frequent occurrence of acquired immunity against infection by M102-like bacteriophages. *Microbiology* **155**, 1966–1976 https://doi.org/10.1099/mic.0.027508-0
33 Garneau, J.E., Dupuis, M.-È., Villion, M., Romero, D.A., Barrangou, R., Boyaval, P. et al. (2010) The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468**, 67–71 https://doi.org/10.1038/nature09523
34 Levy, A., Goren, M.G., Yosef, I., Auster, O., Manor, M., Amitai, G. et al. (2015) CRISPR adaptation biases explain preference for acquisition of foreign DNA. *Nature* **520**, 505–510 https://doi.org/10.1038/nature14302

35 Shiimori, M., Garrett, S.C., Chambers, D.P., Glover, III, C.V.C., Graveley, B.R. and Terns, M.P. (2017) Role of free DNA ends and protospacer adjacent motifs for CRISPR DNA uptake in *Pyrococcus furiosus*. *Nucleic Acids Res.* **45**, 11281–11294 https://doi.org/10.1093/nar/gkx839
36 Díez-Villaseñor, C., Guzmán, N.M., Almendros, C., García-Martínez, J. and Mojica, F.J.M. (2013) CRISPR-spacer integration reporter plasmids reveal distinct genuine acquisition specificities among CRISPR-Cas I-E variants of *Escherichia coli*. *RNA Biol.* **10**, 792–802 https://doi.org/10.4161/rna.24023
37 Amlinger, L., Hoekzema, M., Wagner, E.G.H., Koskiniemi, S. and Lundgren, M. (2017) Fluorescent CRISPR adaptation reporter for rapid quantification of spacer acquisition. *Sci. Rep.* **7**, 10392 https://doi.org/10.1038/s41598-017-10876-z
38 Paez-Espino, D., Morovic, W., Sun, C.L., Thomas, B.C., Ueda, K.-I., Stahl, B. et al. (2013) Strong bias in the bacterial CRISPR elements that confer immunity to phage. *Nat. Commun.* **4**, 1430 https://doi.org/10.1038/ncomms2440
39 Savitskaya, E., Semenova, E., Dedkov, V., Metlitskaya, A. and Severinov, K. (2013) High-throughput analysis of type I-E CRISPR/Cas spacer acquisition in *E. coli*. *RNA Biol.* **10**, 716–725 https://doi.org/10.4161/rna.24325
40 Yosef, I., Shitrit, D., Goren, M.G., Burstein, D., Pupko, T. and Qimron, U. (2013) DNA motifs determining the efficiency of adaptation into the *Escherichia coli* CRISPR array. *Proc. Natl Acad. Sci. U.S.A.* **110**, 14396–41401 https://doi.org/10.1073/pnas.1300108110
41 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 https://doi.org/10.1093/nar/gku226
42 Heler, R., Samai, P., Modell, J.W., Weiner, C., Goldberg, G.W., Bikard, D. et al. (2015) Cas9 specifies functional viral targets during CRISPR-Cas adaptation. *Nature* **519**, 199–202 https://doi.org/10.1038/nature14245
43 Vorontsova, D., Datsenko, K.A., Medvedeva, S., Bondy-Denomy, J., Savitskaya, E.E., Pougach, K. et al. (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
44 Silas, S., Mohr, G., Sidote, D.J., Markham, L.M., Sanchez-Amat, A., Bhaya, D. et al. (2016) Direct CRISPR spacer acquisition from RNA by a natural reverse transcriptase-Cas1 fusion protein. *Science* **351**, aad4234 https://doi.org/10.1126/science.aad4234
45 Staals, R.H.J., Jackson, S.A., Biswas, A., Brouns, S.J.J., Brown, C.M. and Fineran, P.C. (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
46 Heler, R., Wright, A.V., Vucelja, M., Bikard, D., Doudna, J.A. and Marraffini, L.A. (2017) Mutations in Cas9 enhance the rate of acquisition of viral spacer sequences during the CRISPR-Cas immune response. *Mol. Cell* **65**, 168–175 https://doi.org/10.1016/j.molcel.2016.11.031
47 Li, M., Gong, L., Zhao, D., Zhou, J. and 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
48 Rao, C., Chin, D. and Ensminger, A.W. (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
