Skolkovo Institute of Science and Technology

INTERFERENCE AND PRIMED ADAPTATION INTERMEDIATES IN TYPE I CRISPR-CAS SYSTEMS

Doctoral Thesis

by

ANNA SHIRIAEVA

DOCTORAL PROGRAM IN LIFE SCIENCES

Supervisor
Professor Konstantin Severinov

Moscow - 2020

I hereby declare that the work presented in this thesis was carried out by myself at Skolkovo Institute of Science and Technology, Moscow, except where due acknowledgement is made, and has not been submitted for any other degree.

Candidate (Anna Shiriaeva)

Supervisor (Prof. Konstantin Severinov)

Abstract

CRISPR-Cas systems are adaptive immunity systems of prokaryotes. They consist of CRISPR arrays and *cas* genes. CRISPR arrays store short sequences of foreign nucleic acids as 'spacers'. The order of spacers reflects the history of encounters with mobile genetic elements (MGEs). The process of spacer acquisition, CRISPR adaptation, underlies CRISPR-mediated adaptive immunity. Spacers target complementary sequences (protospacers) for destruction by CRISPR interference machinery. CRISPR interference is initiated on a target protospacer containing a short motif called PAM (Protospacer Adjacent Motif). Therefore, during adaptation new spacers should be selected from PAM-containing sequences to ensure efficient degradation of future targets. The PAM is removed from a spacer precursor (prespacer) prior to its integration into the CRISPR array. This mechanism allows cells to avoid the initiation of interference on 'self' CRISPR DNA.

The lengths of spacers and the PAM sequence are specific for a given CRISPR-Cas system. It is known that only double-stranded prespacers are integrated into CRISPR arrays. However, the lengths of prespacer strands, PAM positions within these precursors, and the mechanism of their generation are not known. The fragments generated during CRISPR interference are believed to serve as prespacers during a highly efficient mode of adaptation - primed adaptation. These fragments have not been characterized *in vivo* and it is not known if non-Cas nucleases participate in their production.

In this work we studied primed adaptation by the type I-E and type I-F CRISPR-Cas systems in *Escherichia coli*. Using FragSeq, a high-throughput sequencing approach developed as part of this project for analysis of short DNA fragments generated *in vivo*, we detected asymmetric double-stranded prespacers with short 3'-end overhangs on PAM-derived sides produced by both systems. The efficient generation of prespacers by the type I-E system during primed adaptation requires the presence of intact interference and adaptation modules. Furthermore, at least one of two non-Cas enzymes, a

5′→3′ exonuclease RecJ or a helicase RecBC involved in DNA repair, is required for prespacer production. We show that RecBC and RecJ are involved in trimming of prespacer 5’ ends. In addition, we assess the contribution of host DNA repair nucleases to CRISPR interference and characterize the products they generate under conditions of self-targeting of the *E. coli* genome by the type I-E CRISPR-Cas system.

Publications

1. **Shiriaeva, A.A.**, Savitskaya, E., Datsenko, K.A., Vvedenskaya, I.O., Fedorova, I., Morozova, N., Metlitskaya, A., Sabantsev, A., Nickels, B.E., Severinov, K., et al. (2019). Detection of spacer precursors formed *in vivo* during primed CRISPR adaptation. Nat Commun *10*, 4603.
2. Kurilovich, E., **Shiriaeva, A.**, Metlitskaya, A., Morozova, N., Ivancic-Bace, I., Severinov, K., and Savitskaya, E. (2019). Genome Maintenance Proteins Modulate Autoimmunity Mediated Primed Adaptation by the *Escherichia coli* Type I-E CRISPR-Cas System. Genes *10*, 872.
3. **Shiriaeva, A**., Fedorov, I., Vyhovskyi, D., and Severinov, K. (2020). Detection of CRISPR adaptation. Biochem. Soc. Trans. *48* (1), 257-269 (a review article with original experimental data).
4. Wiegand, T., Semenova, E., **Shiriaeva, A.**, Fedorov, I., Datsenko, K., Severinov, K., Wiedenheft, B. (2020). Reproducible Antigen Recognition by the Type I-F CRISPR-Cas System. The CRISPR Journal. *3* (5), 378-387.

Conferences

1. **Shiriaeva A.**, Savitskaya E., Semenova E., Datsenko K.A., Severinov K. RecBC influences CRISPR Adaptation in the Type I-E CRISPR-Cas System of *Escherichia coli*, Waksman Annual Retreat, Rutgers University, NJ, USA, September 12, 2017: **poster presentation**.
2. **Anna Shiriaeva**, Ekaterina Savitskaya, Kirill Datsenko, Irina Vvedenskaya, Bryce Nickels, Ekaterina Semenova, Konstantin Severinov *In vivo* detection of primed adaptation intermediates in *Escherichia coli*, CRISPR2018, June 20-23, 2018, Vilnius, Lithuania: **poster presentation**.
3. **Anna Shiriaeva** Detection of Spacer Precursors Formed *in vivo* During Primed Adaptation, Rutgers - Wageningen Graduate Student Symposium, June 13, 2019, New Brunswick, USA: **oral presentation**.
4. **Anna Shiriaeva**, Ekaterina Savitskaya, Elena Kurilovich, Anastasia Metlitskaya, Kirill Datsenko, Ekaterina Semenova, Konstantin Severinov Detection of RecBCD-generated DNA products in *Escherichia coli* cells undergoing CRISPR interference, 3rd International Conference on CRISPR Technologies, September 16-18, 2019, Würzburg, Germany: **poster presentation**.

Acknowledgements

I would like to thank my supervisor Professor Konstantin Severinov who inspired me to study CRISPR adaptation, outlined the goals of the presented work, but at the same time provided the freedom to test any other ideas that arose during the implementation of this project, supported collaborations, and participation in academic mobility. This work would have not been possible without the strategic view of Professor Severinov on how my project should develop and his scientific guidance. I am also grateful to Professor Severinov for introducing me to so many wonderful people who helped me with solving specific problems that arose during this project.

I would like to thank Skoltech for its excellent courses, the inspiring atmosphere of a modern rapidly developing university, for supporting long-term academic mobility trips, and participation in international conferences.

Thanks to Skoltech academic mobility programs, I was conducting research in two laboratories at Rutgers University in the USA, Prof. Severinov and Prof. Nickels laboratories. I would like to thank all members of these two amazing laboratories for creating a friendly and productive atmosphere. I would like especially to thank Dr. Ekaterina Semenova who constantly helped me from the first day of my appearance in the laboratory and beyond. The ideas of many experiments would not have been born without our daily discussions with Ekaterina on current experiments and future plans. Another person who I owe many thanks is Prof. Bryce Nickels. I am very grateful to him for allowing me to work in his laboratory, and for the many hours he spent on meticulously editing a draft of our joint paper, which has been greatly improved thanks to his help. I also want to thank Irina Vvedenskaya who is a member of Prof. Nickels laboratory and who taught me useful techniques of HTS library construction.

I was also performing a part of this work at Peter the Great St. Petersburg Polytechnic University in Saint Petersburg. I am very grateful to each member of the analytical center of nano- and biotechnologies of SPbPU and the head of the center Dr. Mikhail Khodorkovskii for creating a productive atmosphere and supporting my research.

When I just entered Skoltech I was working under the supervision of Ekaterina Savitskaya who taught me all the basic methods of studying CRISPR interference and adaptation including the analysis of high-throughput sequencing data and to whom I often turned for advice while performing this project. I am deeply devastated by the fact that Ekaterina is no longer with us and I cannot tell her how grateful I am.

Many other people contributed to this work. I would like to thank Elena Kurilovich, Anastasia Metlitskaya, Natalia Morozova, Iana Fedorova, and Kirill Datsenko who performed some experiments or constructed the necessary strains; Min Tu, Dibyendu Kumar, and Maria Logacheva who performed high-throughput sequencing; Sofia Medvedeva who helped with data analysis; and Konstantin Kuznedelov who suggested some of the experiments described in this work.

I would also like to thank the members of my Individual Doctoral Committee at Skoltech, Prof. Gelfand and Prof. Sergiev, and the Jury Members, Prof. Lukyanov, Prof. Chudakov, Prof. Bailey, Prof. Terns, and Prof. Kotelevtsev for carefully reading this manuscript, and for valuable suggestions for improving the text and presentation of the results.

I would like to thank my parents, Galina Shiriaeva and Alexander Shiriaev, and my dear friends Yuri Kireenko, Maria Sokolova, Iana Fedorova, Oksana Frangulantc, and Ksenia Kuznetsova for their constant moral support and for staying my friends despite my busy lifestyle. I also want to say a lot of thanks to Raga Mohamed, Jomana Mohamed, and Ishita Jain who helped me settle in a foreign country when I arrived in the United States and supported me.

Finally, I would like to thank Dr. Elena Stepchenkova, who was my supervisor while I was doing my bachelor's and master's studies and who I owe my general ability to think and write scientifically, plan and perform experiments and present the obtained results.

Table of Contents

List of Symbols, Abbreviations

5’ app – 5’-adenylpyrophosphoryl

A - adenine

ATP - adenosine triphosphate

BiFC assay - Bimolecular fluorescence complementation assay

bp – base pair

bp/s – base pairs per second

C – cytosine

cAMP - cyclic adenosine monophosphate

Cascade - CRISPR-associated complex for antiviral defense

CFU - colony forming unit

CRISPR- Clustered Regularly Interspaced Short Palindromic Repeats

CRP - cAMP receptor protein

crRNA - CRISPR RNA

Cryo-EM – cryo-electron microscopy

ddC - dideoxycytidine

DNA - deoxyribonucleic acid

dNTPs – deoxynucleotides

DSB - double-strand break

dsDNA - double-stranded DNA

dw – downstream

$Frag^{NT}$ - nontarget-strand-derived fragments

$Frag^{T}$ - target-strand-derived fragments

G – guanine

gDNA – genomic DNA

H-NS - histone-like nucleoid-structuring protein

IHF - integration host factor

IPTG - isopropyl β-D-1-thiogalactopyranoside

IQR - interquartile range

kbp - kilobase pair

K_D - dissociation constant

LB - Luria-Bertani broth

MGEs - mobile genetic elements

nm – nanometer

nt – nucleotide

NT-strand – nontarget strand

OD - optical density

PAM - protospacer adjacent motif

PBS - Phosphate-buffered saline

PCR - Polymerase Chain Reaction

PPS - priming protospacer

pre-crRNA - precursor CRISPR RNA

PS^{NT} – protospacers corresponding to Sp^{NT}

PS^{T} - protospacers corresponding to Sp^{T}

RAMP - Repeat-Associated Mysterious Protein

RNA – ribonucleic acid

SDS - sodium dodecyl sulfate

SEM – standard error of mean

SF1 - Superfamily 1

SF2 - Superfamily 2

Sp^{NT} – spacers whose non-transcribed 'top' strand originates from the NT-strand

Sp^{T} – spacers whose non-transcribed 'top' strand originates from the T-strand

SSB – ssDNA-binding protein

ssDNA – single-stranded DNA

T - thymine

TE buffer - Tris-EDTA buffer

T-strand – target strand

up - upstream

UV - ultraviolet

μm - micrometer

μM – micromolar (micromoles per litre)

List of Figures

Chapter 1. Literature Review

1.1 CRISPR-Cas systems as adaptive immunity systems of prokaryotes

CRISPR-Cas systems are adaptive immunity systems of bacteria and archaea acting against mobile genetic elements like bacteriophages and plasmids (Barrangou et al., 2007; Marraffini and Sontheimer, 2008). Functional CRISPR-Cas systems include one or several CRISPR arrays (Clustered Regularly Interspaced Short Palindromic Repeats) and CRISPR-associated *cas* genes (Figure 1) (Jansen et al., 2002; Mojica et al., 2000). CRISPR arrays consist of repeats separated by unique DNA sequences called **spacers** (Jansen et al., 2002; Mojica et al., 2000). An AT-rich 'leader' sequence is located upstream of the first CRISPR repeat (Jansen et al., 2002). Though the source of most spacers remains unclear, spacers with known origin match sequences of mobile genetic elements: bacteriophages and plasmids (Bolotin et al., 2005; Mojica et al., 2005; Pourcel et al., 2005; Shmakov et al., 2017). In 2007, Barrangou et al. experimentally demonstrated that spacers and *cas* genes protect *Streptococcus thermophilus* from phages containing **protospacers** – regions outside CRISPR arrays whose sequences are identical to the spacers (Barrangou et al., 2007). CRISPR-Cas systems have been found in 43% of complete prokaryotic genomes (85% in archaea and 40% in bacteria) (Makarova et al., 2020). All CRISPR-Cas systems are currently divided into two classes: class 1 includes types I, III, and IV; class 2 includes types II, V, and VI (Makarova et al., 2020). Each CRISPR-Cas type is characterized by the presence of a signature protein which is absent in other types (Makarova et al., 2015). CRISPR-Cas types are further divided into several subtypes (Makarova et al., 2020).

In general terms, CRISPR-Cas systems operate as follows. When bacteriophages inject their genomes into bacteria some cells incorporate pieces of viral DNA (**prespacers**) into the CRISPR array as new spacers (Figure 1) (Barrangou et al., 2007). Acquisition of new spacers from a DNA molecule that was not previously used by the same cell lineage as a source of spacers is termed **naïve adaptation**. Several Cas proteins participate in this process and constitute the module of adaptation: Cas1, Cas2, and, in some subtypes, Cas4, Csn2, and reverse transcriptases (Makarova et al., 2020).

Transcription of the CRISPR array yields a long **pre-crRNA** molecule containing multiple repeats and spacers (Brouns et al., 2008; Deltcheva et al., 2011; Hale et al., 2008). The transcript is further processed into short **crRNAs**, each containing a single spacer and partial repeat sequences (Figure 1) (Brouns et al., 2008; Deltcheva et al., 2011; Hale et al., 2008). crRNAs are bound by proteins of the CRISPR interference module forming an effector complex that searches for protospacers and either cleaves them or recruits other proteins degrading target nucleic acids (Gasiunas et al., 2012; Jinek et al., 2012; Jore et al., 2011; Kazlauskiene et al., 2016; Westra et al., 2012). The effector complexes of class 1 CRISPR-Cas systems are formed by multiple protein subunits while the effector complexes of class 2 are represented by large single proteins Cas9, Cas12, and Cas13, which are also the signature proteins of types II, V, and VI, respectively (Makarova et al., 2020). In different CRISPR-Cas types, effector complexes target either DNA (types I, II, V) or RNA (types III, VI). In type III systems, the recognition of complementary RNA is followed by in-trans cleavage of non-complementary DNA (Estrella et al., 2016; Goldberg et al., 2014; Kazlauskiene et al., 2016; Samai et al., 2015).

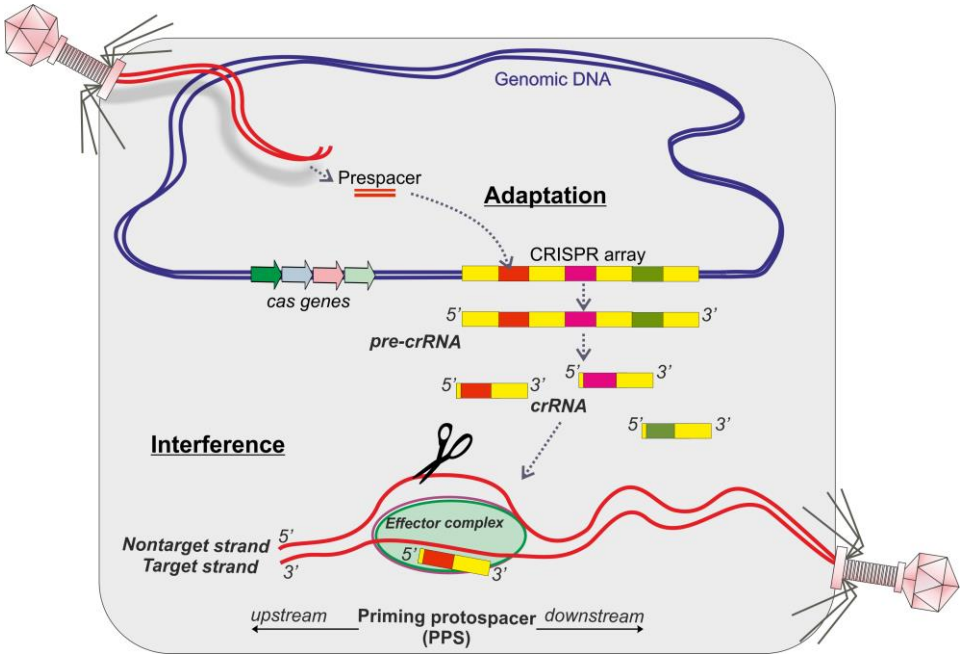


Figure 1. Overview of CRISPR-Cas system functions. Adapted from (Shiriaeva et al., 2018).

In DNA-targeting systems, binding of the targeted protospacer requires prior recognition of a short protospacer adjacent motif (**PAM**) (Deveau et al., 2008; Horvath et al., 2008; Mojica et al., 2009). This mechanism prevents the cleavage of CRISPR arrays that contain spacers complementary to crRNA but lack the PAM sequences. PAM recognition facilitates melting of dsDNA and pairing between a crRNA molecule and the protospacer (Gao et al., 2016; Hayes et al., 2016; Sashital et al., 2012; Sternberg et al., 2014). We will refer to the strand pairing with the crRNA as the target strand (**T-strand**), and the displaced DNA strand as the nontarget strand (**NT-strand**) (Figure 1).

Certain point mutations in protospacers or PAM weaken or abolish CRISPR interference resulting in production of escape phage particles (Barrangou et al., 2007; Deveau et al., 2008; Semenova et al., 2011). To overcome this, type I and type II CRISPR-Cas systems have evolved a mechanism allowing to restore phage resistance – **primed adaptation** (Datsenko et al., 2012; Nussenzweig et al., 2019; Swarts et al., 2012). Primed adaptation is a process of spacer acquisition taking place concurrently with CRISPR interference (Datsenko et al., 2012; Nussenzweig et al., 2019; Swarts et al., 2012). New spacers are selected from the regions flanking the protospacers targeted by the effector complex (Datsenko et al., 2012; Nussenzweig et al., 2019; Swarts et al., 2012). To differentiate between the protospacer initially targeted by the effector complex and new protospacers formed due to primed spacer acquisition, we will refer to the originally targeted protospacer '**priming protospacer**', or **PPS**. New protospacers can be located at both sides of the PPS. We refer to the regions at 5' and 3' sides of the PPS (relative to the PPS sequence in the NT-strand), as '**upstream**' and '**downstream**', respectively (Figure 1).

1.2 CRISPR interference by type I CRISPR-Cas systems

1.2.1 CRISPR interference module of type I CRISPR-Cas systems

Type I systems are the most prevalent and are found in 60% of CRISPR-containing genomes (Makarova et al., 2015). According to the current classification, type I systems are divided into 7 subtypes: I-A, I-B, I-C, I-D, I-E, I-F, and I-G (Makarova et

al., 2020). Type I-F systems are further subdivided into I-F1, I-F2, and I-F3 variants (Makarova et al., 2020).

The effector complex of type I systems is called Cascade (Figure 2A) (Brouns et al., 2008; Jore et al., 2011; Lintner et al., 2011; Nam et al., 2012a; Wiedenheft et al., 2011a, 2011b). It consists of multiple subunits including Cas5, Cas6, and Cas7 proteins belonging to the RAMP superfamily (Repeat-Associated Mysterious Protein) and containing an RNA-recognition motif allowing binding of the complex to crRNA (Haft et al., 2005; Makarova et al., 2006, 2011a; Reeks et al., 2013). The cleavage of pre-crRNA is carried out by Cas5 in type I-C systems or by Cas6 in other systems (Brouns et al., 2008; Carte et al., 2008; Garside et al., 2012; Haurwitz et al., 2010; Lintner et al., 2011; Nam et al., 2012b; Richter et al., 2012b). The cleavage occurs within repeats and yields mature crRNAs, each containing a spacer surrounded by repeat-derived 5' and 3' handles (Figure 2B, Figure 3) (Brouns et al., 2008; Carte et al., 2008; Garside et al., 2012; Haurwitz et al., 2010; Lintner et al., 2011; Nam et al., 2012a; Richter et al., 2012b). In addition to Cas5, Cas6, and Cas7, all Cascade complexes contain a 'large' subunit (>500 residues): Cas10d in subtype I-D and Cas8 in other subtypes (van Duijn et al., 2012; Jore et al., 2011; Lintner et al., 2011; Makarova et al., 2011a, 2015, 2020; Menon et al., 2009; Nam et al., 2012a; Plagens et al., 2014; Wiedenheft et al., 2011a, 2011b). Two subtypes, I-E and I-A, contain 'small' Cas11 subunits (<200 residues) (Jore et al., 2011; Lintner et al., 2011; Majumdar et al., 2015; Plagens et al., 2014; Wiedenheft et al., 2011b).

The signature protein of the type I systems is the Cas3 DExD/H helicase / HD nuclease (Makarova et al., 2011a, 2015, 2020; Sinkunas et al., 2011). Both nuclease and helicase domains are parts of a single Cas3 protein in type I-B, I-C, I-E, and I-G systems (Makarova et al., 2011a, 2015, 2020). In type I-A systems, the two activities are split between two separate proteins: Cas3' (helicase) and Cas3'' (nuclease) (Haft et al., 2005; Makarova et al., 2011a). The large subunit of the type I-D, Cas10d, is fused to an HD domain, which probably originated from the HD nuclease domain of Cas3 (Makarova et al., 2011b). Cas3' helicase is a standalone protein in the type I-D system (Makarova et

al., 2011a, 2015, 2020). In type I-F1 and I-F2 systems, Cas3 is fused to Cas2 (Makarova et al., 2011a, 2015, 2020).

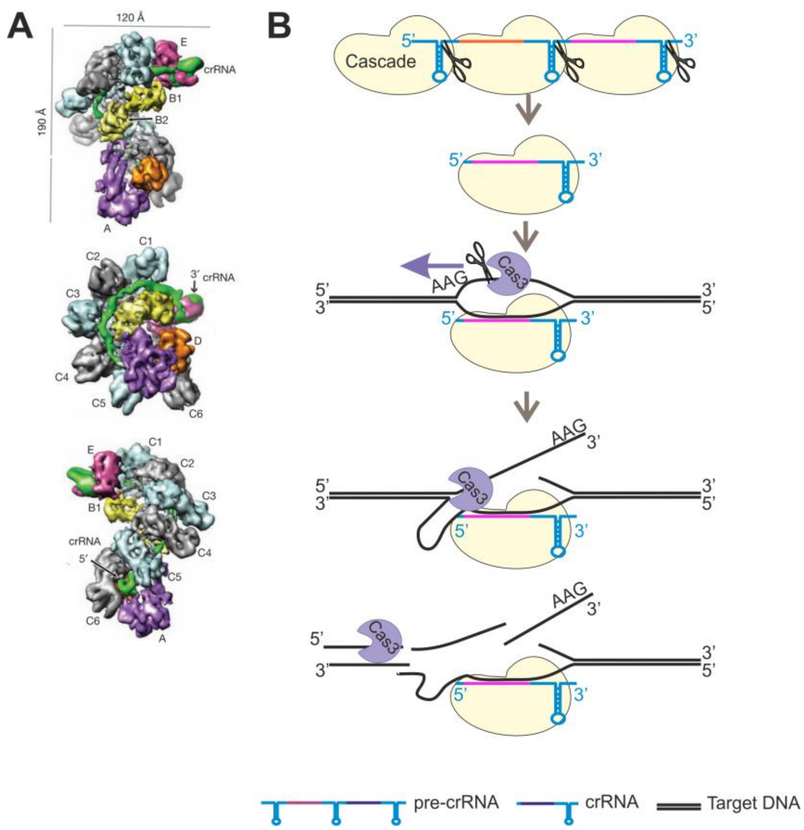


Figure 2. CRISPR interference in the type I-E CRISPR-Cas system. **A.** Structure of the *E. coli* Cascade bound to crRNA. Adapted from (Wiedenheft et al., 2011b) with permission. **B.** Schematic of pre-crRNA processing, target binding, and degradation by the type I-E CRISPR interference module.

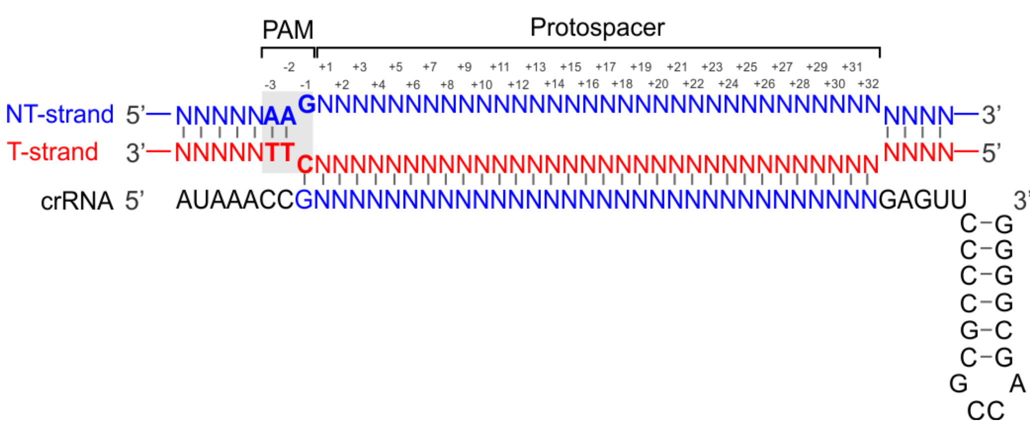


Figure 3. A protospacer targeted by the type I-E crRNA. Protospacer positions are numbered from +1 to +32. PAM positions are numbered from -3 to -1.

Cascade bound to a crRNA locates the PAM in dsDNA, forms base pairs with the complementary T-strand of the target protospacer, and displaces the NT-strand generating an R-loop (Jore et al., 2011; Lintner et al., 2011; Rollins et al., 2015; Sashital et al., 2012; Wiedenheft et al., 2011a). The protospacer-bound Cascade-crRNA complex recruits Cas3, which cleaves the NT-strand and degrades flanking DNA-regions (Majumdar et al., 2017; Nimkar and Anand, 2020; Rollins et al., 2017; Westra et al., 2012). In type I-A systems, Cas3' and Cas3'' are integral subunits of Cascade even in the absence of the target (Majumdar and Terns, 2019; Majumdar et al., 2015; Plagens et al., 2012, 2014).

1.2.2 CRISPR interference in type I-E CRISPR-Cas systems

The *E. coli* K12 strain contains two active CRISPR arrays and 8 *cas* genes belonging to the type I-E CRISPR-Cas system and located on the chromosome in the following order[1]: *cas3* (*ygcB*), *casA* (*cas8* / *cse1* / *ygcL*), *casB* (*cas11* / *cse2* / *ygcK*), *casC* (*cas7* / *cse4* / *ygcJ*), *casD* (*cas5* / *cas5e* / *ygcI*), *casE* (*cas6* / *cse3* / *ygcH*), *cas1* (*ygbT*), *cas2* (*ygbF*) (Brouns et al., 2008; Díez-Villaseñor et al., 2010; Pougach et al., 2010). Two promoters direct *cas* gene transcription: one located upstream of *cas3* and another – upstream of *cas8* (Majsec et al., 2016; Pul et al., 2010). Both promoters are repressed at least partially by H-NS (Majsec et al., 2016; Pul et al., 2010). The *cas8* promoter is also repressed by the CRP-cAMP complex in LB media without glucose (Yang et al., 2014). Due to transcriptional repression, CRISPR interference does not provide substantial protection against phages even when there is a match between a spacer and a protospacer (Pougach et al., 2010). The environmental signals naturally triggering *cas* gene expression remain unknown. Despite this fact, the *E. coli* type I-E system has been extensively studied in strains lacking *hns* (Pougach et al., 2010; Swarts et al., 2012), strains expressing H-NS antagonist LeuO (Westra et al., 2010), or strains overexpressing *cas* genes from inducible promoters (Brouns et al., 2008; Datsenko et al., 2012; Yosef et al., 2012).

[1] The alternative names for the same genes are provided in parentheses

The *E. coli* Cascade is a seahorse-shaped complex composed of 11 subunits with the following stoichiometry: $CasA_1$:$CasB_2$:$CasC_6$:$CasD_1$:$CasE_1$ (Figure 2A) (Jackson et al., 2014a; Jore et al., 2011; Wiedenheft et al., 2011b). CasE cleaves a pre-crRNA molecule to the 3' side of hairpins inside repeats yielding mature crRNAs each containing a 7-nt repeat-derived 5' handle, a 33-nt spacer, and a 21-nt repeat-derived 3' handle (Figure 2B, Figure 3) (Brouns et al., 2008; Jore et al., 2011). In the Cascade-crRNA complex, CasE remains bound to the 3' hairpin of the crRNA, CasD makes sequence-specific interactions with the 5' handle while CasC forms a helical backbone interacting with the crRNA spacer nonspecifically (Figure 2A) (Jackson et al., 2014a; Wiedenheft et al., 2011b). The large Cascade subunit CasA is located in proximity to CasD (Jackson et al., 2014a; Wiedenheft et al., 2011b). Two small CasB subunits are assembled along the $CasC_6$ backbone and do not make contacts with the crRNA (Jackson et al., 2014a; Wiedenheft et al., 2011b).

The Cascade-crRNA complex recognizes a protospacer in either ss- or dsDNA (Jore et al., 2011). For the efficient location of a protospacer by Cascade, the protospacer must be flanked by a PAM sequence. The PAM is recognized as a duplex (Hayes et al., 2016). When talking about the PAM sequence we will refer to the motif located to the 5' side of the targeted protospacer in the NT-strand (Figure 3). Historically, the *E. coli* PAM is regarded as a 3-nt AWG sequence though the last G nucleotide corresponds to the nucleotide which is incorporated into the CRISPR array together with a new spacer, and therefore can be also regarded as the first nucleotide of a protospacer (see section 1.3.2) (Mojica et al., 2009). We follow the existing numbering of PAM nucleotides and refer to the PAM nucleotide closest to the protospacer as -1 position (G of the consensus *E. coli* type I-E PAM), the middle PAM nucleotide as -2 position (W of the consensus *E. coli* type I-E PAM), and the nucleotide which is most distant from the PPS as -3 position (A of the consensus *E. coli* type I-E PAM) (Figure 3). About 20 different PAM variants including the consensus 5'-AWG-3' sequence promote efficient interference in *E. coli* (Fineran et al., 2014; Fu et al., 2017; Musharova et al., 2019; Westra et al., 2012; Xue et al., 2015). PAM sequences prevent autoimmunity allowing the effector complex to

distinguish between a protospacer in a target and the spacer with an identical sequence in the CRISPR array (Semenova et al., 2011; Westra et al., 2012, 2013). The sequence corresponding to the PAM at the repeat/spacer boundary of the CRISPR array is CCG. The introduction of C into the -2 or -3 PAM position decreases the binding of Cascade to the target and abolishes interference (Semenova et al., 2011; Xue et al., 2015).

The PAM sequence is recognized by the large CasA subunit of Cascade (Figure 2A) (Jore et al., 2011; Sashital et al., 2012). The recognition of the PAM locally destabilizes the adjacent protospacer base pairs facilitating DNA unwinding and pairing of the T-strand with the crRNA spacer (van Erp et al., 2015; Hayes et al., 2016; Jore et al., 2011; Sashital et al., 2012; Xiao et al., 2017; Xue et al., 2017). The first 8 bp of protospacers, known as a 'seed' sequence, are critical because R-loop generation proceeds directionally from the seed sequence to the PAM-distal end (Rutkauskas et al., 2015; Semenova et al., 2011). Spacer-protospacer mismatches in the seed region decrease the efficiency of interference (Fineran et al., 2014; Semenova et al., 2011; Xue et al., 2015).

Upon the completion of base pairing between a crRNA spacer and the T-strand of the protospacer Cascade undergoes conformational changes stabilizing the R-loop and triggering the recruitment of Cas3 helicase/nuclease (Figure 2B) (van Erp et al., 2018; Hayes et al., 2016; Hochstrasser et al., 2014; Xiao et al., 2017, 2018; Xue et al., 2016). The type I-E Cas3 proteins combine endonuclease activity on ssDNA (provided by the N-terminal HD-domain) and 3'→5' helicase activity (provided by the C-terminal Superfamily 2 (SF2) domain) (Gong et al., 2014; Huo et al., 2014; Jackson et al., 2014b; Sinkunas et al., 2011). Both activities are required for CRISPR interference (Westra et al., 2012). Upon binding to Cascade-crRNA, Cas3 introduces a nick in the NT-strand of the R-loop and uses the generated 3' terminus to initiate unidirectional unwinding from the PAM-proximal region to several dozen kilobases upstream of the PPS (Figure 2B) (Dillard et al., 2018; Hochstrasser et al., 2014; Loeff et al., 2018; Mulepati and Bailey, 2013; Redding et al., 2015; Xiao et al., 2017, 2018). Single-molecule experiments demonstrate that Cas3 remains bound to Cascade at least for some time while it reels the

NT-strand, generating a loop in the T-strand (Figure 2B) (Dillard et al., 2018; Loeff et al., 2018). In some experiments Cas3 was also demonstrated to move downstream, i. e., opposite to the main direction but the mechanism of strand switching required for such movement remains unknown (Redding et al., 2015).

In bulk biochemical experiments Cas3 digested a Cascade-bound target into fragments of various lengths, from less than 23 nt to 1500 nt (Hochstrasser et al., 2014; Künne et al., 2016; Mulepati and Bailey, 2013). However, very limited degradation was observed in single-molecule studies (Dillard et al., 2018; Loeff et al., 2018; Redding et al., 2015). It was suggested that the difference could be attributed to 10- to 500-fold excess of Cas3 used in the bulk experiments (Loeff et al., 2018). *In vivo* CRISPR interference leads to plasmid loss or degradation of up to 30-kbp segments of phage DNA (Semenova et al., 2016; Strotskaya et al., 2017). It is possible that host nucleases participate in degradation but their contribution to interference has not been studied.

1.3 CRISPR adaptation in type I CRISPR-Cas systems

1.3.1 Genetic requirements for adaptation in various type I systems

CRISPR interference is possible only if there is an appropriate spacer in a CRISPR array that was incorporated during the adaptation stage. CRISPR adaptation (spacer acquisition) was detected in all seven subtypes of type I CRISPR-Cas systems (Almendros et al., 2019; Cady et al., 2012; Erdmann and Garrett, 2012; Kieper et al., 2018; Li et al., 2014; Rao et al., 2016; Yosef et al., 2012). Cas1 and Cas2 proteins are the only two Cas proteins absolutely required for integration of a prespacer into the CRISPR array *in vitro* (Fagerlund et al., 2017; Nuñez et al., 2014, 2015a). *In vivo* expression of *cas1* and *cas2* is sufficient for naïve adaptation in some systems including the type I-E system of *E. coli* (Kieper et al., 2018; Shiimori et al., 2018; Yosef et al., 2012). Naïve adaptation in the type I-F system requires catalytically active Cas3 and the effector complex, though the mechanism for this requirement is unknown (Vorontsova et al., 2015). Most type I systems (except I-E and I-F) encode Cas4 proteins that are not essential for spacer acquisition but enhance the efficiency of naïve adaptation and affect

spacer choice (Almendros et al., 2019; Kieper et al., 2018; Shiimori et al., 2018; Zhang et al., 2019). There is also an example of the type I-A system of *Sulfolobus islandicus* encoding a transcriptional activator of *cas* genes, Csa3a, which is therefore required for naïve adaptation (Liu et al., 2015, 2017).

In addition to *cas1* and *cas2* genes, primed adaptation requires an intact module of CRISPR interference and a preexisting spacer targeting a PPS (Datsenko et al., 2012; Garrett et al., 2020; Li et al., 2014). Primed adaptation is coupled to CRISPR interference and newly acquired spacers originate from the regions adjacent to the PPS (Almendros et al., 2019; Datsenko et al., 2012; Garrett et al., 2020; Li et al., 2014; Rao et al., 2017; Richter et al., 2014; Swarts et al., 2012; Vorontsova et al., 2015). With the exception of the only experimentally characterized type I-G system of *Geobacter sulfurreducens* (Almendros et al., 2019), the orientation of the PPS dictates the orientation of protospacers selected as donors of new spacers during primed adaptation (Datsenko et al., 2012; Li et al., 2014; Rao et al., 2017; Richter et al., 2014; Swarts et al., 2012; Vorontsova et al., 2015). This means that if non-transcribed strand sequences of acquired spacers are mapped to the target, almost all sequences upstream of the PPS will be mapped to one strand, while downstream of the PPS - to the opposite strand.

In experiments revealing robust primed adaptation from a PPS-containing plasmid in the *E. coli* type I-E system, no visible adaptation was observed in a control sample transformed with a plasmid that did not have the PPS (Savitskaya et al., 2013). This means that primed adaptation is much more efficient than naïve adaptation. An approximately 500-fold difference in the efficiencies of primed and naïve adaptation was estimated for the type I-F system (Staals et al., 2016). Both modes of adaptation rely on the same Cas1 and Cas2 proteins for integration of prespacers into the CRISPR array (Fagerlund et al., 2017; Nuñez et al., 2014, 2015a). It is likely that highly efficient prespacer generation by the coordinated action of the interference and adaptation modules results in the greater efficiency of primed adaptation.

1.3.2 Integration of prespacers into the CRISPR array by the Cas1-Cas2 complex

Both naïve and primed adaptation have been extensively studied in the type I-E system, which is also the main subject of this thesis research. Therefore, we will review the mechanisms of adaptation in the type I-E system in greater detail.

Naïve adaptation in the type I-E system requires the presence of *cas1*, *cas2*, and the CRISPR-array with an adjacent leader sequence (Díez-Villaseñor et al., 2013; Yosef et al., 2012). The Cas1 and Cas2 proteins form a symmetrical butterfly-shaped complex composed of two Cas2 subunits sandwiched between two Cas1 dimers (Figure 4A, Figure 5A) (Nuñez et al., 2014, 2015b). This complex catalyzes the integration of a 33-bp double-stranded oligonucleotide mimicking a prespacer into the CRISPR array placed on a plasmid (Nuñez et al., 2015a). The integration reaction proceeds via two nucleophilic attacks at the boundaries of the first repeat by the 3'-hydroxyl groups of the oligonucleotide (Figure 5B) (Nuñez et al., 2015a). The rate of integration at the leader/repeat junction in the 'top' strand is ≈14-fold higher than at the repeat/spacer junction in the 'bottom' strand (Nuñez et al., 2016). Based on this observation, it was suggested that the top-strand reaction occurs first (Nuñez et al., 2016). After the completion of the second reaction, the two strands of the repeat are separated by the integrated prespacer that results in an intermediate with two single-stranded gaps in the top and bottom strands of the first and second repeat, respectively (Figure 5B) (Arslan et al., 2014; Nuñez et al., 2015a). DNA polymerase I is a candidate for filling the gaps based on its requirement for naïve and primed adaptation (Ivančić-Baće et al., 2015). The ligase sealing the nicks is not identified.