49 Musharova, O., Vyhovskyi, D., Medvedeva, S., Guzina, J., Zhitnyuk, Y., Djordjevic, M. et al. (2018) Avoidance of trinucleotide corresponding to consensus protospacer adjacent motif controls the efficiency of prespacer selection during primed adaptation. *MBio* **9**, 02169-18 https://doi.org/10.1128/mBio.02169-18
50 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 https://doi.org/10.1016/j.celrep.2018.02.103
51 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**, 814–824 https://doi.org/10.1016/j.molcel.2018.05.002
52 Almendros, C., Nobrega, F.L., McKenzie, R.E. and Brouns, S.J.J. (2019) Cas4-Cas1 fusions drive efficient PAM selection and control CRISPR adaptation. *Nucleic Acids Res.* **47**, 5223–5230 https://doi.org/10.1093/nar/gkz217
53 Paez-Espino, D., Sharon, I., Morovic, W., Stahl, B., Thomas, B.C., Barrangou, R. et al. (2015) CRISPR immunity drives rapid phage genome evolution in *Streptococcus thermophilus*. *MBio* **6**, 00262-15 https://doi.org/10.1128/mBio.00262-15
54 Nivala, J., Shipman, S.L. and Church, G.M. (2018) Spontaneous CRISPR loci generation in vivo by non-canonical spacer integration. *Nat. Microbiol.* **3**, 310–318 https://doi.org/10.1038/s41564-017-0097-z
55 Jackson, S.A., Birkholz, N., Malone, L.M. and Fineran, P.C. (2019) Imprecise spacer acquisition generates CRISPR-Cas immune diversity through primed adaptation. *Cell Host Microbe* **25**, 250–260 https://doi.org/10.1016/j.chom.2018.12.014
56 Krivoy, A., Rutkauskas, M., Kuznedelov, K., Musharova, O., Rouillon, C., Severinov, K. et al. (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
57 Shipman, S.L., Nivala, J., Macklis, J.D. and Church, G.M. (2016) Molecular recordings by directed CRISPR spacer acquisition. *Science* **353**, aaf1175 https://doi.org/10.1126/science.aaf1175
58 McKenzie, R.E., Almendros, C., Vink, J.N.A. and Brouns, S.J.J. (2019) Using CAPTURE to detect spacer acquisition in native CRISPR arrays. *Nat. Protoc.* **14**, 976–990 https://doi.org/10.1038/s41596-018-0123-5
59 Schmidt, F., Cherepkova, M.Y. and Platt, R.J. (2018) Transcriptional recording by CRISPR spacer acquisition from RNA. *Nature* **562**, 380–385 https://doi.org/10.1038/s41586-018-0569-1
60 Semenova, E., Kuznedelov, K., Datsenko, K.A., Boudry, P.M., Savitskaya, E.E., Medvedeva, S. et al. (2015) The Cas6e ribonuclease is not required for interference and adaptation by the *E. coli* type I-E CRISPR-Cas system. *Nucleic Acids Res.* **43**, 6049–6061 https://doi.org/10.1093/nar/gkv546
61 Semenova, E., Savitskaya, E., Musharova, O., Strotskaya, A., Vorontsova, D., Datsenko, K.A. et al. (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
62 Kuznedelov, K., Mekler, V., Lemak, S., Tokmina-Lukaszewska, M., Datsenko, K.A., Jain, I. et al. (2016) Altered stoichiometry *Escherichia coli* cascade complexes with shortened CRISPR RNA spacers are capable of interference and primed adaptation. *Nucleic Acids Res.* **44**, 10849–10861 https://doi.org/10.1093/nar/gkw914
63 Musharova, O., Klimuk, E., Datsenko, K.A., Metlitskaya, A., Logacheva, M., Semenova, E. et al. (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
64 Musharova, O., Sitnik, V., Vlot, M., Savitskaya, E., Datsenko, K.A., Krivoy, A. et al. (2019) Systematic analysis of Type I-E *Escherichia coli* CRISPR-Cas PAM sequences ability to promote interference and primed adaptation. *Mol. Microbiol.* **111**, 1558–1570 https://doi.org/10.1111/mmi.14237

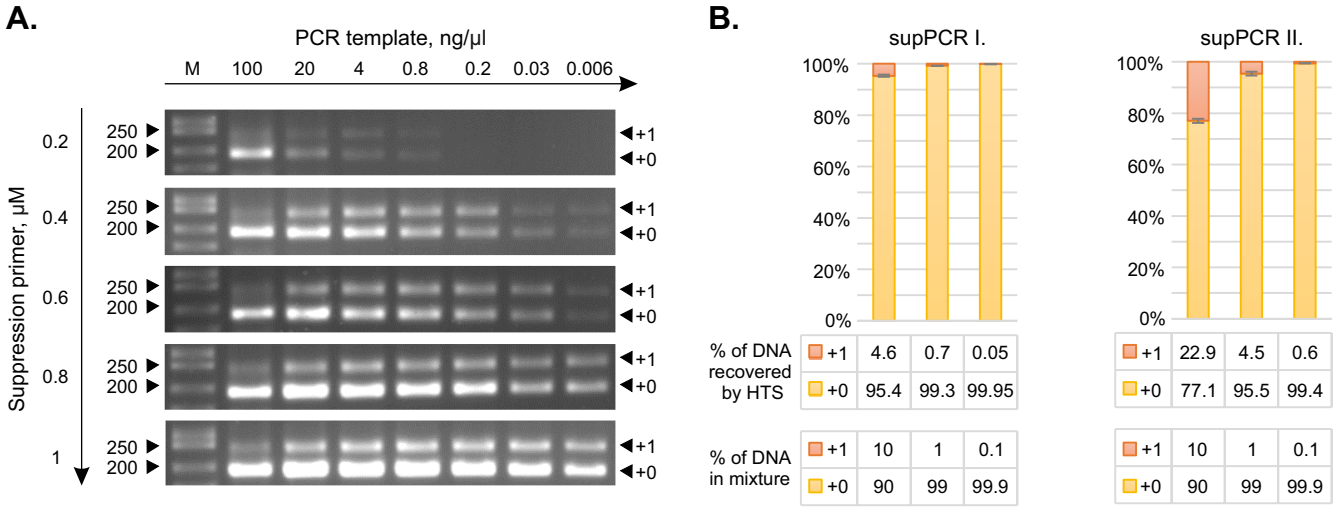
65 Strotskaya, A., Semenova, E., Savitskaya, E. and Severinov, K. (2015) Rapid multiplex creation of *Escherichia coli* strains capable of interfering with phage infection through CRISPR. *Methods Mol. Biol.* **1311**, 147–159 https://doi.org/10.1007/978-1-4939-2687-9_9
66 Siebert, P.D., Chenchik, A., Kellogg, D.E., Lukyanov, K.A. and Lukyanov, S.A. (1995) An improved PCR method for walking in uncloned genomic DNA. *Nucleic Acids Res.* **23**, 1087–1088 https://doi.org/10.1093/nar/23.6.1087
67 Diatchenko, L., Lau, Y.F., Campbell, A.P., Chenchik, A., Moqadam, F., Huang, B. et al. (1996) Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries. *Proc. Natl Acad. Sci. U.S.A.* **93**, 6025–6030 https://doi.org/10.1073/pnas.93.12.6025
68 Shagin, D.A., Lukyanov, K.A., Vagner, L.L. and Matz, M.V. (1999) Regulation of average length of complex PCR product. *Nucleic Acids Res.* **27**, e23 https://doi.org/10.1093/nar/27.18.e23
69 Ivančić-Baće, I., Cass, S.D., Wearne, S.J. and Bolt, E.L. (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
70 Fagerlund, R.D., Wilkinson, M.E., Klykov, O., Barendregt, A., Pearce, F.G., Kieper, S.N. et al. (2017) Spacer capture and integration by a type I-F Cas1-Cas2-3 CRISPR adaptation complex. *Proc. Natl Acad. Sci. U.S.A.* **114**, E5122–E5128 https://doi.org/10.1073/pnas.1618421114
71 Wright, A.V., Liu, J.-J., Knott, G.J., Doxzen, K.W., Nogales, E. and Doudna, J.A. (2017) Structures of the CRISPR genome integration complex. *Science* **357**, 1113–1118 https://doi.org/10.1126/science.aao0679
72 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**, 48–59 https://doi.org/10.1016/j.molcel.2018.03.003
73 Drabavicius, G., Sinkunas, T., Silanskas, A., Gasiunas, G., Venclovas, Č. and Siksnys, V. (2018) Dnaq exonuclease-like domain of Cas2 promotes spacer integration in a type I-E CRISPR-Cas system. *EMBO Rep.* **19**, e45543 https://doi.org/10.15252/embr.201745543
74 Radovčić, M., Killelea, T., Savitskaya, E., Wettstein, L., Bolt, E.L. and Ivančić-Baće, I. (2018) CRISPR–cas adaptation in *Escherichia coli* requires RecBCD helicase but not nuclease activity, is independent of homologous recombination, and is antagonized by 5′ ssDNA exonucleases. *Nucleic Acids Res.* **46**, 10173–10183 https://doi.org/10.1093/nar/gky799
75 Wilkinson, M., Drabavicius, G., Silanskas, A., Gasiunas, G., Siksnys, V. and Wigley, D.B. (2019) Structure of the DNA-bound spacer capture complex of a type II CRISPR-Cas system. *Mol. Cell* **75**, 90–101 https://doi.org/10.1016/j.molcel.2019.04.020
76 Wang, J., Li, J., Zhao, H., Sheng, G., Wang, M., Yin, M. et al. (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
77 Rollie, C., Graham, S., Rouillon, C. and White, M.F. (2018) Prespacer processing and specific integration in a type I-A CRISPR system. *Nucleic Acids Res.* **46**, 1007–1020 https://doi.org/10.1093/nar/gkx1232
78 Moch, C., Fromant, M., Blanquet, S. and Plateau, P. (2017) DNA binding specificities of *Escherichia coli* Cas1-Cas2 integrase drive its recruitment at the CRISPR locus. *Nucleic Acids Res.* **45**, 2714–2723 https://doi.org/10.1093/nar/gkw1309
79 Shiriaeva, A.A., Savitskaya, E., Datsenko, K.A., Vvedenskaya, I.O., Fedorova, I., Morozova, N. et al. (2019) Detection of spacer precursors formed in vivo during primed CRISPR adaptation. *Nat. Commun.* **10**, 4603 https://doi.org/10.1038/s41467-019-12417-w
80 Zhelkovsky, A.M. and McReynolds, L.A. (2012) Structure-function analysis of *Methanobacterium thermoautotrophicum* RNA ligase - engineering a thermostable ATP independent enzyme. *BMC Mol. Biol.* **13**, 24 https://doi.org/10.1186/1471-2199-13-24