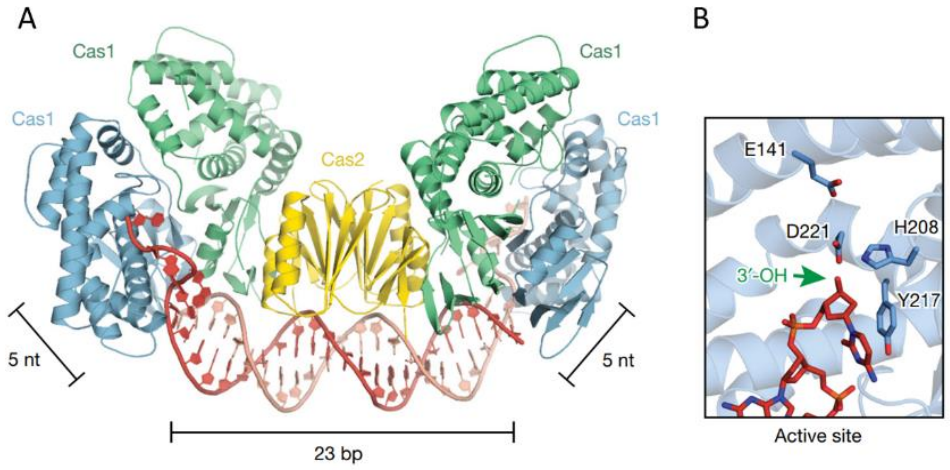


Figure 4. The architecture of the Cas1-Cas2 complex. **A.** The overall structure of the Cas1-Cas2 complex bound to a 23-bp duplexed oligonucleotide with two 5-nt 3' overhangs. **B.** Configuration of the Cas1 active site coordinating the 3' end. Adapted from (Nuñez et al., 2015b) with permission.

The formation of the Cas1-Cas2 complex is required for adaptation since mutations disrupting the interaction of Cas1 and Cas2 abolish spacer integration *in vitro* and *in vivo* (Nuñez et al., 2014, 2015a). Though the biochemical activity of purified *E. coli* Cas2 has not been tested so far, some of its homologs have RNAse and DNase activities (Beloglazova et al., 2008; Dixit et al., 2016; Ka et al., 2014, 2017; Nam et al., 2012b). However, it is unlikely that the nuclease activity of *E. coli* Cas2 is involved in spacer acquisition since the substitution of a conserved metal-coordinating residue does not influence prespacer integration (Nuñez et al., 2014, 2015a). The *E. coli* Cas1 protein is an endonuclease cleaving ssRNA, ssDNA, and dsDNA (Babu et al., 2011). The active site is located in the C-terminal domain, and substitutions of conserved metal-coordinating amino acids (E141A, H208A, D218A, D221A) abolish the nuclease activity (Figure 4B) (Babu et al., 2011). The same residues are required for spacer integration *in vitro* and *in vivo* (Nuñez et al., 2014, 2015a; Yosef et al., 2012).

The Cas1-Cas2 complex binds to various double-stranded oligonucleotides of approximately spacer size (Moch et al., 2017; Nuñez et al., 2015b, 2015a; Wang et al., 2015). The best substrate for integration *in vitro* is a 23-bp duplex flanked by two 5-nt 3' overhangs or a double-stranded 33-bp molecule with five terminal base pairs at both sides splayed due to mismatches (Nuñez et al., 2015b, 2015a). In a crystal structure of the

Cas1-Cas2 complex bound to the preferred *in vitro* substrate, the duplex region lies on the Cas2 dimer (Figure 4A) (Nuñez et al., 2015b). At the interface of Cas2 and Cas1, the Cas1 Y22 residue splits the double helix displacing the 5' end from the Cas1-Cas2 complex and directing the 3' end into the Cas1 'arginine channel' leading to the Cas1 active site within the same subunit (Nuñez et al., 2015b). Therefore, only two of four Cas1 subunits are catalytically active and bind nucleophilic 3' ends (Nuñez et al., 2015b).

Initial *in vitro* experiments demonstrated the integration of a duplex oligonucleotide into a CRISPR array on a supercoiled but not on a linearized or relaxed plasmid (Nuñez et al., 2015a). Moreover, *in vitro* integration happened after every repeat of the array while a strong preference for the insertion at the leader/repeat boundary was revealed *in vivo* (Nuñez et al., 2015a; Yosef et al., 2012). This discrepancy was attributed to an additional non-Cas protein called IHF during *in vivo* spacer acquisition (Nuñez et al., 2016). IHF is a histone-like protein that binds to an AT-rich sequence and bends DNA by >160° (Craig and Nash, 1984; Rice et al., 1996). The IHF binding site was found in the CRISPR leader sequence (Nuñez et al., 2016). Binding of IHF to the leader stimulates the recruitment of Cas1-Cas2 (Figure 5B) (Yoganand et al., 2017). IHF bends the leader sequence bringing in contact the non-catalytic Cas1 subunit with a sequence in the leader (Wright et al., 2017; Yoganand et al., 2017). It is speculated that due to the bent structure of supercoiled plasmids, Cas1-Cas2 does not require IHF in locating the leader of plasmid-borne CRISPR arrays (Yoganand et al., 2017).

Fully or partially double-stranded 33-bp oligonucleotides were used in most *in vitro* experiments, but the precise structure of prespacers *in vivo* remains unknown. Sequencing of newly acquired spacers revealed that ≈40% and ≈95% of parental protospacers are associated with the 5'-AAG-3' PAM in naïve and primed adaptation, respectively (Radovcic et al., 2018; Savitskaya et al., 2013; Yosef et al., 2012, 2013).

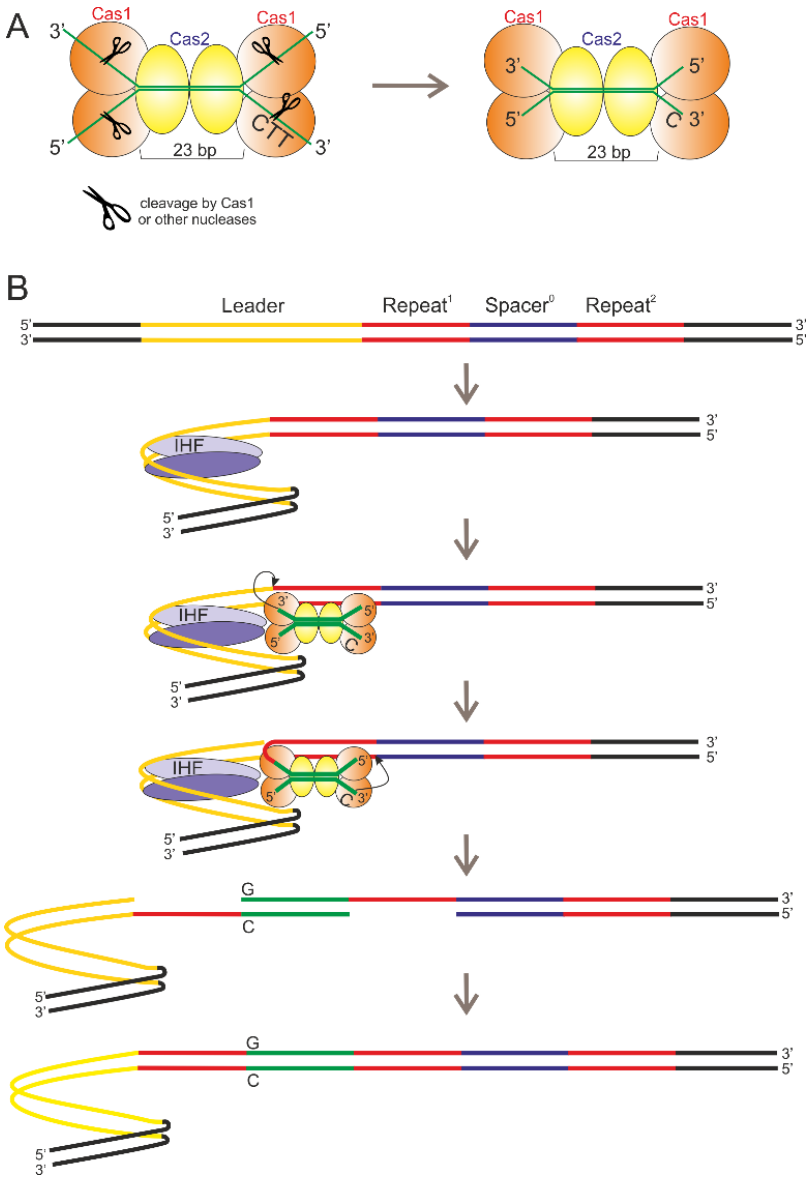


Figure 5. Model of prespacer generation and integration into the CRISPR array. **A.** Schematic representation of the Cas1-Cas2 complex bound to a PAM-containing prespacer (shown in green) with long 3' and 5' overhangs that are trimmed by Cas1 or other cellular nucleases to produce a mature prespacer suitable for integration. **B.** Prespacer integration into the CRISPR array by Cas1-Cas2 assisted by IHF bound to the leader sequence. Adapted from (Shiriaeva et al., 2018).

Spacers are inserted into the CRISPR array in a specific orientation so that the base in PAM -1 position (a G in the case of the *E. coli* consensus PAM) becomes the last base of the first repeat (Datsenko et al., 2012; Swarts et al., 2012). High preference for AAG-flanked protospacers during naïve adaptation in cells devoid of the interference module highlights the existence of an intrinsic PAM specificity within the adaptation

module (Radovcic et al., 2018; Yosef et al., 2012, 2013). A crystal structure of the Cas1-Cas2 complex bound to a 23-bp duplex flanked by two 10-nt 3' single-stranded overhangs, each containing the 3'-TTC-5' sequence complementary to the most prevalent 5'-AAG-3' PAM, was determined (Wang et al., 2015). This structure revealed a pocket in the Cas1 C-terminal domain that binds to the 3'-TTC-5' sequence in a base-specific manner (Wang et al., 2015). In addition, Cas1 cut the 3'-TTC-5' sequence between the C and T bases (Wang et al., 2015). Based on these observations, Wang et al. suggested a model according to which Cas1-Cas2 binds to a TTC-containing spacer precursor, incises it between the C and T bases generating the PAM-derived 3' end with a terminal C nucleotide, and makes the second cut in the opposite strand at a distance of 33 bp from the PAM-derived C (Figure 5A) (Wang et al., 2015).

The mechanism of prespacer trimming has been a subject of extensive studies over the past 3 years and the model of prespacer trimming by Cas1 proposed by Wang et al. is currently in doubt (see below). Several groups reported that they were unable to detect cleavage of prespacer 3' ends by the Cas1-Cas2 complex (Kim et al., 2020; Ramachandran et al., 2020; Yoganand et al., 2019). Instead, it was shown that other nucleases trim the 3' ends while Cas1-Cas2 actually limits the degradation. Drabavicius et al. characterized *in vitro* a type I-E CRISPR-Cas system of *Streptococcus thermophilus* DGCC7710. Unlike the type I-E system of *E. coli*, the Cas2 protein of *S. thermophilus* is fused to a DnaQ domain possessing 3'→5' exonuclease activity (Drabavicius et al., 2018). The *S. thermophilus* Cas1/Cas2-DnaQ complex trimmed long single-stranded 3' overhangs attached to duplex oligonucleotides through the exonuclease activity of DnaQ and integrated the processed substrates into a supercoiled plasmid even if it lacked CRISPR (Drabavicius et al., 2018). Inactivation of the DnaQ active site inhibited the integration reaction (Drabavicius et al., 2018). Inspired by these results other groups set out to determine which 3'→5' exonucleases may be responsible for trimming prespacer 3' ends in *E. coli*. Based on *in vitro* studies, two exonucleases from the DnaQ superfamily, ExoT and the proofreading ε subunit of PolIII, are potential candidates for this role (Kim et al., 2020; Ramachandran et al., 2020).

It should be noted that type I systems other than I-E and I-F encode the Cas4 protein, which, in some cases, is fused to Cas1 (Hudaiberdiev et al., 2017; Makarova et al., 2015). Analysis of spacer acquisition by type I-A and type I-B CRISPR-Cas systems of *Pyrococcus furiosus* with a single module of adaptation (Shiimori et al., 2018) and by the type I-G system of *Geobacter sulfurreducens* (Almendros et al., 2019) revealed that though Cas4 is not required for adaptation, it enhances spacer acquisition efficiency. A similar result was obtained for the type I-D adaptation module of *Synechocystis* sp. 6803 transferred to a heterologous *E. coli* host (but only if the host contained a deletion of *recB* or *recC*) (Kieper et al., 2018). Cas4 is required for primed adaptation in the type I-B system of *Haloarcula hispanica* as revealed by PCR followed by agarose gel-electrophoresis that allows for detection of spacer acquisition in at least 1% of cells (Li et al., 2014). Therefore, it can not be ruled out that a low level of spacer acquisition was retained by *Δcas4* cells of *H. hispanica* but could not be detected by the method used. A similar situation was described for the type I-A system of *Sulfolobus islandicus* where Cas4 was considered essential for adaptation in an early work (Liu et al., 2017) but low-efficiency spacer acquisition was detected for *Δcas4* mutants later using a high-throughput sequencing approach (Zhang et al., 2019).

Most strikingly, in all experiments where the sequences of newly acquired spacers from strains lacking *cas4* or expressing mutant *cas4* genes were determined, the corresponding protospacers were not flanked by correct PAMs (Almendros et al., 2019; Kieper et al., 2018; Shiimori et al., 2018; Zhang et al., 2019). Moreover, the inactivation of Cas4 led to the acquisition of spacers that were, on average, longer than spacers acquired by strains with the intact adaptation modules, at least in some of the studied CRISPR-Cas systems (I-D of *Synechocystis* sp. 6803 and I-A/I-B systems of *P. furiosus*) (Kieper et al., 2018; Shiimori et al., 2018). An influence of Cas4 on spacer lengths was also shown for the type I-A system of *S. islandicus*, though contradictory results were obtained: while an increase in an average spacer length was demonstrated for strains with point mutations in *cas4*, shorter spacers were acquired by cells with a deletion of *cas4* (Zhang et al., 2019).

Though the exact biochemical activities of purified Cas4 vary between different systems, DNA unwinding, exonuclease, and endonuclease activities have been reported (Lee et al., 2018, 2019; Lemak et al., 2013, 2014; Zhang et al., 2012). Experiments *in vitro* demonstrate that Cas4 proteins from the type I-C and type I-A systems trim the 3' overhangs of partially duplex oligonucleotides in a PAM-dependent manner in the presence of Cas1-Cas2 (Lee et al., 2018, 2019; Rollie et al., 2018). In line with this, Cas4 residues that are likely required for nuclease activity are also essential for PAM specificity *in vivo* (Almendros et al., 2019; Kieper et al., 2018; Shiimori et al., 2018). Together, these results provide a basis for a model according to which the Cas1-Cas2 complex of the Cas4-containing CRISPR-Cas systems binds to a prespacer containing a correct PAM, which is recognized and trimmed by Cas4. In contrast, the Cas1-Cas2 complex of the type I-E or the I-F system lacking Cas4 has the ability to recognize a correct PAM on its own but likely relies on host nucleases in the generation of prespacer 3' ends.

Little is known about the processing of prespacer 5' ends. It was shown that the type I-E Cas1-Cas2 complex protects 63-bp double-stranded oligonucleotides from full degradation by the bacteriophage T5 5'→3' exonuclease and yields DNA fragments approximately of spacer size (Yoganand et al., 2019). *E. coli* 5'→3' exonucleases have not been tested for their ability to trim prespacer 5' ends so far.

The existing model of prespacer integration into the CRISPR array yields a correctly oriented spacer only if we assume that the attack with the hydroxyl group of the PAM-derived C (PAM -1 position) occurs as the second integration reaction into the bottom strand (Figure 5B). However, a 33-bp prespacer starting with a G/C pair is integrated into the CRISPR array in both orientations with equal efficiency, as was shown by *in vivo* experiments from the Church laboratory (Shipman et al., 2016). Shipman et al. electroporated oligonucleotides of various structures into *E. coli* cells containing a single CRISPR array in the genome and overexpressing *cas1* and *cas2* genes from a plasmid (Shipman et al., 2016). The authors expected that the oligonucleotides would be bound by the Cas1-Cas2 complex as prespacers and integrated into the CRISPR array, which

indeed was observed (Shipman et al., 2016). The efficiency of integration and orientation of newly acquired spacers was determined by high-throughput sequencing of the PCR amplicons derived from extended and non-extended arrays (Shipman et al., 2016). The authors demonstrated that when a 35-bp double-stranded oligonucleotide starting with the complete PAM 5'-AAG-3'/3'-TTC-5' was used instead of the 33-bp substrate starting with a G/C pair, the efficiency of integration increased ≈5-fold and more than 90% of integration events resulted in insertion of correctly oriented spacers (Shipman et al., 2016). A similar result was observed for a 45-bp oligonucleotide with 10 additional base pairs upstream of 5'-AAG-3'/3'-TTC-5' (Shipman et al., 2016). Integration of oligonucleotides into the CRISPR array by Cas1-Cas2 argues that if PAM-containing DNA fragments are produced by cellular processes, these fragments can be trimmed to the length of mature spacers and incorporated into the CRISPR array, at least during naïve adaptation.

1.3.3 Possible mechanisms of prespacer generation during naïve adaptation

Cellular processes feeding Cas1-Cas2 with substrates for naïve adaptation include DNA replication and repair. Levy et al. studied naïve adaptation by introducing the *cas1* and *cas2* genes on a plasmid into an *E. coli* strain lacking *cas* genes but containing a CRISPR array in the genome (Levy et al., 2015). Since the interference module was not present, spacers originating from plasmid and genomic DNA were preserved in populations. Using this system, Levy et al. demonstrated that naïve adaptation depends on replication (Levy et al., 2015). In a *dnaC2* mutant unable to initiate replication at 39 °C but replicating at 30 °C, the percentage of expanded arrays dropped at least 200-fold at non-permissive temperature (Levy et al., 2015). High-throughput sequencing of spacers acquired by wild-type cells revealed that the corresponding protospacers are not randomly distributed throughout the genome (Levy et al., 2015). Two protospacer hotspots were detected in the terminus region close to the *terC* and *terA* sites (Levy et al., 2015). The third hotspot was adjacent to the CRISPR array (Levy et al., 2015). In all three cases, the region enriched with protospacers stretched ≈5-50 kbp from the listed sites in the direction opposite to the direction of replication fork movement and was

limited by the first encountered Chi site – a sequence regulating double-strand break repair via RecBCD pathway (Dabert et al., 1992; Dixon and Kowalczykowski, 1993; Levy et al., 2015). Two replication forks progressing from the origin meet in the terminus region, move past each other, and get stalled by the Tus protein bound to *ter* sequences (Neylon et al., 2005). Integration of a prespacer into the CRISPR array creates gaps in both strands (Figure 5B) (Arslan et al., 2014). If a replication fork reaches a single-strand break, a double-strand break (DSB) is produced (Kuzminov, 1995, 2001). Therefore, a DSB should be produced when the replication fork reaches the CRISPR array with the unrepaired gap. In cells expressing the I-SceI endonuclease, a new protospacer hotspot appears near a single I-SceI cleavage site introduced in the genome (Levy et al., 2015). Mutations of *recB, recC*, and *recD* genes, which encode the subunits of the RecBCD complex initiating double-strand break repair, reduce naïve adaptation (Ivančić-Baće et al., 2015; Levy et al., 2015; Radovcic et al., 2018). Moreover, Cas1 was shown to co-purify with RecB, RecC and another double-strand break repair protein RuvB (Babu et al., 2011). Taken together these results suggest that the sources of prespacers during naïve adaptation are regions adjacent to free DNA ends produced in result of DSBs or replication fork stalling. DNA repair proteins are involved in naïve adaptation.