Supplementary Data



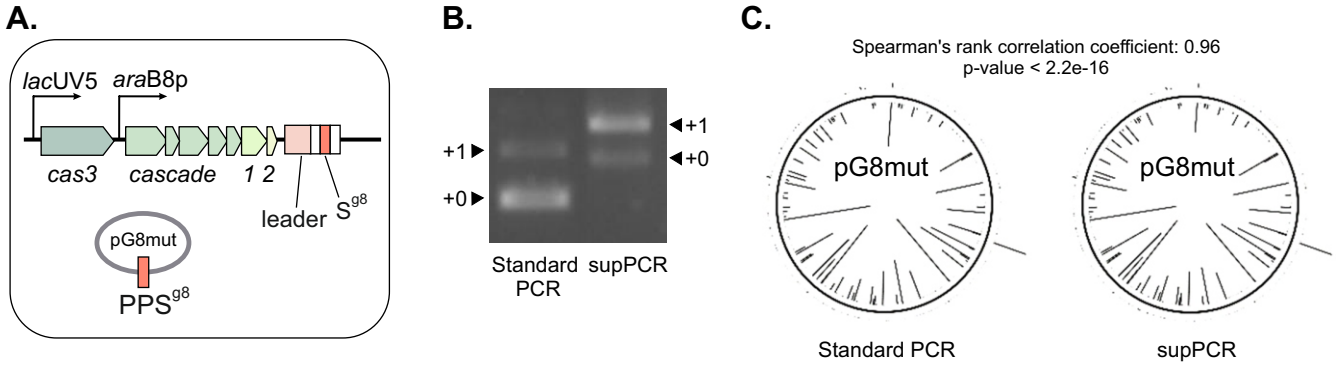
Supplementary Figure 1. Analysis of primed adaptation by SENECA reveals no biases in amplification of individual spacers compared with standard PCR.