An interplay between DNA repair and CRISPR adaptation is further supported by evidence from other type I systems:

- In *Pyrococcus furiosus* containing type I-A, I-B, and III-B CRISPR-Cas systems with a shared adaptation module, the protospacer hotspots coincide with regions where free DNA termini are expected (the double-strand origin of plasmid rolling circle replication, CRISPR arrays, and active transposons) or regions where recombination between the chromosome and plasmids occurs (Shiimori et al., 2017). The enrichment of protospacers was also detected for highly transcribed rRNA genes presumably due to frequent nicking of the displaced DNA strand in R-loops (Shiimori et al., 2017).
- The enrichment of type I-G and I-D protospacers was detected near *ter* sites (Almendros et al., 2019; Kieper et al., 2018).

- In the type I-A system of *Sulfolobus islandicus*, a transcriptional activator of *cas* operons also binds to promoters of DNA repair genes and activates their transcription (Liu et al., 2017).
- Positive association between some DNA repair genes and type I-B, I-C, I-E, and I-F systems was discovered in genomes of Proteobacteria and Firmicutes (Bernheim et al., 2019).

To gain an insight into the possible mechanisms of the interaction between naïve adaptation and CRISPR adaptation we will review the pathways of DSB repair in *E. coli*.

1.3.4 DNA double-strand break repair as a potential source of spacer precursors

1.3.4.1 An overview of DSB repair by homologous recombination

Double-strand breaks (DSBs) may appear in DNA directly after exposure to DNA damaging agents such as γ- or X-rays (Figure 6A) (Roots et al., 1985). Such breaks are called two-ended DSBs. Alternatively, when a replication fork encounters a single-strand nick in template DNA, a one-ended DSB is produced (Figure 6B) (Kuzminov, 1995, 2001). DSB repair in *E. coli* requires homologous dsDNA and RecA (Clark and Margulies, 1965; Krasin and Hutchinson, 1977). RecA is a strand exchange protein that binds to a single-stranded 3'-terminated tail and catalyzes the search for a homologous region and exchange of DNA strands between the two molecules, generating a D-loop (Anderson and Kowalczykowski, 1997a; Cassuto et al., 1980; Cox and Lehman, 1981; Forget and Kowalczykowski, 2012; Lesterlin et al., 2014; McEntee et al., 1979; Ragunathan et al., 2011; Shibata et al., 1979; West et al., 1980). PriA is recruited to the D-loop where it initiates DNA synthesis from the invading 3'-ended single-stranded tail using the intact homologous partner as a template (Heller and Marians, 2006; Michel and Sandler, 2017; Xu and Marians, 2002). This allows for restoring the broken sequence (Figure 6). Finally, the Holliday junctions (HJs) are resolved according to the classical models of homologous recombination (Figure 6) (Resnick, 1976; Szostak et al., 1983).

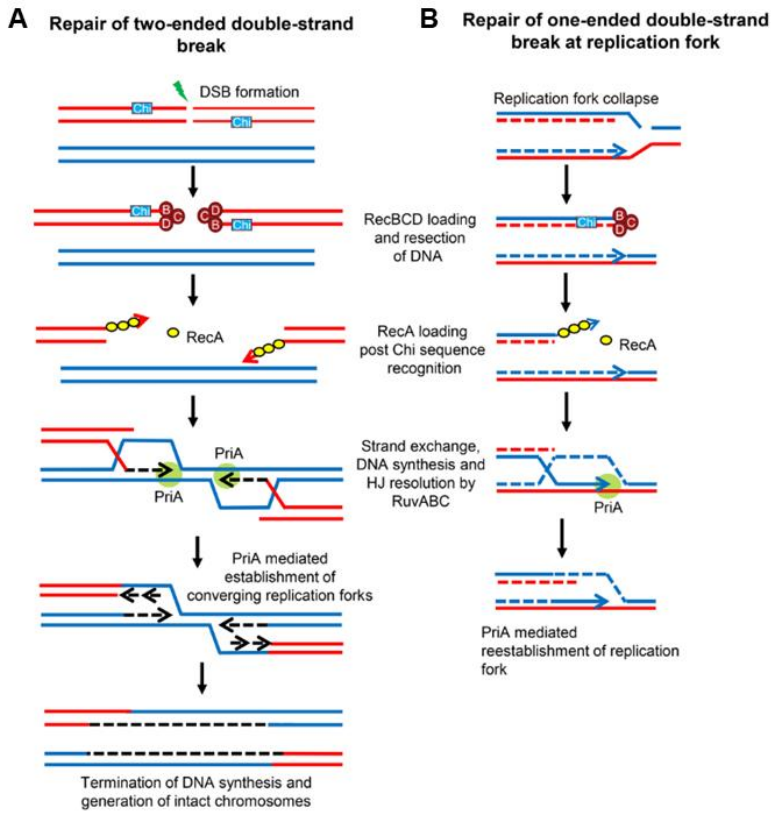


Figure 6. Overview of RecBCD-dependent homologous recombination. The repair of a two-ended DSB generated due to exposure to DNA-damaging agents (**A**) or a one-ended DSB formed upon encountering a nick by a replication fork (**B**) is shown. In both cases, the double-strand ends are resected to generate 3'-ended single-stranded tails terminated at Chi sequences. The 3' overhangs are covered with the RecA protein, which catalyzes the invasion into homologous duplexes. The generated D-loops are targeted by the PriA protein (Michel and Sandler, 2017) that initiates an assembly of a primosome, ultimately leading to the establishment of converging replication forks restoring the sequence lost during end resection (**A**) or the restart of the replication fork (**B**). Holliday junctions (HJs) are resolved to produce two intact DNA molecules. Dotted lines indicate newly synthesized strands, solid lines indicate template strands. Arrows indicate the 3' ends of the leading strands. The figure was adapted from (Sinha et al., 2020) distributed under the terms of the Creative Commons CC BY license (no permission is required).

At physiological pH, RecA nucleates the formation of a filament on SSB-coated ssDNA slowly, indicating the need for accessory factors (Bell et al., 2012). Besides, a single-stranded 3'-terminated tail should be produced from a double-strand end prior to RecA loading (Figure 6). There are several pathways in *E. coli* providing suitable DNA substrates and facilitating RecA nucleation. The RecBCD pathway is responsible for 99% of homologous recombination events in wild-type cells (Barbour et al., 1970; Dillingham and Kowalczykowski, 2008; Kushner et al., 1971; Repar et al., 2013; Willetts and Mount,

1969). In cells devoid of RecBCD, additional suppressor mutations activate the RecFOR or RecE pathways (Barbour et al., 1970; Horii and Clark, 1973; Kushner et al., 1971; Smith, 1989).

1.3.4.2 RecBCD pathway of homologous recombination

RecBCD, also known as Exonuclease V, is a heterotrimeric enzyme composed of RecB, RecC, and RecD subunits encoded by *recB*, *recC*, and *recD* genes, respectively (Figure 7) (Amundsen et al., 1986; Wright et al., 1971). The purified RecBCD complex is active as Mg^{2+}- and ATP-dependent exonuclease on dsDNA and, to a lesser extent, on ssDNA (Goldmark and Linn, 1970, 1972; Wright et al., 1971). RecBCD also exhibits Mg^{2+}-dependent endonuclease activity on ssDNA but this activity[2] is ≈30-50-fold lower than the RecBCD exonuclease activities[3] (Goldmark and Linn, 1970, 1972; Wright et al., 1971). Closed circular or nicked dsDNA is not degraded by the enzyme (Goldmark and Linn, 1972; Karu et al., 1973). RecBCD slowly degrades gapped circles (Karu et al., 1973; Taylor and Smith, 1985). Presumably, these events are initiated by inefficient endonucleolytic cleavage in the single-stranded regions followed by highly efficient degradation of linearized DNA by the exonuclease activity (Taylor and Smith, 1985).

In line with the inability of RecBCD to cleave circular dsDNA, very little binding of dsDNA with no accessible ends (circular dsDNA or linear dsDNA with terminal hairpins) was observed (Taylor and Smith, 1985, 1995a). The preferred substrate for RecBCD binding is linear dsDNA with blunt or nearly blunt ends containing a few nucleotides on the 5' or 3' end (Taylor and Smith, 1985, 1995a). The enzyme binds to such ends with $K_D \approx 0.1\text{-}7$ nM with the tightest binding reported for 4-nt 5' overhangs (Taylor and Smith, 1995a). Binding to single-stranded oligonucleotides is weaker ($K_D \approx$ 50-250 nM) than to duplex substrates (Taylor and Smith, 1995a).

Though the RecBCD exonuclease activity is always accompanied by ATP hydrolysis (Goldmark and Linn, 1972), the opposite is not true (Rosamond et al., 1979).

[2] Endonuclease activity is measured in units where one unit makes 1 nmole of circular single-stranded fd phage DNA susceptible to exonuclease I in 30 min (Goldmark and Linn, 1972).
[3] Exonuclease activity is measured in units where one unit converts 1 nmole of *E. coli* DNA into acid-soluble fragments in 30 min (Goldmark and Linn, 1972).

The ATPase activity requires the presence of DNA but can proceed under conditions inhibiting the nuclease activities, for example in the presence of Ca^{2+}, at high ATP concentration or when DNA treated with agents causing interstrand crosslinks is used (Karu and Linn, 1972; Karu et al., 1973; Rosamond et al., 1979). Electron microscopy of dsDNA treated with RecBCD under these conditions revealed duplex molecules with single-stranded tails and/or single-stranded loops suggesting that RecBCD is a helicase (Braedt and Smith, 1989; Muskavitch and Linn, 1982; Rosamond et al., 1979; Taylor and Smith, 1980). A limited number of nicks were detected in each strand (≈5 per a 40-kbp T7 genome) (Rosamond et al., 1979; Taylor and Smith, 1980). No looped or tailed structures could be observed for circular dsDNA (Taylor and Smith, 1980).

These results set a basis for a model according to which RecBCD binds to a double-stranded end, unwinds the two strands, and degrades them (Figure 8) (MacKay and Linn, 1974; Roman and Kowalczykowski, 1989a; Rosamond et al., 1979). RecBCD is a helicase containing two motor subunits (Dillingham et al., 2003; Taylor and Smith, 2003). RecD and the amino-terminal region of RecB contain motifs of superfamily 1 (SF1) helicases (Gorbalenya et al., 1988, 1989). RecD translocates in 5'→3 direction (Dillingham et al., 2003) while RecB has the opposite 3'→5' polarity (Figure 8) (Boehmer and Emmerson, 1992; Phillips et al., 1997). Remarkably, though no sequence similarity was revealed, there is a structural similarity between SF1 helicases and RecC suggesting that this subunit could have evolved from an ancestral helicase (Singleton et al., 2004). The nuclease active site of RecBCD resides in the C-terminal domain of the RecB subunit (Figure 7) (Singleton et al., 2004; Wang et al., 2000; Yu et al., 1998a, 1998b).

The simultaneous work of the two helicases accounts for a high rate and processivity of RecBCD (Dillingham et al., 2005). Depending on reaction conditions, values varying in the range of several hundred - several thousand bp/s and 7-38 kbp have been reported for the DNA unwinding rate and processivity, respectively (Bianco et al., 2001; Dillingham et al., 2005; Handa et al., 2005; Korangy and Julin, 1994; Roman and Kowalczykowski, 1989b; Spies et al., 2003; Taylor and Smith, 1980, 2003). The two

RecBCD helicase subunits progress at different rates (Figure 8). It was initially reported that RecB unwinds DNA at 20-25% of the RecD rate (Korangy and Julin, 1994; Taylor and Smith, 2003). The difference in the rates results in the appearance of a loop ahead of the slower RecB subunit (Figure 8) (Braedt and Smith, 1989; Taylor and Smith, 2003). However, later it was shown that RecB and RecD motors require different Mg^{2+} and ATP concentrations for maximal speed and therefore the lead motor may switch depending on reaction conditions (Dillingham et al., 2005; Spies et al., 2005).

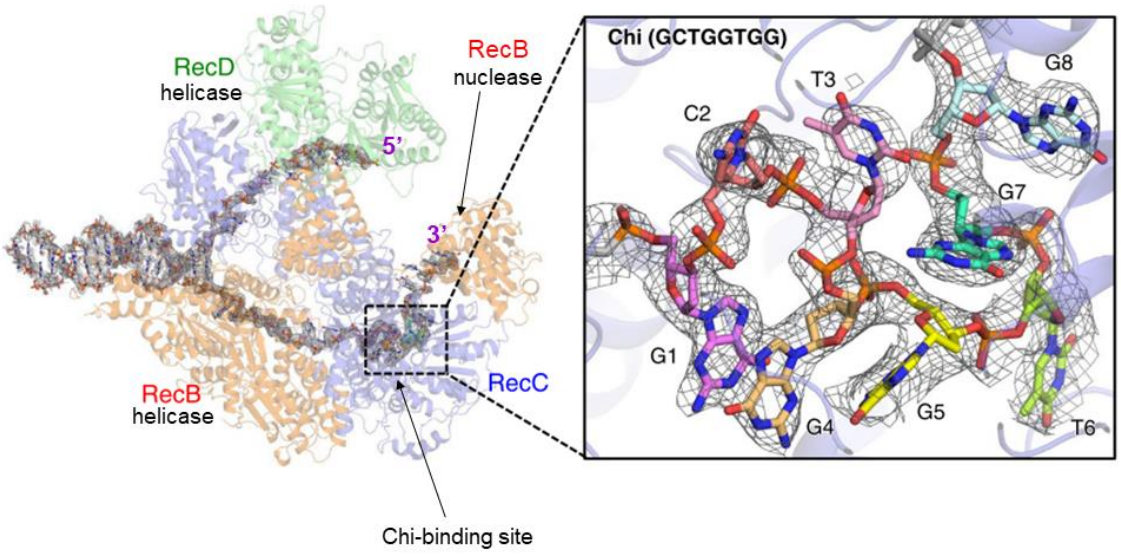


Figure 7. Structure of the RecBCD complex in the Chi-recognized state. The inset shows the density for the Chi bases. Adapted from (Cheng et al., 2020) with permission.

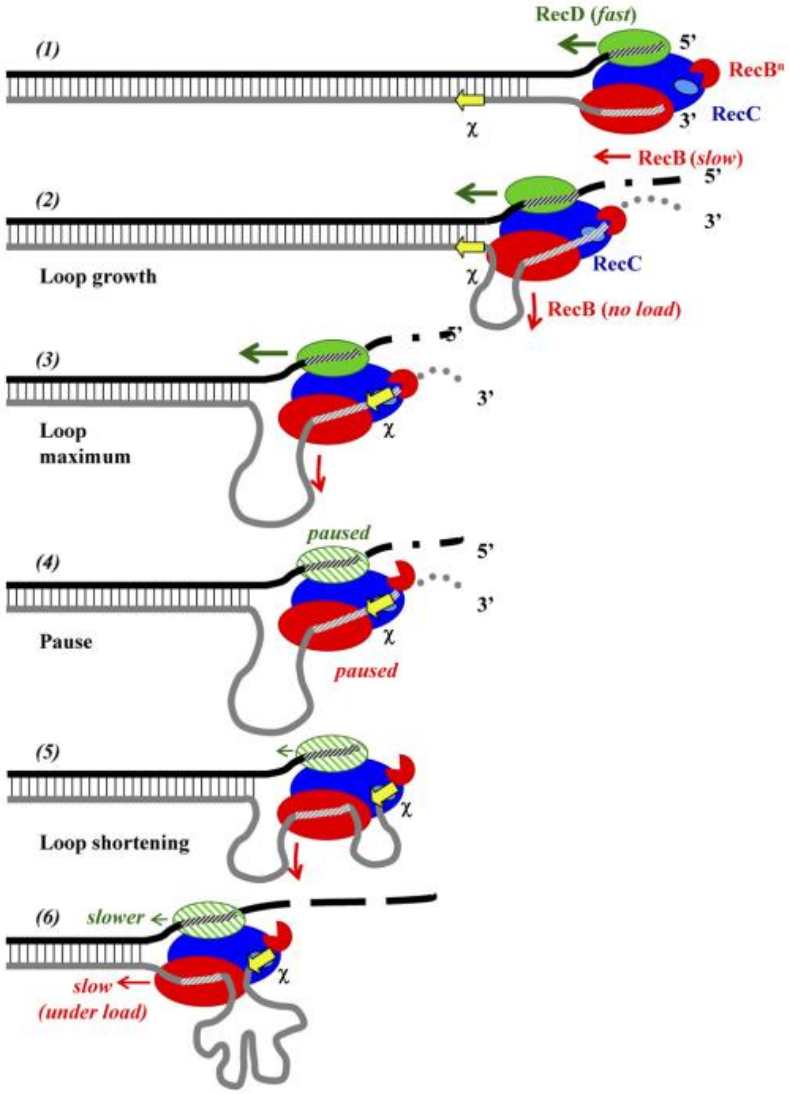


Figure 8. A model of the initial steps of DSB repair via the RecBCD-pathway. The two motors of RecBCD, RecB (red) and RecD (green), translocate along the two DNA strands at different speeds (1). A loop is generated ahead of the slower RecB motor (2). A single Chi site (yellow arrow) is recognized by the RecC subunit (blue) in the unwound 3'-terminated strand (3). The conformation of the complex is changed upon Chi recognition resulting in inhibition of RecD helicase activity (4). The Chi-terminated 3' end remains bound by the RecC subunit while RecB continues unwinding resulting in a new loop generated at the interface of the RecB helicase and RecC Chi-recognition domains (5). This new loop is further used as a substrate for RecA loading by RecB (not shown). Since RecB is the only subunit operating as a helicase upon encountering the Chi site, the loop ahead of RecB disappears (6). The RecB nuclease domain (a Packman-shaped red figure at the rear of the complex) randomly cleaves both strands once they leave the motors (1-6). Due to the greater proximity of the nuclease domain to the 3' end, the 3'-terminated strand is degraded more extensively prior to Chi recognition (2-4).
The figure is taken from (Spies et al., 2007) with permission.

In the initiation complex of RecBCD bound to the end of dsDNA, RecB and RecC subunits are tightly locked with each other (Figure 7) (Singleton et al., 2004). About 12 bp of duplex DNA lie on an 'arm' region of RecB ahead of the point where the two strands are split by a 'pin' in the RecC subunit (Singleton et al., 2004). The 3'-terminated tail heads towards the RecB helicase domain while the 5'-terminated tail passes through a channel in RecC towards the RecD motor (Figure 7, Figure 8) (Ganesan and Smith, 1993; Singleton et al., 2004). The coordinated action of the two motors drives DNA unwinding by pulling the strands across the pin (Singleton et al., 2004). As the 5'-terminated tail leaves the RecD helicase domain, it appears near the nuclease domain of RecB (Singleton et al., 2004). The 3'-terminated tail exits the RecB helicase domain, enters a channel in RecC, and then also emerges near the nuclease domain (Singleton et al., 2004). The availability of both strands for digestion explains the RecBCD dsDNA exonuclease activity.

The destructive mode of RecBCD activity needs to be attenuated to allow for the generation of single-stranded 3' tails suitable for RecA loading. The switching of RecBCD activity occurs at an 8-nt sequence 5'-GCTGGTGG-3' called a Chi site (Figure 7, Figure 8) (Dabert et al., 1992; Dixon and Kowalczykowski, 1993; Smith et al., 1981). Genetic studies demonstrated that Chi sites stimulate recombination exclusively through the RecBCD pathway (Stahl and Stahl, 1977). To be functional, a Chi site needs to be properly oriented with respect to a DSB (Kobayashi et al., 1982). RecBCD recognizes the Chi site in the 3'-terminated strand approaching the motif from the 3' side (Taylor et al., 1985). During unwinding of dsDNA, the Chi-containing 3'-terminated tail passes through the RecB helicase domain and enters the RecC subunit where the 8-nt Chi sequence is recognized (Figure 7, Figure 8) (Bianco and Kowalczykowski, 1997; Cheng et al., 2020). Binding to the Chi site makes the complex pause for a few seconds (Spies et al., 2003, 2007). A recent cryo-EM structure of the RecBCD complex revealed conformational changes at the interface of RecC and the RecB nuclease domain upon Chi binding (Cheng et al., 2020). The C-terminal domain of RecB rotates towards the Chi-recognition site bringing the nuclease active site near the 4^{th} nucleotide to the 3' side of the Chi (Figure 7)

(Cheng et al., 2020). In line with this, *in vitro* experiments revealed a Chi-dependent nick introduced by RecBCD 4-6 nucleotides to the 3' side of the motif (Ponticelli et al., 1985; Taylor et al., 1985). Slightly different results were obtained at higher Mg^{2+} concentration when the nick was shifted inside the Chi motif (Taylor and Smith, 1995b). Wherever the Chi-dependent cleavage occurs, it is a final nick introduced into the Chi-containing strand: the nuclease activity of RecBCD on the 3'-terminated strand to the 5' side of the motif is attenuated by Chi cleavage ≈500-fold (Dixon and Kowalczykowski, 1993). The degradation of the 5'-terminated strand is, on the contrary, upregulated (Figure 8) (Anderson and Kowalczykowski, 1997b). Single-molecule experiments revealed that the complex continues unwinding dsDNA past Chi but at a rate 2-fold lower than before Chi recognition (Spies et al., 2003, 2007). This change is attributed to the switch in the lead motor from the fast RecD to the slower RecB subunit (Spies et al., 2007). According to a currently accepted model, the unwinding of dsDNA after Chi accompanied by degradation of the 5' end leads to the production of a recombinogenic 3' overhang (Anderson and Kowalczykowski, 1997b). The RecB subunit facilitates the loading of RecA onto the growing 3' tail (Anderson and Kowalczykowski, 1997a). It is proposed that the Chi sequence at the end of the 3' overhang remains bound to the RecC subunit for some time resulting in a single-stranded loop formed between the RecB helicase and RecC Chi-binding domains (Figure 8) (Spies et al., 2007).

Extensive degradation of DNA up to Chi by the RecBCD complex seems counterintuitive if we think about RecBCD as a DNA repair enzyme. However, the enzyme has an additional function: it degrades linear exogenous DNA that appears in a cell upon cleavage of phage DNA by restriction endonucleases, injection of linear phage genomes, or rolling-circle replication of phage/plasmid DNA (Dabert et al., 1992; Enquist and Skalka, 1973; Kuzminov and Stahl, 1997; Kuzminov et al., 1994; Silverstein and Goldberg, 1976; Simmon and Lederberg, 1972). An expected frequency of an 8-nt Chi sequence is 1 site per ≈30 kbp of dsDNA. Hence, plasmids and small phages should have zero or few Chi sites. To protect their DNA, many phages, for example, λ and P22, encode RecBCD-inhibiting proteins (Enquist and Skalka, 1973; Karu et al., 1975;

Murphy et al., 1987; Sakaki et al., 1973). In lambdoid phage genomes encoding RecBCD inhibitors, Chi sites are underrepresented while genomes of related phages that lack the inhibitors are enriched with Chi suggesting that phages unable to block RecBCD use Chi sites for protection or, alternatively, rely on the host recombination system (Bobay et al., 2013).

In the *E. coli* genome, the Chi site is overrepresented and appears once per every ≈6 kb (Blattner et al., 1997; Bobay et al., 2013; El Karoui et al., 1999; Halpern et al., 2007). In addition, Chi distribution is not random and biased towards the strand which is replicated as leading within a given region (Blattner et al., 1997; Burland et al., 1993; Médigue et al., 1993). These observations support a model of RecBCD involvement in the reestablishment of collapsed replication forks (Figure 6B). When a replication fork encounters a nick on the template strand, RecBCD binds to the generated double-stranded end and degrades the broken chromosome arm, moving towards the origin, until it recognizes a Chi site in the 3'-terminated strand (this is the strand enriched with Chi) (Blattner et al., 1997; Kuzminov, 1995; Kuzminov et al., 1994). The invasion of a RecA-coated strand into the intact homologous arm results in reassembling of the replication fork (Figure 6B) (Kuzminov, 1995; Kuzminov et al., 1994).

Whether RecBCD degrades dsDNA before Chi recognition or only unwinds it is not important for recombination as long as the Chi is recognized and the 3'-terminated strand is nicked in its vicinity. However, the fate of dsDNA to the 3' side of Chi is of great interest for studies of prespacer generation. Levy et al. demonstrated that when a single DSB is introduced into the chromosome of cells undergoing naïve adaptation the region between the DSB and the closest appropriately oriented Chi site serves as a donor of spacers at an efficiency higher than for other genomic regions (Levy et al., 2015). The authors proposed that the fragments generated by RecBCD before the recognition of Chi are reannealed and bound by the Cas1-Cas2 complex for further processing to become prespacers. The weakness of this model is that the products of digestion by RecBCD *in vitro* are greatly affected by reaction conditions and have not been characterized *in vivo* so far.

The products of RecBCD *in vitro* cleavage contain 5'-phosphates and 3'-hydroxyls (Goldmark and Linn, 1972; Karu et al., 1973; Wright et al., 1971). The analysis of nucleotide composition of cleavage sites revealed no absolute base specificity though enrichment with purines at the 5' terminus and pyrimidines at the second nucleotide on the 3' end was shown (Goldmark and Linn, 1972; Taylor et al., 1985).

In early bulk biochemical assays under conditions stimulating exonuclease activity, native or denatured *E. coli* DNA was cleaved into acid-soluble oligonucleotides shorter than 10 nt (Goldmark and Linn, 1972; Wright et al., 1971); no mono- and dinucleotides were detected (Goldmark and Linn, 1972). Such short fragments cannot be utilized as spacer precursors since the length of a mature spacer in *E. coli* is 33 bp (Ishino et al., 1987). Subsequent experiments revealed that longer RecBCD digestion products can be obtained under different experimental conditions (increased ATP concentration and ionic strength) (Karu et al., 1973). These products contained a mixture of duplex fragments several kbp long and shorter single-stranded molecules no longer than 400 nt (median length ≈135 nt) (Karu et al., 1973; MacKay and Linn, 1974). Later works studying the RecBCD response to Chi were conducted under two distinct sets of conditions by Smith and Kowalczykowski laboratories and gave different results with respect to the degree of dsDNA degradation (Dixon and Kowalczykowski, 1993; Ponticelli et al., 1985; Taylor et al., 1985). Eggleston and Kowalczykowski demonstrated that the ratio of Mg^{2+} to ATP concentrations rather than the absolute values is critical for the RecBCD exonuclease activity (Eggleston and Kowalczykowski, 1993). ATP chelates Mg^{2+} (Wilson and Chin, 1991) and, therefore, little of free Mg^{2+} should remain in cells if ATP is in excess, resulting in decreased RecBCD nuclease activity. Side by side comparison of digestion products generated by RecBCD on a 2.3-kbp dsDNA molecule in the absence of Chi revealed that under conditions of high Mg^{2+}/ATP ratio the 3'-terminated strand is fully degraded while little if any cleavage occurs on the 5'-terminated strand (Taylor and Smith, 1995b). Under conditions of low Mg^{2+}/ATP ratio, both strands mostly remain intact (Taylor and Smith, 1995b). Overall, these observations demonstrate that the RecBCD nuclease activity is highly susceptible to reaction

conditions. Given that the concentrations of free Mg^{2+} and ATP in cytoplasm remain uncertain, it is not clear if RecBCD indeed produces short fragments longer than 33 bp that can be used as spacer precursors during naïve adaptation. Alternatively, it can be assumed that RecBCD-mediated DNA unwinding increases the efficiency of PAM recognition by Cas1-Cas2. Another possibility is that fragments generated due to secondary cleavage reactions by other nucleases on long DNA fragments produced by RecBCD serve as spacer precursors.

1.3.4.3 RecFOR pathway of homologous recombination

RecBCD is the main enzyme performing end resection due to its high affinity to dsDNA ends, high speed, and processivity. However, in the absence of the RecBCD nuclease activity, multiple nucleases get involved in end processing. Their interaction is complex with some enzymes facilitating or interfering with recombination depending on the genetic context.

RecFOR is an alternative pathway of homologous recombination (Horii and Clark, 1973; Kushner et al., 1971). The 3'-terminated single-strand tail in this pathway is produced due to the activity of the 5'→3' exonuclease RecJ (Figure 9) (Handa et al., 2009; Lovett and Clark, 1984; Lovett and Kolodner, 1989). The RecJ exonuclease degrades single-stranded DNA to mononucleotides (Han et al., 2006). It is active on ss- but not dsDNA and requires a single-stranded tail of at least 7 nt for binding (Han et al., 2006; Lovett and Kolodner, 1989). The RecJ nuclease activity is enhanced if ssDNA is covered by SSB (Han et al., 2006). These observations highlight the need for a helicase unwinding DNA from a double-stranded end prior to RecJ loading. The partner helicase is the RecQ protein, which binds to a blunt end or an end with a 3' overhang and unwinds the duplex moving along the 3'-terminated strand (Figure 9) (Morimatsu and Kowalczykowski, 2014; Umezu et al., 1990). Therefore, RecQ is a functional analog of RecBCD (after Chi recognition) but it moves at a slower rate of ≈1-84 bp/s (Harmon and Kowalczykowski, 2001; Zhang et al., 2006). The loading of RecA is facilitated either by the RecFOR complex at the ss/dsDNA junction or by the RecOR complex at the SSB-coated single-stranded tail (Figure 9) (Bell et al., 2012; Handa et al., 2009; Morimatsu

and Kowalczykowski, 2003; Sakai and Cox, 2009; Umezu and Kolodner, 1994; Umezu et al., 1993).

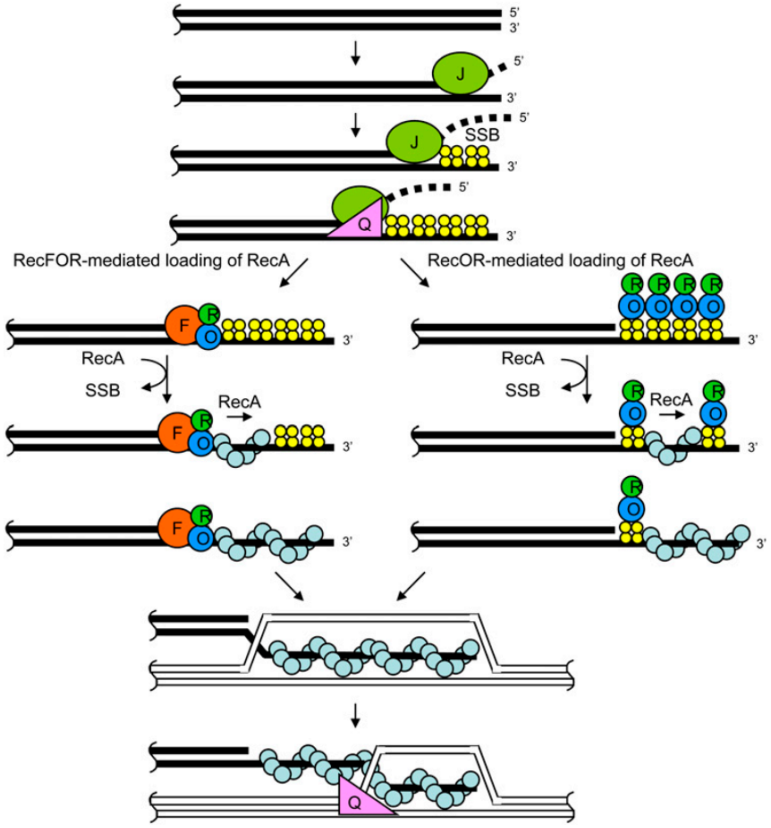


Figure 9. RecFOR and RecOR pathways of homologous recombination. The 3'-terminated single-stranded tail is generated due to the activities of the RecQ helicase and RecJ 5'→3' exonuclease. The RecFOR or RecOR complexes facilitate the loading of RecA onto the generated overhang. The picture is taken without changes from a paper of Handa et al., 2009 distributed under a Creative Commons License (no permission is required) (Handa et al., 2009).

Though the RecFOR pathway is functional in wild-type cells and normally participates in post-replication repair (Ganesan and Seawell, 1975; Tseng et al., 1994), the survival after UV-exposure and recombination drop by at least two orders of magnitude upon inactivation of the RecBCD-pathway suggesting that the contribution of the RecFOR pathway to DSB repair is minor (Barbour et al., 1970; Kushner et al., 1971; Repar et al., 2013; Willetts and Mount, 1969). Additional mutations restoring the wild-type level of recombination and viability were obtained in *recB*$^{-}$ and *recC*$^{-}$ strains and were named *sbc* ("suppressors of *rec* *BC*") (Barbour et al., 1970; Gibson et al., 1992;

Kushner et al., 1971; Lloyd and Buckman, 1985). Simultaneous inactivation of two nucleases, ExoI (encoded by the *sbcB* gene also known as *xonA*) and SbcCD (encoded by the *sbcC* and *sbcD* genes), fully restores the ability of *recB*$^{-}$*recC*$^{-}$ cells to recombine via the RecFOR pathway (Gibson et al., 1992; Horii and Clark, 1973; Kushner et al., 1971; Lloyd and Buckman, 1985). It was suggested that the exonuclease activities of ExoI and SbcCD might interfere with the stability of recombinogenic 3' ends (Connelly and Leach, 1996; Kushner et al., 1971).

ExoI is an exonuclease digesting ssDNA in the 3'→5' direction to mononucleotides (Lehman, 1960; Lehman and Nussbaum, 1964). SbcCD is an ATP-dependent dsDNA 3'→5' exonuclease and an ATP-independent ssDNA endonuclease (Connelly and Leach, 1996; Connelly et al., 1997, 1999). It cleaves hairpin structures 5' of the loop and removes 5' overhangs adjacent to duplex DNA if the length of the single-stranded region is at least 10 nt (Connelly et al., 1998, 1999). *In vitro* the SbcCD complex also cleaves DNA close to a biotinylated end bound by avidin, suggesting that one of SbcCD functions *in vivo* may be the removal of proteins blocking DNA ends (Connelly et al., 2003). While initial experiments were conducted on relatively short DNA molecules, unexpected SbcCD activity was demonstrated for duplex molecules ≥ 40 bp with a free end or a hairpin (Lim et al., 2015). Instead of exonucleolytic degradation from the 3' end, DSBs at distances that are multiples of 10-11 bp from the ends were observed (Lim et al., 2015). A new term, 'DNA end-dependent binary endonuclease', was coined for this type of activity (Lim et al., 2015). The authors proposed a model according to which a single complex senses a double-stranded end or a hairpin and stimulates the assembly of a filament composed of multiple complexes arranged on the duplex, some of which cleave DNA (Lim et al., 2015). Importantly, electron microscopy revealed that the SbcCD complex consists of two globular domains separated by a filamentous rod (Connelly et al., 1998). The diameter of the globular domain is ≈3.4 nm, which corresponds to ≈10.5 bp of B-form dsDNA supporting the hypothesis of Lim et al. (Connelly et al., 1998; Lim et al., 2015). The exact mechanism of DNA end-dependent binary endonuclease activity remains unknown.

Given that RecJ and ExoI digest DNA to mononucleotides it is unlikely that these enzymes produce substrates for adaptation. However, partial cleavage of DNA by the SbcCD binary endonuclease activity can potentially result in the generation of dsDNA fragments of 30-40 bp suitable for adaptation.

1.3.4.4 DSB repair in ΔrecD mutant cells

In *recD* null mutant cells, the RecBC complex is formed which is devoid of the exonuclease but retains the helicase activity, which is, however, approximately 4-5-times slower than that of the wild-type enzyme (Biek and Cohen, 1986; Masterson et al., 1992; Palas and Kushner, 1990; Rinken et al., 1992). Despite the differences in the activities of RecBCD and RecBC complexes, *recD* mutant cells are fully viable and proficient in recombination (Biek and Cohen, 1986; Chaudhury and Smith, 1984; Lloyd et al., 1988; Lovett et al., 1988). This phenotype is due to the ability of the complex to unwind dsDNA and constitutively load RecA onto the 3' end regardless of Chi sites (Churchill et al., 1999). Genetic analysis revealed a dramatic increase in UV-sensitivity and decreased frequency of recombination of a double *recD recJ* mutant suggesting that RecJ can substitute the missing RecBCD activities (Lloyd et al., 1988; Lovett et al., 1988). These results suggest that a hybrid pathway of homologous recombination operates in *ΔrecD* cells where the 3'-terminated single-stranded tail is produced due to the 5'→3' exonuclease activity of RecJ while DNA unwinding and RecA loading is facilitated by RecBC (Amundsen and Smith, 2003).

Additional mutations introduced into *ΔrecD* cells affect recombination and survival after treatment with DNA-damaging factors to a different extent. The most pronounced impairment is observed in *ΔrecD* cells lacking RecJ and ExoVII. The viability of these cells after γ-irradiation is at least four orders of magnitude lower than that of the *ΔrecD* cells and the frequency of recombination drops 100-1000-fold (Dermić, 2006; Dermić et al., 2006). ExoVII is a single-strand specific exonuclease degrading ssDNA in 5'→3' and 3'→5' directions to oligonucleotides of ≈2-25 nt with the majority of the products being in the range of 4-10 nt (Chase and Richardson, 1974a, 1974b). Inactivation of ExoVII does not affect cell viability and recombination in *ΔrecD* cells,

while the deletion of *recJ* decreases the survival rate after γ-irradiation ≈5-fold and decreases recombination frequency measured in conjugation experiments ≈10-fold (Dermić, 2006; Dermić et al., 2006). These results demonstrate that though ExoVII can partially substitute RecJ in *ΔrecD* cells, its contribution is minor and the resection of 5'-terminated strands mostly relies on RecJ. Surprisingly, the number of transconjugants and the survival of γ-irradiated cells is dramatically decreased in *ΔrecD ΔrecJ ΔsbcB* mutant (≈3-orders of magnitude decrease compared with the *ΔrecD ΔrecJ* mutant) suggesting that RecJ and ExoI are synergistic in *recD* cells (Dermić, 2006). A milder but still reproducible effect is observed in transduction with P1 or λ phages where recombination drops ≈7-fold compared with the double mutant *ΔrecD ΔrecJ* (Dermić, 2006; Dermić et al., 2006). ExoI and RecJ have opposite polarities and the mechanism of ExoI involvement in recombination remains unknown. The involvement of ExoI and RecJ into degradation of linear DNA was also demonstrated when *ΔrecD* cells were infected with T4 gene *2* mutant phage, which lacks a protein protecting DNA ends from degradation (Oliver and Goldberg, 1977; Rinken et al., 1992).

As was noted before, RecJ and ExoI are unlikely to produce substrates for adaptation. The contribution of ExoVII to cell viability and recombination in *ΔrecD* cells is minor, therefore little if any fragments should be produced by this enzyme during DSB repair in *ΔrecD* cells. In agreement with these expectations, naïve adaptation in *ΔrecD* is strongly decreased (Radovcic et al., 2018). Surprisingly, *recA* mutation restored adaptation in *recD* cells suggesting that the RecBC helicase activity is important for spacer acquisition in this genetic context (Radovcic et al., 2018). The mechanism is not clear. The authors demonstrated that adaptation in *recD recA* cells was dependent on various nucleases including ExoI and SbcCD but the plasmid encoding *cas1* and *cas2* was unstable in these mutants and therefore it is difficult to draw conclusions since the level of *cas1* and *cas2* expression could have been different (Radovcic et al., 2018).