A. Schematic representation of an *E. coli* strain KD263_FaqI cell used to monitor primed adaptation by SENECA. The *lac*UV5 and *ara*B8p promoters direct inducible transcription of *cas* genes (shown as block arrows). The engineered CRISPR array contains the leader (light peach rectangle), three CRISPR repeats (white rectangles), spacer S^{FaqI} with an FaqI recognition site (turquoise rectangle), and spacer S^{g8} (dark peach rectangle) matching priming protospacer PPS^{g8} (dark peach rectangle) of plasmid pG8mut. **B.** Agarose gel electrophoresis analysis of primed adaptation in induced KD263_FaqI culture by standard PCR and SENECA. Amplicons corresponding to unexpanded and expanded arrays revealed by standard PCR (left lane) are labeled +0 and +1, respectively. Amplicons corresponding to arrays expanded by one or two spacers amplified by SENECA (right lane) are labeled +1 and +2. **C.** HTS of acquired spacers revealed by standard PCR and SENECA. +1 amplicons from panel **B** were subjected to HTS. Spacers were extracted from reads and mapped to the pG8mut plasmid. Each bar represents an individual protospacer. Bars oriented inside or outside the plasmid circles represent protospacers located on different DNA strands. The bar heights are proportional to the number of HTS reads containing corresponding spacers. Spearman's rank correlation coefficient between the numbers of spacers obtained by standard protocol and SENECA is shown above.



Supplementary Figure 2. Optimizing supPCR conditions to obtain increased sensitivity of CRISPR adaptation detection.

A. Genomic DNA purified from an *E. coli* strain with one additional spacer (+1) was mixed with DNA from a parental strain with unexpanded (+0) array in the ratio of 10% to 90% and used as a template for the first stage of supPCR. The products were diluted to concentrations indicated at the top and used as a template for the second stage supPCR in the presence of a suppression primer at concentrations indicated on the left. Products of the second stage supPCR corresponding to unexpanded (+0) and expanded (+1) arrays separated by agarose gel electrophoresis are shown. M, molecular-weight size marker. **B.** Genomic DNA purified from an *E. coli* strain with one additional spacer (+1) was mixed with DNA from a strain with unexpanded (+0) array in proportions indicated at the bottom. The mixtures were used as a template for the first and second stages of supPCR ("supPCR I" and "supPCR II") followed by high-throughput sequencing. Percentages of recovered reads corresponding to unexpanded and expanded CRISPR arrays are indicated in bar plots and tables immediately below.



Supplementary Figure 3. Analysis of primed adaptation by supPCR reveals no biases in amplification of individual spacers compared with standard PCR.

A. Schematic representation of an *E. coli* strain KD263 cell used to monitor primed adaptation by supPCR. The *lac*UV5 and *ara*B8p promoters direct inducible transcription of *cas* genes (shown as block arrows). The engineered CRISPR array contains the leader (light peach rectangle), two CRISPR repeats (white rectangles), and spacer S^{g8} (dark peach rectangle) matching priming protospacer PPS^{g8} (dark peach rectangle) of plasmid pG8mut. **B.** Agarose gel electrophoresis analysis of primed adaptation in induced KD263 culture by standard PCR and supPCR. Amplicons corresponding to unexpanded and expanded arrays are labeled +0 and +1, respectively. **C.** HTS of acquired spacers revealed by standard PCR and supPCR. The +1 amplicons from panel **B** were subjected to HTS. See Supplementary Figure 1 panel C legend for details.