1.3.4.5 RecET pathway of homologous recombination

In addition to RecBCD and RecFOR pathways of homologous recombination, there is a RecET pathway. The main players of the RecET pathway – RecE (ExoVIII)

and RecT proteins - are encoded by the Rac prophage (Kaiser and Murray, 1979; Low, 1973). Similarly to the RecFOR pathway, the RecET pathway was discovered when suppressor mutations called *sbcA* were obtained in *recB*$^-$ and *recC*$^-$ cells (Barbour et al., 1970). The *sbcA* mutations activate expression of the *recE* gene encoding an ATP-independent nuclease ExoVIII, which is not produced in wild-type cells (Barbour et al., 1970; Kushner et al., 1974).

RecE (ExoVIII) is a 5'→3' exonuclease degrading one strand of dsDNA to mononucleotides (Joseph and Kolodner, 1983a, 1983b). RecT promotes the annealing of complementary single-stranded DNA (Hall et al., 1993). Together, RecE and RecT promote strand exchange reactions in a RecA-independent manner (Hall and Kolodner, 1994). However, in some cases, hybrid pathways with the involvement of RecFOR, RecA, RecJ and RecQ were described (Gillen et al., 1981; Lovett and Clark, 1984; Luisi-DeLuca et al., 1989; Mahdi and Lloyd, 1989).

Taken together, the available data on the activity of various DSB repair nucleases suggest that RecBCD and SbcCD nucleases are the likely candidates to produce fragments for naïve adaptation. However, the products of their activity *in vivo* are yet to be characterized. Exonucleases like SbcB and RecJ digesting DNA to mononucleotides can not produce substrates suitable for adaptation and may, in fact, reduce the number of fragments available for binding by Cas1-Cas2. However, it can be also hypothesized that some of the mentioned exonuclease activities are involved in trimming the ends of prespacers after Cas1-Cas2 binds to fragments longer than 33 bp.

1.3.5 Possible mechanisms of prespacer generation during primed adaptation

Primed adaptation requires all components of the type I-E CRISPR-Cas system: Cas1, Cas2, Cascade, Cas3, and crRNA targeting a PPS available in a cell (Datsenko et al., 2012).

A low percentage of CRISPR arrays are extended via primed adaptation if the PPS has a proper PAM sequence and there is a full correspondence between a spacer and the protospacer (Xue et al., 2015). A single or a few mutations introduced into the PPS or the adjacent PAM sequence increase spacer yield (Datsenko et al., 2012; Fineran et al.,

2014). This phenomenon is interpreted as a backup mechanism allowing cells that have lost the ability to interfere with phage infection to restore the resistance by acquiring new interference-proficient spacers (Datsenko et al., 2012; Fineran et al., 2014).

There are several characteristics of primed adaptation in type I-E systems.

1) Approximately 95% of spacers acquired by primed adaptation target protospacers flanked by the 5'-AAG-3' PAM (Datsenko et al., 2012; Savitskaya et al., 2013).
2) More than 90% of spacers originate from regions flanking the PPS (Datsenko et al., 2012). If a plasmid or small circular replicon (for example, M13 phage DNA) is used as a target, prespacers can be selected from any part of the target DNA without apparent preference for protospacers directly adjacent to the PPS, although some spacers are integrated with higher efficiency than the others regardless of their position (Datsenko et al., 2012; Savitskaya et al., 2013; Swarts et al., 2012). If a long DNA molecule is used as a target (for example, λ or T5 phage genomes), it becomes evident that protospacers are selected with lower efficiency as the distance from the PPS increases (Strotskaya et al., 2017). Two gradients of protospacer selection efficiency, upstream and downstream of the PPS, are observed (Strotskaya et al., 2017). The total number of spacers originating from the upstream region is higher than from the downstream region reflecting the preferential movement of Cas3 in 3'→5' direction along the NT-strand upstream of the PPS (Figure 2B, Figure 10) (Mulepati and Bailey, 2013; Strotskaya et al., 2017).
3) The orientation of protospacers is markedly different for the two regions (Figure 10) (Strotskaya et al., 2017). If we map the non-transcribed 'top' strand of spacers in the CRISPR array to spacer source DNA, most matches in the upstream region will be located on the NT-strand (we will denote these protospacers as PS^{NT} and the corresponding spacers – Sp^{NT}) while most matches downstream of the PPS will be located on the T-strand (we will denote these protospacers as PS^{T} and the corresponding spacers – Sp^{T}) (Strotskaya et al., 2017). Early studies of the type I-E primed adaptation, where small circular PPS-containing replicons were used, did not reveal bidirectional spacer acquisition with inversed protospacer gradients and

reported that most of the acquired spacers match the NT-strand (Sp^{NT}) (Datsenko et al., 2012; Fineran et al., 2014; Savitskaya et al., 2013; Swarts et al., 2012). This is probably due to the higher efficiency of degradation/adaptation upstream of the PPS combined with the overlap of two gradients on small circular replicons (Mulepati and Bailey, 2013; Sinkunas et al., 2013; Strotskaya et al., 2017).

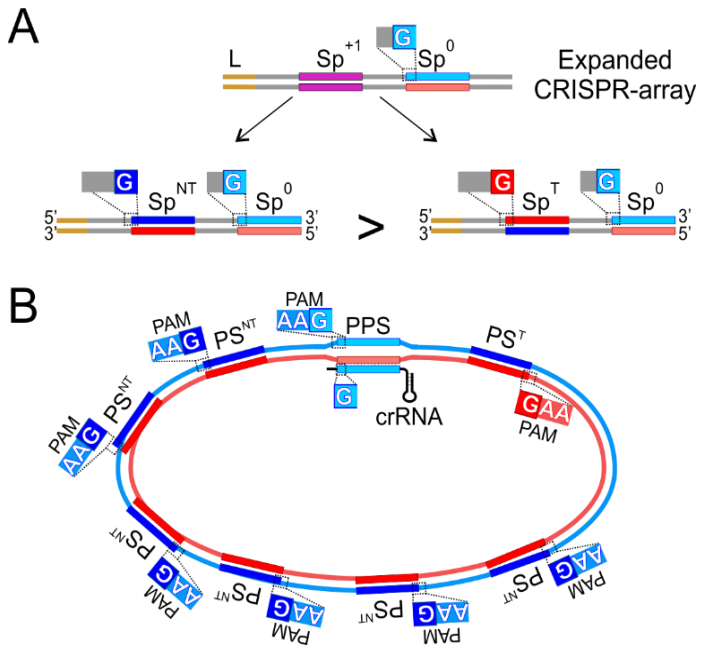


Figure 10. Spacers acquired during primed adaptation and their corresponding protospacers. **A.** Two orientations (Sp^{NT} and Sp^{T}) with respect to the strand targeted by newly acquired spacers (Sp^{+1}) incorporated into the CRISPR array during primed adaptation initiated by Sp^{0}-containing crRNA in the type I-E CRISPR-Cas system. **B.** Positions and orientations of protospacers (PS^{NT} and PS^{T}) selected during primed adaptation initiated at the PPS.

The mechanism of primed adaptation is not understood. One model proposed that the Cas3 nuclease/helicase cleaves target DNA into fragments used by Cas1-Cas2 as spacer precursors (Swarts et al., 2012). In support of this model, plasmid-derived products of Cas3-mediated cleavage were bound by Cas1-Cas2 and integrated into the CRISPR array *in vitro* (Künne et al., 2016). The authors demonstrated a difference in lengths of the fragments generated by Cas3 on the opposite strands: the NT-strand was digested into fragments 60-100 nt long and the T-strand was cleaved into shorter 30-70-nt fragments (Künne et al., 2016). Most cleavage products on both strands had T as the last nucleotide on their 3' ends (Künne et al., 2016). It led the authors to propose that Cas3

products were enriched with 3'-TTC-5' on their 3' termini and therefore formed better substrates for Cas1-Cas2 (Künne et al., 2016). However, this model does not explain the bias observed for the orientation of protospacers *in vivo* and it is not clear if the reported difference in product lengths has something to do with this phenomenon.

Another model suggests that after binding of the Cascade-crRNA to the PPS, Cas3 and Cas1-Cas2 form a larger complex that slides along DNA searching for and excising prespacers (Datsenko et al., 2012; Dillard et al., 2018; Redding et al., 2015). The main function of Cas3 in this model is to deliver the Cas1-Cas2 complex to remote protospacers. Single-molecule experiments demonstrated that in the absence of Cas1-Cas2, Cas3 stays bound to Cascade at least for some time (Dillard et al., 2018; Loeff et al., 2018; Redding et al., 2015). In turn, Cascade remains tightly bound to the PPS leading to the generation of a DNA loop between Cas3 and Cascade while Cas3 translocates (Dillard et al., 2018; Loeff et al., 2018; Redding et al., 2015). In *Thermobifida fusca* type I-E system Cas3 remained bound to Cascade in ≈50% of cases (Dillard et al., 2018). When a labeled Cas1-Cas2 complex was added into the system, a primed acquisition complex (PAC) composed of Cascade, Cas3, and Cas1-Cas2 was formed (Dillard et al., 2018). The addition of Cas1-Cas2 stabilized Cas3-Cascade interaction and no rupture between Cas3 and Cascade was observed (Dillard et al., 2018). Cas1-Cas2 stayed bound to Cas3/Cascade in ≈90% of cases (Dillard et al., 2018). The formation of the PAC in *T. fusca* containing target DNA was confirmed *in vivo* using BiFC assay (Dillard et al., 2018).

The model suggesting the assembly of the PAC is supported by the fact that in a closely related type I-F CRISPR-Cas system Cas3 is fused to Cas2 and the hybrid Cas2/Cas3 protein forms a complex with Cas1 (Fagerlund et al., 2017; Richter et al., 2012a; Rollins et al., 2017). A target-bound Csy complex (type I-F Cascade) recruits a standalone Cas2/3 protein or the complex of Cas1 with Cas2/3 (Rollins et al., 2017). Similarly to the type I-E, an interference-deficient spacer stimulates primed adaptation in the type I-F (Richter et al., 2014; Vorontsova et al., 2015). Prespacers are selected from the region up- and downstream of the PPS but the strand bias is opposite to what is seen

in the type I-E and there is a different PAM sequence – 5'-GG-3' (Richter et al., 2014; Vorontsova et al., 2015). It is not understood how the strand biases are formed. Assuming that the PAC is formed in both systems, one can speculate that the architectures of the two complexes are different and the PAMs are recognized by the type I-E and I-F Cas1 proteins in the opposite strands.

Whichever model is true, it remains unknown if prespacers generated during priming have any specific structure. It is also not clear which nuclease activities are involved in prespacer excision. The controversy about the cleavage of the prespacer 3' ends by the type I-E Cas1 has already been discussed in section 1.3.2 of this review. The purified type I-F Cas1-Cas2/3 complex trims the 3' ends next to the 5'-GG-3' PAM sequence and integrates them into a CRISPR-containing plasmid (Fagerlund et al., 2017). No data on *in vivo* prespacer trimming and integration exist for the type I-F system.

It was reported that the Cas3 nuclease activity is attenuated by Cas1 or Cas1-Cas2 in the type I-F and I-E, respectively (Dillard et al., 2018; Redding et al., 2015; Rollins et al., 2017). However, increased cleavage at the sites of stalling was reported for the Cas3/Cascade complexes encountering DNA-bound proteins (Dillard et al., 2018). Following this line of thought, it could be speculated that stalling at protospacers bound by Cas1-Cas2 within the PAC might also increase the probability of cleaving DNA by Cas3 in the vicinity of the protospacer.

Since primed adaptation is coupled to interference, DNA cleaved by Cas3 may serve as an entry point for host DNA repair enzymes, which may participate in prespacer trimming. Primed adaptation was examined by PCR in a set of single deletion mutants of DSB repair genes including *recB, recJ,* and *sbcD* (Ivančić-Baće et al., 2015). Neither mutation fully abolished spacer acquisition but it was not ruled out that some mutations led to a decrease in adaptation efficiency (Ivančić-Baće et al., 2015). Given that DSB repair enzymes are redundant, it is plausible that single deletions might not be enough to reveal the dependence of primed adaptation on host nucleases even if this dependence exists.

Chapter 2. Project Objectives

The progress in unraveling the mechanisms of CRISPR adaptation *in vivo* is mostly due to studies that address spacer acquisition efficiency, analyze the source of spacers, their lengths, and nucleotide determinants of efficient acquisition. These experiments deal with spacers incorporated into CRISPR arrays, which can be amplified by standard PCR with primers annealing to the leader sequence and a pre-existing spacer (Shiriaeva et al., 2020). Unlike the acquired spacers that remain steady within CRISPR arrays and accumulate over time, spacer precursors are transient and therefore difficult to detect. No method of high-throughput analysis of prespacers generated *in vivo* has been proposed so far. Some suggestions about prespacer structure can be made based on *in vitro* studies of Cas1-Cas2 binding to oligonucleotides and their integration into the CRISPR array (Moch et al., 2017; Nuñez et al., 2015b, 2015a; Wang et al., 2015). *In vivo* experiments on electroporation of spacer-size oligonucleotides into *E. coli* cells expressing type I-E *cas1* and *cas2* demonstrated that only double-stranded prespacers are functional, and the presence of the PAM sequence is essential for proper integration (Shipman et al., 2016). However, these experiments do not provide details about the length of each strand of prespacers and the position of the PAM within it.

Using primer extension assay, Musharova et al. detected two cuts in genomic DNA at the PAM-distal and PAM-proximal boundaries of two protospacers frequently used as spacer donors during primed adaptation (Musharova et al., 2017). However, the primer extension products were detected only for one protospacer strand containing the 5’-AAG-3’ PAM sequence and the fate of the TTC-containing strand remains unknown.

The mechanism of transition from CRISPR interference to primed adaptation is not determined and it is hypothesized that Cas3 may generate raw material for the production of prespacers (Künne et al., 2016; Swarts et al., 2012). *In vivo* products of CRISPR interference have not been characterized. Recent data suggest that host nucleases may be involved in prespacer trimming (Drabavicius et al., 2018; Kim et al., 2020; Ramachandran et al., 2020; Yoganand et al., 2019). In our laboratory, Elena

Kurilovich showed that RecBCD widens the gap produced by Cas3 during CRISPR interference (unpublished data). However, it is not known if any fragments fueling adaptation are produced during this process.

Given the apparent lack of understanding of prespacer generation and maturation processes *in vivo*, we set a goal to characterize interference and primed adaptation intermediates formed *in vivo* in type I CRISPR-Cas systems.

The following objectives were pursued.

1. To develop a protocol for purification of short DNA fragments generated *in vivo* and their high-throughput analysis.
2. To determine if spacer precursors or CRISPR interference intermediates can be detected using the developed approach under conditions of CRISPR interference and primed adaptation by the type I-E system of *Escherichia coli*.
3. To evaluate the impact of the interference and adaptation modules on the production of the detected fragments.
4. To assess the impact of host DNA repair nucleases on the generation of spacer precursors and CRISPR interference intermediates.
5. To characterize CRISPR interference and primed adaptation intermediates produced by the type I-F system and compare them with the intermediates detected in the type I-E system.

Chapter 3. Materials and Methods

3.1 Bacterial strains and plasmids

The *E. coli* strains used in this study are listed in Appendices, Table 1. The red recombinase-mediated gene-replacement technique was used to obtain strains KD403, KD518, and KD753 (Datsenko and Wanner, 2000). Self-targeting strains with deletions of *recB*, *recC*, *recD*, *sbcB*, *sbcD*, or *recJ* genes were obtained using P1 transduction (Moore, 2011). KD403 was used as a recipient. Strains with single deletions of *recB*, *recC*, *recD*, *sbcB*, *sbcD*, and *recJ* genes from the Keio collection were used as donor strains (Baba et al., 2006).

A plasmid pCas1+2 for expression of type I-E *cas1* and *cas2* genes as well as plasmids pCas and pCsy for expression type I-F *cas* and *csy* genes were described in (Vorontsova et al., 2015; Yosef et al., 2012).

3.2 Growth conditions

For analysis of CRISPR-mediated self-targeting by the type I-E system, an overnight culture of the KD403 strain grown at 37°C in LB medium was diluted 100-fold into 10 ml fresh LB and incubated at 37°C until OD_{600} reached 0.3. The culture was divided into two portions, *cas* genes inducers, IPTG and L-(+)-arabinose, were added at 1 mM concentration to one portion, and cultures with and without inducers were incubated at 37°C for 7 hr. At various time points postinduction, cells were plated with serial dilutions on 1.5% LB agar plates for counting colony-forming units (CFU) or were monitored using fluorescent microscopy.

In assays using strains KD403, KD518, KD753, KD263, and KD403 derivatives with single or double deletions of DNA repair genes that were followed by sequencing of total genomic DNA, short DNA fragments or newly acquired spacers, similar conditions of culture growth and *cas* genes induction were applied, except that cultures were grown at 30°C. Five hours postinduction, 10 ml of cells were pelleted by centrifugation at 3000 x *g* for 5 min at 4°C, washed with 10 ml of PBS, pelleted by centrifugation at 3000 x *g*

for 5 min at 4°C and resuspended in 1 ml of PBS. Cells were divided into 125-µl aliquots and stored at -70°C before they were used for DNA isolation.

For analysis of short DNA fragments generated during self-targeting by the type I-F system, cultures of strain KD675 transformed with plasmids pCas and pCsy were grown at 37°C in LB supplemented with 100 µg/ml ampicillin and 50 µg/ml spectinomycin. Overnight cultures were diluted 200-fold into 10 ml of LB without antibiotics, grown at 37°C until OD_{600} reached 0.3 and supplemented with 1mM IPTG and 1mM L-(+)-arabinose. Cells were harvested 24 hr postinduction and prepared for DNA isolation as described above for strains KD403, KD518, KD753, KD263, and DNA repair mutant derivative of KD403.

3.3 Fluorescence microscopy

Cultures grown with or without induction of *cas* gene expression were analyzed using a LIVE/DEAD viability kit (Thermo Scientific) at 5 hr after induction. Viable cells in each culture were detected by the addition of 20 µM SYTO9, a green fluorescent dye that can penetrate through intact cell membranes. Non-viable cells in each culture were detected by the addition of 20 µM propidium iodide dye, which cannot enter viable cells. Sample chambers were made using a microscope slide (Menzel-Gläser) with two strips on the upper and lower edges formed by double-sided sticky tape (Scotch TM). To obtain a flat substrate required for high-quality visualization of bacteria, a 1.5% agarose solution was placed between tape strips and covered with another microscopic slide. After solidification of the agarose, the upper slide was removed and several agarose pads were formed. 1 µl of each cell suspension (with and without induction) was placed on an agarose pad. The microscopic chamber was sealed using coverslip (24 x 24 mm, Menzel-Gläser).

Fluorescence microscopy was performed using Zeiss AxioImager.Z1 upright microscope. Fluorescence signals in green (living cells) and red (dead cells) fluorescent channels were detected using Zeiss Filter Set 10 and Semrock mCherry-40LP filter set respectively. Fluorescent images of self-targeting cells were obtained using Cascade II:1024 back-illuminated EMCCD camera (Photometrics). The microscope was

controlled using AxioVision Microscopy Software (Zeiss). All image analysis was performed using ImageJ (Fiji) with ObjectJ plugin used for measurements of cell length (Vischer et al., 2015).

3.4 High-throughput sequencing of total genomic DNA

Total genomic DNA was purified by GeneJET Genomic DNA Purification Kit (ThermoFisher Scientifc). Sequencing libraries were prepared either by NEBNext® Ultra™ II DNA Library Prep Kit for Illumina (NEB) or by Accel-NGS® 1S Plus DNA Library Kit (Swift Biosciences) and sequenced on a NextSeq 500 platform.

Raw reads were analyzed in R with ShortRead and Biostrings packages (Morgan et al., 2009). Reads with no more than 2 bases with quality < 20 were mapped to the KD403 reference genome using Unipro UGENE platform (Okonechnikov et al., 2012). Bowtie2 was used as a tool for alignment with end-to-end alignment mode and 1 mismatch allowed (Langmead and Salzberg, 2012). The BAM files were analyzed by Rsamtools package and reads with the MAPQ score equal to 42 were selected and used for downstream coverage analysis (Li et al., 2009). Mean coverage over non-overlapping 1-kb or 10-kb bins was calculated and normalized to the total coverage (the sum of means).

3.5 High-throughput sequencing of newly acquired spacers

Cell lysates were prepared by resuspending cells in water and heating at 95°C for 5 min. Cell debris was removed from lysates by centrifugation at 16 x g for 1 min. For analysis of spacer acquisition in strains KD263 and KD403 lysates were used in PCR reactions containing primers LDR-F2 (ATGCTTTAAGAACAAATGTATACTTTTAG) and Ec_minR (CGAAGGCGTCTTGATGGGTTTG) (25 cycles, T_a=52°C) (Appendices, Table 2). Reaction products were separated by agarose gel electrophoresis (Figure 14B). To obtain amplicons derived from extended CRISPR arrays in KD403 and its derivatives, PCR reactions were performed using primers LDR-F2 (ATGCTTTAAGAACAAATGTATACTTTTAG) and autoSp2_R (AATAGCGAACAACAAGGTCGGTTG) (30 cycles, T_a=52°C) (Appendices, Table 2).

Reaction products were separated by agarose gel electrophoresis and the amplicon derived from the extended array was purified from the gel using a GeneJET Extraction Kit (ThermoFisher Scientifc) and sequenced on a NextSeq 500 system.

Bioinformatic analysis was performed in R using ShortRead and Biostrings packages (Morgan et al., 2009). Spacer sequences were extracted from the reads containing two or more CRISPR repeats. Spacers of length 33 bp were mapped to the KD403 genome to identify 33-bp protospacer sequences with 0 mismatches allowed. Spacers that aligned to a single position in the chromosome were used to determine protospacer distribution along the genome. Spacers arising from protospacers due to potential slippage or flippage were removed from the analysis (Shmakov et al., 2014).

3.6 Prespacer efficiency assay

Prespacer efficiency assay was performed according to the following protocol (Shipman et al., 2016). An overnight culture of BL21-AI cells containing a plasmid pCas1+2 was diluted 30-fold into 9 ml LB supplemented with 50 μg/ml streptomycin, 13 mM L-(+)-arabinose, and 1 mM IPTG and grown at 37°C for 2 hours. Cells were harvested by centrifugation at +4°C (1 ml of cells per 1 transformation), washed twice with cold water, and resuspended in 50 μl of a solution containing 3.125 μM complementary oligonucleotides (Appendices, Table 3). Electroporation was carried out in 1 mm gap cuvette at a voltage of 1.8 kV. 3 ml of LB supplemented with 50 μg/ml streptomycin were added to electroporated cells and the cultures were incubated at 37°C for 2 hours. Lysates of cell cultures were prepared and used in PCR reactions containing a primer BLCRdir complementary to the leader sequence (GGTAGATTGTGACTGGCTTAAAAAATC) and a primer BLCRreverse complementary to the preexisting spacer in the array (GTTTGAGCGATGATATTTGTGCTC), respectively (Appendices, Table 2). Amplicons corresponding to extended and nonextended CRISPR arrays were isolated using GeneJET PCR Purification Kit (ThermoFisher Scientifc) and sequenced on a NextSeq 500 platform. Bioinformatic analysis was performed in R using ShortRead and Biostrings packages (Morgan et al., 2009). Reads containing the bases with Phred quality

< 14 were removed from the analysis and reads containing at least one CRISPR repeat were further analyzed. Newly acquired spacers were extracted from the expanded reads and mapped to the genome, plasmid, and transforming oligonucleotide sequences with 2 mismatches allowed. 33 bp oligo-derived spacers that were cut between AA and G before integration were considered as properly processed. For simplicity only properly processed oligo-derived spacers inserted into the CRISPR array in a direct (GCCCAATTTACTACTCGTTCTGGTGTTTCTCGT) or reverse (ACGAGAAACACCAGAACGAGTAGTAAATTGGGC) orientation were included in the analysis.

3.7 Isolation of DNA fragments generated *in vivo*

Total genomic DNA was isolated from cultures of strains KD403, KD518, KD753, KD263, KD403 *ΔrecB*, KD403 *ΔrecC*, KD403 *ΔrecD*, KD403 *ΔrecJ*, KD403 *ΔsbcB*, KD403 *ΔsbcD*, KD403 *ΔrecB ΔrecJ*, KD403 *ΔrecB ΔsbcB*, KD403 *ΔrecB ΔsbcD*, and KD675 by collecting 1.25 ml of cell suspensions by centrifugation, resuspending cells in 125 µl of PBS, adding 2 ml lysis buffer (0.6% SDS, 12 µg/ml proteinase K in 1x TE buffer) and incubating at 55°C for 1 hr. Two milliliters of phenol:chloroform:isoamyl alcohol (25:24:1) (pH 8) were added to the lysate, the solution was gently mixed, and the aqueous and organic phases separated by centrifugation at 7000 x *g* for 10 min at room temperature. The upper aqueous phase containing total genomic DNA was collected and the residual phenol was removed by the addition of 2 ml of chloroform:isoamyl alcohol (24:1). The solution was gently mixed, centrifuged at 7000 x *g* for 10 min at room temperature, the upper DNA-containing fraction was transferred into a fresh tube, 0.2 M NaCl, 15 µg/ml of Glycoblue (Invitrogen), and two volumes of cold 100% ethanol were added, and the solution was incubated at -80°C overnight. Precipitated DNA was recovered by centrifugation at 21000 x *g* for 30 min at 4°C. Pellets were washed twice with 80% ethanol, resuspended in 200 µl of 1x TE buffer, and treated with 1 mg/ml RNase A at 37°C for 30 min to remove residual RNA. DNA was isolated by phenol:chloroform:isoamyl alcohol extraction and ethanol precipitation as described above.

DNA fragments < 700 bp in length were isolated from 9 µg of total genomic DNA using a Select-a-Size DNA Clean & Concentrator kit (Zymo Research) according to the manufacturer's recommendations ("double size selection protocol"). To ensure the binding of fragments <50 bp to the column filter, the volume of 100% ethanol added to the fraction prior to on-filter purification was increased from 290 µl to 600 µl. DNA fragments were eluted with 2 x 50 µl of elution buffer, pooled, and purified by ethanol precipitation. 100 µl DNA was mixed with 10 µl of 3 M NaOAc (0.1xV), 1 µl of 10 mg/ml glycogen (0.01xV), and 330 µl of 100% ethanol, vortexed, and incubated overnight at -80°C. DNA was recovered by centrifugation at 21000 x *g* for 30 min at 4°C. Pellets were washed 3 times with 80% cold ethanol, air-dried for 5 min, and resuspended in 5 µl of nuclease-free water.

3.8 High-throughput sequencing of DNA fragments: FragSeq

3.8.1 The libraries of DNA fragments purified from KD403, KD518, KD753, KD263, and KD675

The results of sequencing these short DNA fragments are presented in Figure 15-23, Figure 26.

The DNA oligo i116 that served as a 3' adapter was adenylated using 5' DNA Adenylation Kit (NEB), purified by ethanol precipitation as above, and diluted to 10 µM with nuclease-free water (Appendices, Table 4).

DNA fragments < 700 bp (in 5 µl water) were heat-denatured at 95°C for 5 min, cooled to 65°C, and mixed with 0.5 µM adenylated oligo i116, 1x NEBuffer 1, 5 mM $MnCl_2$, and 10 pmol of thermostable 5' App DNA/RNA ligase (NEB) in a 10-µl reaction volume. The mixture was incubated at 65°C for 1 hr, heated at 90°C for 3 min, and cooled to 4°C on ice. Ligated products were combined with 1x T4 RNA ligase buffer, 12% PEG 8000, 10 mM DTT, 60 µg/ml BSA, and 10 U of T4 RNA ligase 1 (NEB) in a 25-µl reaction volume. Since ATP was omitted from the ligation mixture and the 5' end of the i116 adapter was pre-adenylated but the fragments were not, self-ligation of fragments was prevented. The reaction was incubated at 16°C for 16 hr, 25 µl of 2x

loading dye was added, and products were separated by electrophoresis on 10% 7 M urea slab gels (equilibrated and run in 1x TBE buffer). The gel was stained with SYBR Gold nucleic acid gel stain, bands were visualized on a UV transilluminator, and products of 40 to 500 nt were excised from the gel and recovered as described in Vvedenskaya et al., 2015. Briefly, the excised gel slice was crushed, 400 µl of 0.3 M NaCl in 1x TE buffer was added, and the mixture incubated at 70°C for 10 min. The eluate was collected using a Spin-X column. After the first elution step the elution procedure was repeated, eluates were pooled, and DNA was isolated by ethanol precipitation and resuspended in 15 µl of nuclease-free water.

Next, the 3' adapter-ligated DNA fragments were adenylated using the 5' DNA Adenylation Kit (NEB) in a 20-µl reaction following the manufacturer's recommendations. Nuclease-free water was added to 100 µl, DNA fragments were purified by ethanol precipitation and resuspended in 5 µl of nuclease-free water. The two-step ligation procedure described above was repeated using 5 µl of adenylated 3'-ligated DNA fragments, 0.5 µM of barcoded oligos i112, i113, i114, or i115 that served as 5' adapters (barcodes were used as internal controls; Appendices, Table 4), 10 pmol of thermostable 5' App DNA/RNA ligase at the first ligation step, and 10 U of T4 RNA ligase 1 at the second ligation step. Reactions were stopped by the addition of 25 µl of 2x loading dye, and products were separated by electrophoresis on 10% 7 M urea slab gels (equilibrated and run in 1x TBE buffer). DNA products of 70 to 500 nt in size were excised and eluted from the gel as described above, isolated by ethanol precipitation, and resuspended in 20 µl of nuclease-free water.

To amplify DNA, 2 to 8 µl of adapter-ligated DNA fragments were added to a mixture containing 1x Phusion HF reaction buffer, 0.2 mM dNTPs, 0.25 µM Illumina RP1 primer, 0.25 µM Illumina RPI index primer, and 0.02 U/µl Phusion HF polymerase in a 30-µl reaction. PCR was performed with an initial denaturation step of 30 sec at 98°C, amplification for 15 cycles (denaturation for 10 sec at 98°C, annealing for 20 sec at 62°C and extension for 15 sec at 72°C), and a final extension for 5 min at 72°C Amplicons were isolated by electrophoresis using a non-denaturing 10% slab gel

(equilibrated and run in 1x TBE). The gel was stained with SYBR Gold nucleic acid gel stain and species of 150 to 300 bp (or up to 700 bp for paired-end sequencing) were excised. DNA products were eluted from the gel with 600 µl of 0.3 M NaCl in 1xTE buffer at 37°C for 3 hr, purified by ethanol precipitation, and resuspended in 25 µl of nuclease-free water. Barcoded libraries were sequenced on Illumina NextSeq 500 platform (high output) in single-end (1x150 bp) or paired-end (2x75 bp or 2x150 bp) modes (Supplementary Figure 1).

Bioinformatic analysis was performed in R using ShortRead and Biostrings packages (Morgan et al., 2009). Reads with no more than 2 bases with Phred quality < 20 were included in the analyses. After adapter trimming, all reads were compared to each other to reveal clusters of overamplified reads containing the same insert and combination of unique molecular identifiers conjugated to adapters. One insert per each cluster was used for further alignment to the KD403 reference genome with 2 mismatches allowed. Since some of the libraries were sequenced only in a single-end 1x150-bp mode (Supplementary Figure 1A), and most paired-end sequenced inserts were shorter than 100 nt (Supplementary Figure 1B) we further analyzed only reads that uniquely aligned to the genome and were within the 16-100 nt length span.

Logos were generated using the ggseqlogo package (Wagih, 2017). To determine the significance of nucleotide enrichment, pLogo was used, fragments mapping to genomic positions 2400000-2800000 nt were used as background (O'Shea et al., 2013).

3.8.2 The libraries of DNA fragments purified from KD403 and its DNA repair mutant derivatives

The results of the sequencing of DNA fragments purified from KD403 and its DNA repair mutant derivatives are presented in Figure 28-Figure 31, Figure 34, Figure 35. The libraries were prepared using Accel-NGS® 1S Plus DNA Library Kit (Swift Biosciences) with modifications to the standard protocol recommended by the manufacturer to retain small fragments (≥40 bp) The libraries were sequenced on Illumina NextSeq 500 platform (high output) in a 2x75 bp paired-end read mode (Supplementary Figure 2).

Bioinformatic analysis was performed in R using ShortRead and Biostrings packages (Morgan et al., 2009). Reads with no more than 3 bases with Phred quality < 20 were included in the analyses. During the first stage of library preparation following the protocol of Accel-NGS® 1S Plus DNA Library Kit (Swift Biosciences), a low complexity tail with an average length of 8 bases (and up to ≈12 bases) mostly composed of C and T nucleotides is ligated to the 3' end of each fragment. To account for these tails, for each fragment we mapped the first 30 nucleotides of the forward read corresponding to the 5' end of the fragment (which do not include the 3'-end tail if the size of a fragment is at least 30 nt) with 3 mismatches allowed. The first 30 nt of the reverse read, which include the tail and the 3'-end nucleotides of the fragment, were mapped without mismatches allowed. If the read was not aligned, the first nucleotide from the reverse read was removed and the adjacent 30 nt were mapped again. The trimming and alignment of 30 nucleotides from the beginning of the reverse read was repeated until the read was aligned but not more than 15 times. The positions of each forward and reverse reads of aligned pairs on the chromosome were compared. If the two reads were properly oriented relative to each other and the distance between their 5' ends was less than 1000 nt, the 5'-end positions of the forward and reverse reads were regarded as the positions of fragments 5' ends and 3' ends, respectively.

Chapter 4. Results

4.1 *In Vivo* Detection of Primed Adaptation Intermediates in Type I CRISPR-Cas systems

The results presented in section 4.1 are published in:

Shiriaeva, A.A., Savitskaya, E., Datsenko, K.A., Vvedenskaya, I.O., Fedorova, I., Morozova, N., Metlitskaya, A., Sabantsev, A., Nickels, B.E., Severinov, K., et al. (2019). Detection of spacer precursors formed *in vivo* during primed CRISPR adaptation. Nat Commun *10*, 4603.

Self-targeting strains used in this chapter were constructed by K. Datsenko. Fluorescence microscopy was performed by N. Morozova (Figure 12A). Experiments on the quantification of colony-forming units were performed by I. Fedorova (Figure 12A). Experiments on self-targeting by the type I-F system were performed by E. Semenova (Figure 26B).

The author performed all experiments on self-targeting and oligo transformation in the type I-E system presented in Figure 12B - Figure 25. The author performed the analysis of high-throughput sequencing data for all results presented in chapter 4.1. The author prepared FragSeq libraries described in this chapter. The author purified DNA for high-throughput sequencing of total genomic DNA; genomic DNA libraries were prepared by Waksman Genomics Core Facility, Rutgers University, USA. The author prepared PCR amplicons of extended CRISPR arrays for further high-throughput sequencing; the libraries were prepared by Waksman Genomics Core Facility, Rutgers University, USA. HTS of all libraries including FragSeq libraries was performed at Waksman Genomics Core Facility, Rutgers University, USA. The author's work described in this chapter was performed in Konstantin Severinov and Bryce Nickels laboratories at Waksman Institute of Microbiology, Rutgers University, USA.

4.1.1 A genetic system for studying CRISPR-mediated self-targeting of *E. coli* genome

A derivative of an *E. coli* K12 strain KD403 with *cas* genes under the control of inducible promoters and a single CRISPR array containing a spacer complementary to a chromosomally located non-essential *yihN* gene was constructed (we will refer to this strain as "wild type", or "*wt*") (Figure 11). The *yihN* protospacer (further referred to as "PPS", for "*p*riming *p*roto*s*pacer") is preceded with the interference-proficient consensus PAM 5'-AAG-3'. There is a single mismatch between the spacer in the crRNA and the PPS, located immediately downstream of the PAM, at position +1.

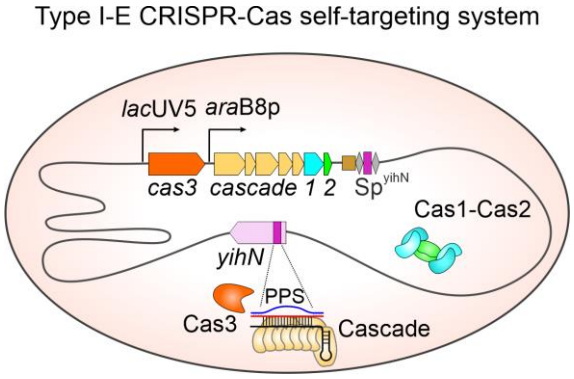


Figure 11. The type I-E self-targeting system. Shaded oval, an *E. coli* cell; grey line, chromosome; orange, tan, blue, and green pentagons, *cas* genes; brown rectangle, CRISPR-array leader sequence; grey diamonds, repeats; purple rectangle, spacer targeting *yihN* (Sp^{yihN}); *lac*UV5 and *ara*B8p, promoters; mauve pentagon, *yihN*; PPS, priming protospacer within *yihN*; blue line, nontarget strand; red line, target strand; black line, crRNA. This figure is published in (Shiriaeva et al., 2019).

Previously published data revealed that plasmids harboring a protospacer with a single mismatch at the +1 position are subject to interference (Semenova et al., 2016). In agreement with this observation, cells undergoing CRISPR self-targeting formed fewer colonies compared to non-induced cells (Figure 12A). The number of CFUs in induced cultures started to drop 3 hours after the addition of *cas* gene inducers and reached its minimum after 5 hours (Figure 12A). Fluorescent microscopy revealed a dramatic increase in cell lengths in induced cultures (Figure 12A). Surprisingly, more than 96% (287 out of 296) of induced cells remained alive as judged by staining with SYTO9

(stains live cells green) and propidium iodide (stains dead cells red) (Figure 12A) (Boulos et al., 1999; López-Amorós et al., 1995; Stocks, 2004).