Supplementary Methods

Bacterial strains and plasmids

The *E. coli* strains used in this study are listed in Supplementary Table 1. The KD263_FaqI strain is a derivative of KD263 strain [1] with an additional spacer S^{FaqI} containing FaqI recognition site (Fig. 3A, Supplementary Fig. 1A). To obtain KD263_FaqI, primed adaptation was induced in KD263 transformed with a pG8mut_FaqI plasmid described below (Fig. 3A). After induction of primed adaptation, individual colonies that lost Amp-resistance were selected and tested by PCR with primers Ec_LDR_F and M13_G8 (Supplementary Table 2). Amplicons corresponding to arrays expanded by a single spacer-repeat unit were subjected to Sanger sequencing. Screening of ~100 clones allowed to identify several colonies with the S^{FaqI} insertion.

Plasmid pG8mut_FaqI is a derivative of pG8mut [2] with a modified hot protospacer HS1 (GCTTTCCCTATAGTGAGTCGTATTAGAGCTTGG) [3]. Modified HS1, PS^{FaqI} (<u>GTCCC</u>CCCTATAGTGAGTCGTATTAGAGCTTGG), contains the FaqI recognition site (underlined). To construct pG8mut_FaqI, plasmid pG8mut was amplified with primers Faq_HS1_plF and Faq_HS1_plR (Supplementary Table 2) using iProof High-Fidelity DNA Polymerase (Bio-Rad) according to the manufacturer's instructions. The resulting PCR product was purified and treated with 0.05 U/µl DpnI (Thermo Fisher Scientific). After 5'-phosphorylation with T4 polynucleotide kinase (Thermo Fisher Scientific) DNA was precipitated, diluted in 5 µl of sterile Milli-Q water and circularized with Gibson Assembly Master Mix (NEB) following manufacturer's instructions. The ligation mixture was transformed into competent *E. coli* DH5α cells and plasmids with desired sequence were selected by Sanger sequencing.

Primed adaptation assay

The same protocol was used to induce primed adaptation in KD263 transformed with pG8mut_FaqI to obtain the KD263_FaqI strain, in KD263_FaqI transformed with pG8mut to monitor spacer acquisition by SENECA, or in KD263 transformed with pG8mut to monitor spacer acquisition by supPCR.

Overnight cultures obtained from individual transformant colonies were grown in liquid LB supplemented with 100 µg/ml ampicillin, diluted 100-fold in LB without antibiotics and grown at 37 °C until culture OD_{600} reached 0.3-0.4. CRISPR interference/primed adaptation was induced by the addition of 1 mM IPTG and 1 mM L-(+)-arabinose as described [4]. After 6

hours growth cells were either plated to select the KD263_FaqI strain (above) or processed for total genomic DNA purification by GeneJET Genomic DNA Purification Kit (Thermo Fisher Scientific) and subsequent analysis of spacer acquisition by standard PCR, SENECA, or supPCR.

Detection of spacer acquisition by standard PCR

CRISPR arrays were amplified by PCR in a 25-µl reaction containing 30 ng of total genomic DNA, 2.5 U of Taq DNA Polymerase (αFerment), 1× Taq buffer (αFerment), 0.2 mM dNTPs (Thermo Fisher Scientific), 0.5 µM primer annealing to the leader sequence (Ec_LDR_F, Supplementary Table 2), and 0.5 µM primer annealing to the preexisting spacer (M13_G8 for KD263 strain or HS1_qPCR_adaptation for KD263_FaqI strain, Supplementary Table 2). The PCR cycling conditions were as follows: 3 min at 95 °C followed by 32 cycles of 15 sec at 95 °C, 30 sec at 55 °C, and 30 sec at 72 °C with a final extension of 5 min at 72 °C.

Detection of spacer acquisition by SENECA

For analysis of adaptation by SENECA, CRISPR arrays were first amplified by PCR in a 25-µl reaction containing 30 ng of total genomic DNA, 2.5 U of Taq DNA Polymerase (αFerment), 1× Taq buffer (αFerment), 0.2 mM dNTPs (Thermo Fisher Scientific), 0.5 µM Ec_LDR_F primer (anneals to the leader sequence), and 0.5 µM HS1_qPCR_adaptation primer annealing to S^{FaqI} spacer (Supplementary Table 2). The PCR cycling conditions were as follows: 3 min at 95 °C followed by 35 cycles of 15 sec at 95 °C, 30 sec at 50 °C, and 30 sec at 72 °C with a final extension of 5 min at 72 °C.

The PCR product was purified with the GeneJET PCR Purification Kit (Thermo Fisher Scientific), treated with FaqI restriction endonuclease and ligated to Illumina adapter according to [5]. Oligonucleotides FaqI_adapter_F and FaqI_adapter_R (Supplementary Table 2) were annealed by heating at 95 ˚C for 5 min and cooling the mixture to 25 ˚C to form a double-stranded Illumina adapter (FaqI_adapter) with an overhang mimicking the sticky end generated by FaqI. The cleavage of CRISPR array amplicons by FaqI and simultaneous ligation of the FaqI_adapter was performed in a 50-µl reaction containing 20 – 200 ng of DNA, 0.01 µM FaqI_adapter, 1 mM ATP (NEB), 1 mM DTT (VWR Life Science AMRESCO), 0.2 U of T4 DNA Ligase (Thermo Fisher Scientific), 1 µL of FastDigest FaqI restriction endonuclease (Thermo Fisher Scientific), and 1× FastDigest buffer (Thermo Fisher Scientific). 99 cycles of 3 min at 37 °C and 3 min at 20 °C were performed. Reaction products were purified with the GeneJET PCR Purification Kit (Thermo Fisher Scientific), resuspended in 15 µl of MQ. 3 µl

aliquots (3ng/µl) were used for amplification of CRISPR arrays ligated to the FaqI_adapter in a 100-µl reaction with 2.5 U of Taq DNA Polymerase (αFerment), 1× Taq buffer (αFerment), 0.2 mM dNTPs (Thermo Fisher Scientific), 0.5 µM primer FaqI_adapter_PRIMER annealing to the Illumina adapter, and 0.5 µM primer CRISPR_array_for annealing to CRISPR repeat not cleavable by the FaqI (Figure 2D, oligonucleotide sequences are available in Supplementary Table 2). The PCR cycling conditions were as follows: 3 min at 95 °C followed by 35 cycles of 15 sec at 95 °C, 30 sec at 52 °C, and 30 sec at 72 °C with a final extension of 5 min at 72 °C.

Detection of spacer acquisition by supPCR

For analysis of adaptation by supPCR, two consecutive PCR reactions were performed. At the first stage, CRISPR arrays were amplified by standard PCR in a 25-µl reaction containing 20 ng of total genomic DNA, 2.5 U of Taq DNA Polymerase (αFerment), 1× Taq buffer (αFerment), 0.2 mM dNTPs (Thermo Fisher Scientific), and 0.5 µM primers containing the ITR sequence on 5' ends and annealing to the leader sequence (KD263_Ist_F, Supplementary Table 2) and preexisting spacer (KD263_Ist_R, Supplementary Table 2). The PCR cycling conditions were as follows: 3 min at 95 °C followed by 20 cycles of 15 sec at 95 °C, 30 sec at 55 °C, and 30 sec at 72 °C with a final extension of 5 min at 72 °C.

The amplified DNA was diluted 100-fold and 1 µl of the mixture was used as a template for the second stage PCR reaction performed in a total volume of 20 µl and containing 1× Q5 high-fidelity polymerase master mix (NEB) and 0.4 µM suppression primer (ITR_primer, Supplementary Table 2). The PCR cycling conditions were as follows: 30 sec at 98 °C followed by 40 cycles of 30 sec at 98 °C, 30 sec at 57.3 °C, and 30 sec at 72 °C with a final extension of 2 min at 72 °C.

High-throughput sequencing of CRISPR arrays

HTS libraries were generated using NEBNext Ultra II DNA Library Prep Kit (NEB) and sequenced at the MiSeq Illumina platform in a pair-end 150-bp read mode. Raw sequencing data were analyzed using ShortRead and BioStrings packages [6,7]. Sequencing reads were filtered for quality scores of ≥20 and reads containing two repeats (with up to two mismatches) were selected. Reads that contained 33-bp sequences between two parts of CRISPR repeats (AGCGGGGATAAACCG and GTGTTCCCCGCGCC) were next selected. The 33-bp segments were considered as spacers. For reads corresponding to multiple spacer acquisition events, only spacers that were acquired first were selected for further analysis. Spacers were

next mapped onto the pG8mut plasmid with no mismatches allowed. Read mapping and spacer statistics analysis was performed with R [1]. Graphical representation was carried out using the EasyVisio tool developed by Ekaterina Rubtsova.

Supplementary Table 1. Strains used in this study

Name	Description	Source
KD263	K-12 F^+, *lacUV5-cas3 araBp8-cse1*, CRISPR I: repeat-S^{g8}-repeat, CRISPR II deleted. S^{g8} (CTGTCTTTCGCTGCTGAGGGTGACGATCCCGC) targets g8 gene of M13 phage.	Shmakov *et al.*[1]
KD263_FaqI	A derivative of KD263 with an additional spacer S^{FaqI} in CRISPR I: repeat-S^{FaqI}-repeat-S^{g8}-repeat. S^{FaqI} (<u>GTCCC</u>CCCTATAGTGAGTCGTATTAGAGCTTGG) contains the FaqI recognition site (underlined).	This study

Supplementary Table 2. Oligonucleotides used in this study

Name	Sequence (5' to 3')	Description
Faq_HS1_plF	CCTGCAGGCATGCAAGTCCCCCCTATAGTGAGTCGTATTA	pG8mut_FaqI construction
Faq_HS1_plR	TAATACGACTCACTATAGGGGGGACTTGCATGCCTGCAGG	
Ec_LDR_F	AAGGTTGGTGGGTTGTTTTTATGG	Anneals to the leader sequence (standard PCR; SENECA 1st stage PCR)
M13_G8	GGATCGTCACCCTCAGCAGCG	Anneals to S^{g8} spacer (standard PCR)
HS1_qPCR_adaptation	CCAAGCTCTAATACGAC	Anneals to S^{FaqI} spacer (standard PCR; SENECA 1st stage PCR)
Faq_adapter_F	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	Two strands of FaqI_adapter for SENECA
Faq_adapter_R	AGCGAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC	
Faq_adapter_PRIMER	GTGACTGGAGTTCAGACG	Anneals to FaqI_adapter (SENECA 2nd stage PCR)
CRISPR_array_for	AGCGGGGATAAACCG	Anneals to CRISPR repeat (SENECA 2nd stage PCR)
KD263_Ist_F	GTAATACGACTCACTATAGGGCACGGTTGCGGAAATGTTACATTAAGGTTGG	The underlined region anneals to the leader sequence (supPCR 1st stage)
KD263_Ist_R	GTAATACGACTCACTATAGGGCACGGTTGCCCCTCAGCAGCGAAAGACAG	The underlined region anneals to S^{g8} spacer (supPCR 1st stage)
ITR_primer	GTAATACGACTCACTATAGGGC	Suppression primer (supPCR 2nd stage)

Supplementary References

1. 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(9)**, 5907–5916
2. Datsenko, K.A., Pougach, K., Tikhonov, A., Wanner, B.L., Severinov, K. and Semenova, E. (2012) Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. *Nat. Commun.* **3**, 945
3. Musharova, O., Sitnik, V., Vlot, M., Savitskaya, E., Datsenko, K.A., Krivoy, A. et al. (2019) Systematic analysis of Type I-E Escherichia coli CRISPR-Cas PAM sequences ability to promote interference and primed adaptation. *Mol. Microbiol.* **111(6)**, 1558–1570
4. Semenova, E., Savitskaya, E., Musharova, O., Strotskaya, A., Vorontsova, D., Datsenko, K.A. et al. (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(27)**, 7626–7631
5. Schmidt, F., Cherepkova, M.Y., Platt, R.J. (2018) Transcriptional recording by CRISPR spacer acquisition from RNA. *Nature.* **562(7727)**, 380–385
6. 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(19)**, 2607–2608
7. Pagès, H., Aboyoun, P., Gentleman, R., DebRoy, S. (2017) Biostrings: Efficient manipulation of biological strings. R package version 2.46.0.