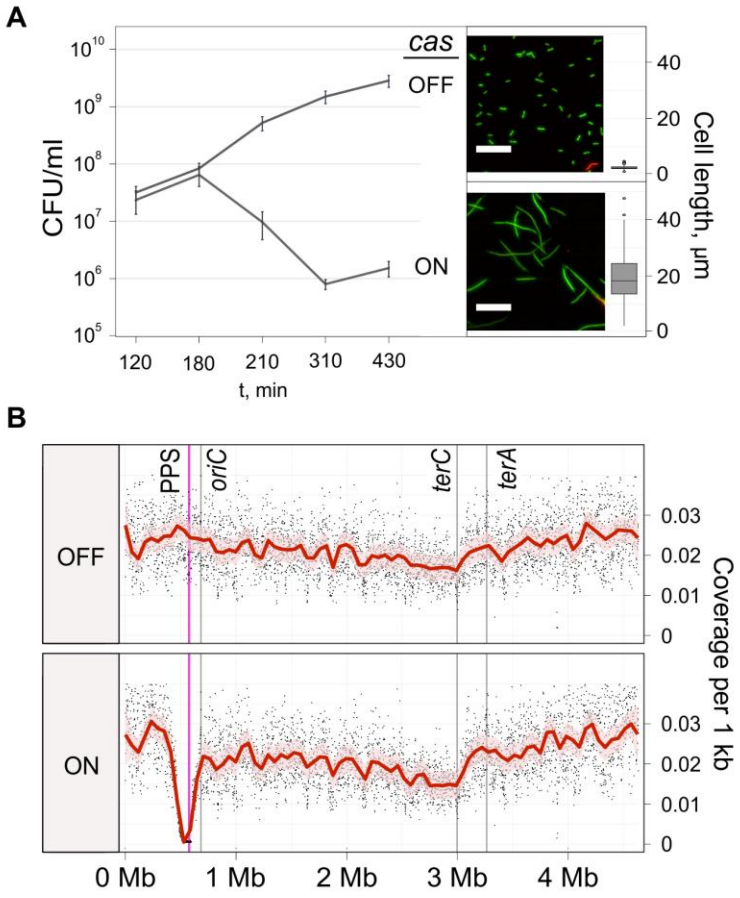


Figure 12. Self-targeting of the genome by the type I-E CRISPR-Cas system leads to CRISPR interference. **A.** Growth curves for self-targeting cultures in which *cas* gene expression is induced (ON) or not induced (OFF). Mean±SEM of CFU/ml values obtained in four biological replicates are shown for indicated time points postinduction. Green, viable cells; red, non-viable cells; scale bar, 20 μm. Boxplot: the central line, median; hinges, the first and third quartiles; whiskers, 1.5 x IQR; n = 125. **B.** Effect of self-targeting on genomic DNA content. *oriC*, replication origin; *terA* and *terC*, sites of replication termination; dot, coverage per 1 kb; red line, Loess smoothing; pink shading, 99% confidence interval. This figure is published in (Shiriaeva et al., 2019).

Cell filamentation is a known phenotype of SOS response resulting from various types of DNA damages including DSBs (Meddows et al., 2005). To test for the presence of DSBs, high-throughput sequencing of genomic DNA purified from induced and uninduced cultures was performed (Figure 12B). As expected, genomic coverage in

uninduced cells was evenly distributed with a gradual decline from the origin of replication towards the *ter* sites (Figure 12B). A similar decline was observed in DNA prepared from cultures undergoing CRISPR interference but, in addition, a dramatic drop was evident in the region surrounding the PPS (Figure 12B). The coverage began to decline ≈200 kbp up- and ≈100 kbp downstream of the PPS gradually approaching its lowest value in the immediate vicinity of the PPS. DNA coverage was also analyzed in nontargeting cells that did not have the self-targeting spacer; expressed catalytically inactive Cas1 H208A, or nuclease-deficient Cas3 H74A (Figure 13A) (Babu et al., 2011; Westra et al., 2012). Cas1 inactivation did not have any effect on genome content near the PPS while inactivation of Cas3 prevented the loss of DNA making the distribution similar to that observed in uninduced cells or in induced control cells lacking the self-targeting spacer (Figure 13A).

The dramatic drop in genomic coverage around the PPS apparently reveals the extent to which DNA can be degraded by Cas3 – alone or aided by host nucleases. The initial protocol used for HTS library preparation allowed the sequencing of double-stranded DNA only (Figure 13B). To rule out a possibility that DNA in proximity to the PPS was single-stranded and thus was lost during library construction, we followed a different protocol compatible with single-stranded DNA. Strand-specific DNA sequencing revealed a drop in genomic coverage near the PPS similar to that observed for the dsDNA-specific protocol (Figure 13B).

Thus, when a spacer matching the bacterial genome is placed into the CRISPR array, CRISPR interference leads to the degradation of up to several hundred kbp of genomic DNA around the targeted protospacer. It leads to the blockage of cell division resulting in cell filamentation. However, cells remain alive during at least several hours after activation of *cas* gene expression.

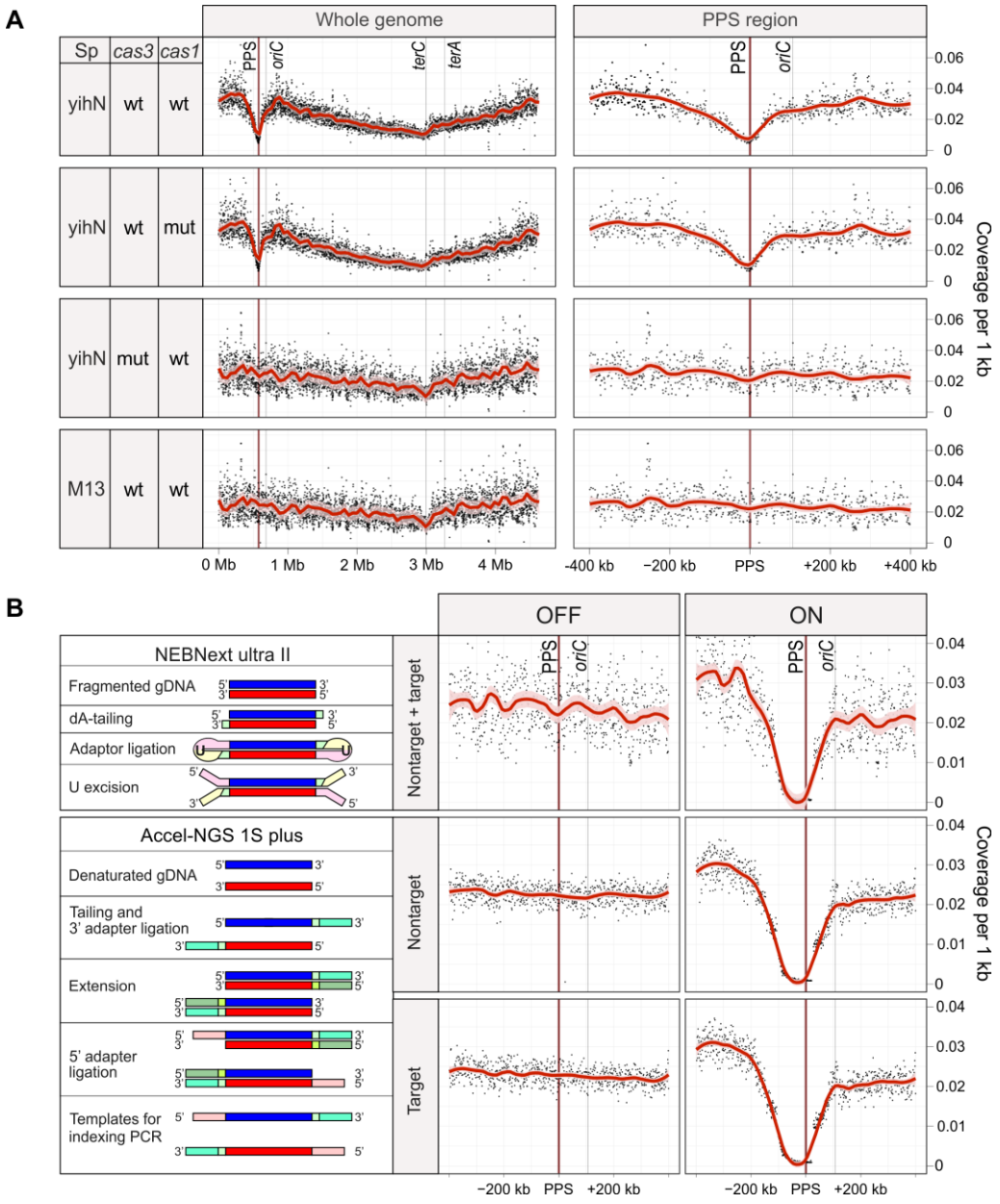


Figure 13. HTS analysis of genomic DNA purified from self-targeting and nontargeting cultures. **A.** HTS analysis of genomic DNA: effects of disruptions in components of interference or adaptation modules. Graph of sequence coverage per 1 kb for the whole genome (left) or PPS^{yihN} region (right) in the indicated strains. *oriC*, site of replication origin; *terA* and *terC*, sites of replication termination; dot, coverage per 1 kb (mean of 3 biological replicates); red line, Loess smoothing; pink shading, 99% confidence interval. *cas1* mut, gene encoding Cas1^{H208A}, *cas3* mut, gene encoding Cas3^{H74A}. **B.** HTS analysis of genomic DNA purified from self-targeting *wt* cultures: comparison of library construction methods. Left, steps in library construction using a NEBNext ultra II kit (analysis of double-stranded DNA) or Accel NGS 1S plus kit (analysis of single-stranded and double-stranded DNA). Right, PPS-region coverage plots obtained for wild-type cells. The 100% values on the Y-axis correspond to the total coverage with genomic DNA reads mapped to the genome. This figure is published in (Shiriaeva et al., 2019).

4.1.2 Primed spacer acquisition during self-targeting

CRISPR interference is associated with primed adaptation, a process during which new spacers are acquired from the target DNA, and the orientation of new spacers is dictated by the orientation of the PPS (Datsenko et al., 2012; Swarts et al., 2012). PCR analysis with primers annealed to the leader sequence and Sp^{yihN} in the CRISPR array revealed the acquisition of additional spacers in induced self-targeting cultures (Figure 14A,B). Judging by relative intensities of amplified bands, 23±2% of cells acquired an extra spacer by 5 hours postinduction. No spacer acquisition was detected in uninduced cells or cells where Sp^{yihN} was replaced with Sp^{M13}, which did not have a target in the described experiments (Sp^{M13} targets the M13 phage genome). These results suggest that newly adapted spacers are acquired during primed adaptation.

To determine the source of newly acquired spacers, we purified PCR fragments corresponding to the expanded CRISPR arrays and subjected them to high-throughput sequencing (Figure 14C,D). Alignment of spacers to the genome revealed, that 98.8±0.3% of spacers originated from the genomic region spanning 100 kbp up- and 100 kbp downstream of the PPS (Figure 14D). The number of protospacers decreased gradually as the distance from the PPS increased. 58.1±0.1% of spacers were mapped to the 100-kbp region upstream of the PPS while 40.7±0.3% originated from the 100-kbp downstream region. The orientation of protospacers was opposite for the upstream and downstream regions. 98.5±0.1% of spacers mapping to the upstream region corresponded to PS^{NT} (57.3±0.2% of all protospacers) while 98.3±0.3% of spacers mapping to the downstream region corresponded to PS^{T} (40±0.4% of all protospacers) (Figure 14C,D). 97.5±0.3% of PS^{NT} upstream of the PPS and 97.8±0.3% of PS^{T} downstream of the PPS were flanked by 5'-AAG-3' (Figure 14D). Taken together, these results demonstrate that the self-targeting strain acquires spacers in a bidirectional, orientation-dependent manner characteristic of primed adaptation described for the *E. coli* I-E system (Datsenko et al., 2012; Savitskaya et al., 2013; Swarts et al., 2012).

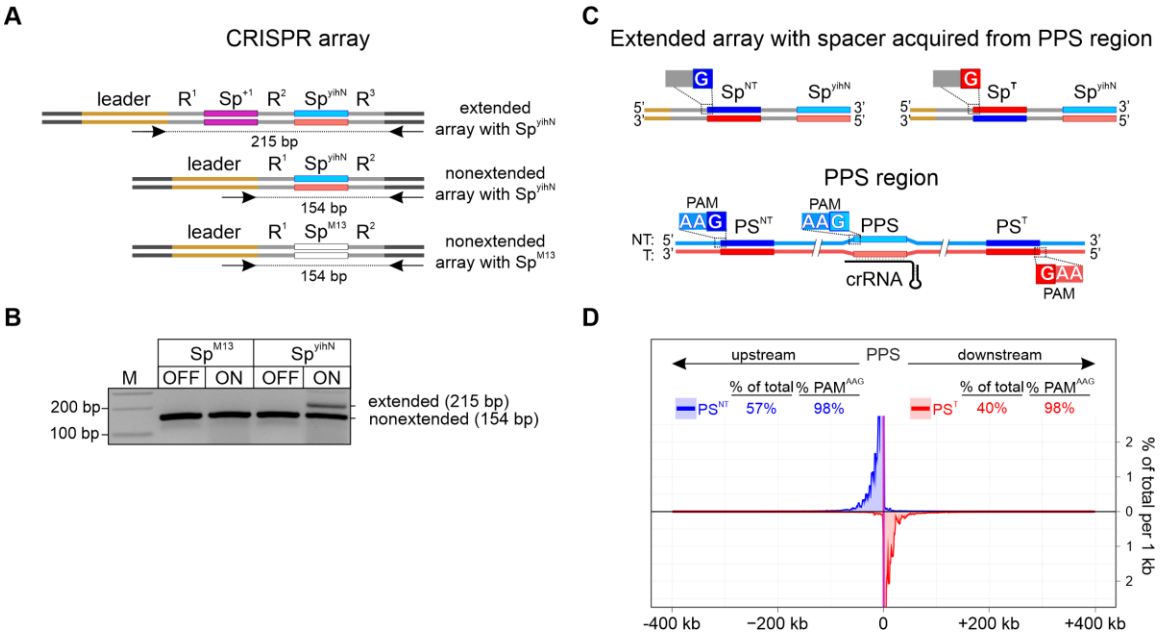


Figure 14. Self-targeting of the genome by the type I-E CRISPR-Cas system leads to primed adaptation. **A.** The scheme depicts an extended CRISPR array containing Sp^{yihN} and an acquired spacer Sp^{+1}, a nonextended array containing Sp^{yihN} only, or an array containing a spacer targeting M13 phage (Sp^{M13}). Blue line, non-transcribed strand of Sp^{yihN}; red line, transcribed strand of Sp^{yihN} (directs synthesis of crRNA); R, repeats. Arrows below arrays represent the positions of primers used for PCR whose products are shown in **B**; sizes of PCR amplicons are indicated. **B.** PCR analysis of cells containing an array with Sp^{yihN} or Sp^{M13} shown in **A**. M, double-stranded DNA marker. **C.** The scheme depicts extended arrays with spacers acquired from protospacers in the PPS-region. Sp^{NT}, a spacer with the non-transcribed strand derived from the nontarget strand (NT, blue) and the transcribed strand derived from the target strand (T, red); Sp^{T}, a spacer with the non-transcribed strand derived from the target strand (T, red) and the transcribed strand derived from the nontarget strand (NT, blue). PS^{NT}, a protospacer for Sp^{NT}; PS^{T}, a protospacer for Sp^{T}. **D.** High-throughput sequencing analysis of spacers acquired during self-targeting. The plot shows the percentage of spacers per 1 kb of the genome derived from PS^{NT} (blue) or PS^{T} (red). The 100% value on the Y-axis corresponds to the total number of spacers mapped to the genome. The width of blue and red lines in the plot corresponds to mean±SEM values obtained in three biological replicates. This figure is published in (Shiriaeva et al., 2019).

4.1.3 Detection of DNA fragments specific for primed adaptation

The exact mechanism of prespacer generation during primed adaptation and the structure of prespacers *in vivo* are not determined. We set out to detect prespacers generated *in vivo* and their possible longer precursors in self-targeting cells using a high-throughput sequencing approach. To achieve this goal, we developed FragSeq - a protocol for strand-specific high-throughput sequencing of short single-stranded and double-stranded fragments generated *in vivo* (Figure 15). The procedure starts with the

purification of total DNA using phenol/chloroform extraction to retain small fragments that may be lost if silica column-based DNA purification methods are applied. The purified DNA is then filtered using a commercial kit allowing for selection of fragments shorter than ~700 bp in length (see Materials and Methods). After enrichment with short fragments, DNA is denatured, and single-stranded adapters are consecutively ligated to the 3' and 5' ends by ligases capable of ligating single-stranded DNA. The libraries are then amplified and sequenced on Illumina platforms. Since no tailing is applied prior to adapter ligation, the procedure allows mapping of 5' and 3' ends with one-nucleotide resolution. In addition, both adapters have unique molecular identifiers consisting of 9 or 11 random nucleotides that are used during downstream analysis to eliminate fragments overamplified during library preparation.

Using this approach, we sequenced and analyzed 16-100-nt DNA fragments purified after incubation with *cas* genes inducers from the *wt* self-targeting strain containing Sp^{yihN}, its derivatives expressing inactive $Cas1^{H208A}$ or $Cas3^{H74A}$, and the nontargeting strain with Sp^{M13} (Figure 16). Fragments produced in all strains were mapped throughout the whole genome. A prominent sharp peak of short fragments mapping 25 kbp up- and 25 kbp downstream of the PPS (indicated with a bracket in Figure 16B,C) on both target and nontarget strands was revealed in the *wt* self-targeting strain. No such peak was detected in the *cas1* mutant, the *cas3* mutant, and the nontargeting strain (Figure 16B,C). A much broader shallow enrichment area was seen in both the *wt* and the *cas1* mutant but absent in other strains (Figure 16B, see also page 87 below).

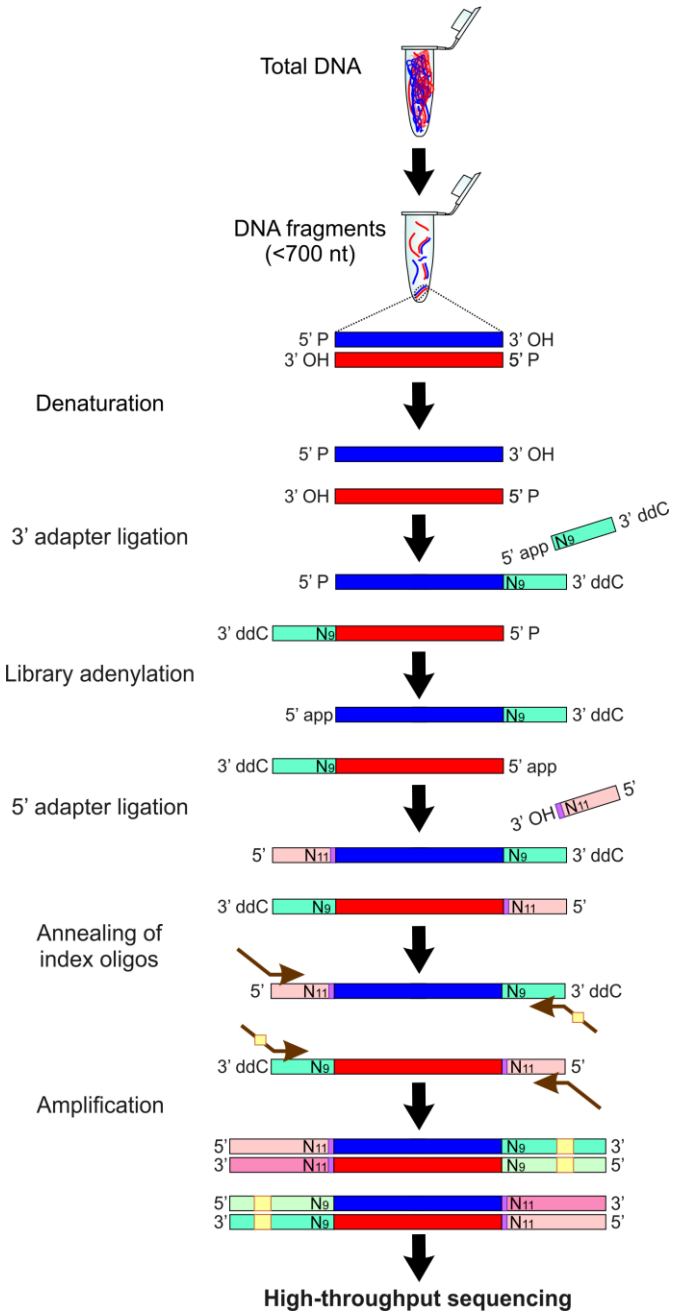


Figure 15. Strand-specific, high-throughput sequencing of DNA fragments, “FragSeq.” Steps in library construction. 5' app, adenylated 5' end; 3' ddC, blocked 3' end; N9 and N11, unique molecular identifiers on 3' and 5' adapters; purple rectangle, 4-nt barcode on 5' adapter; yellow rectangle, index. This figure is published in (Shiriaeva et al., 2019).

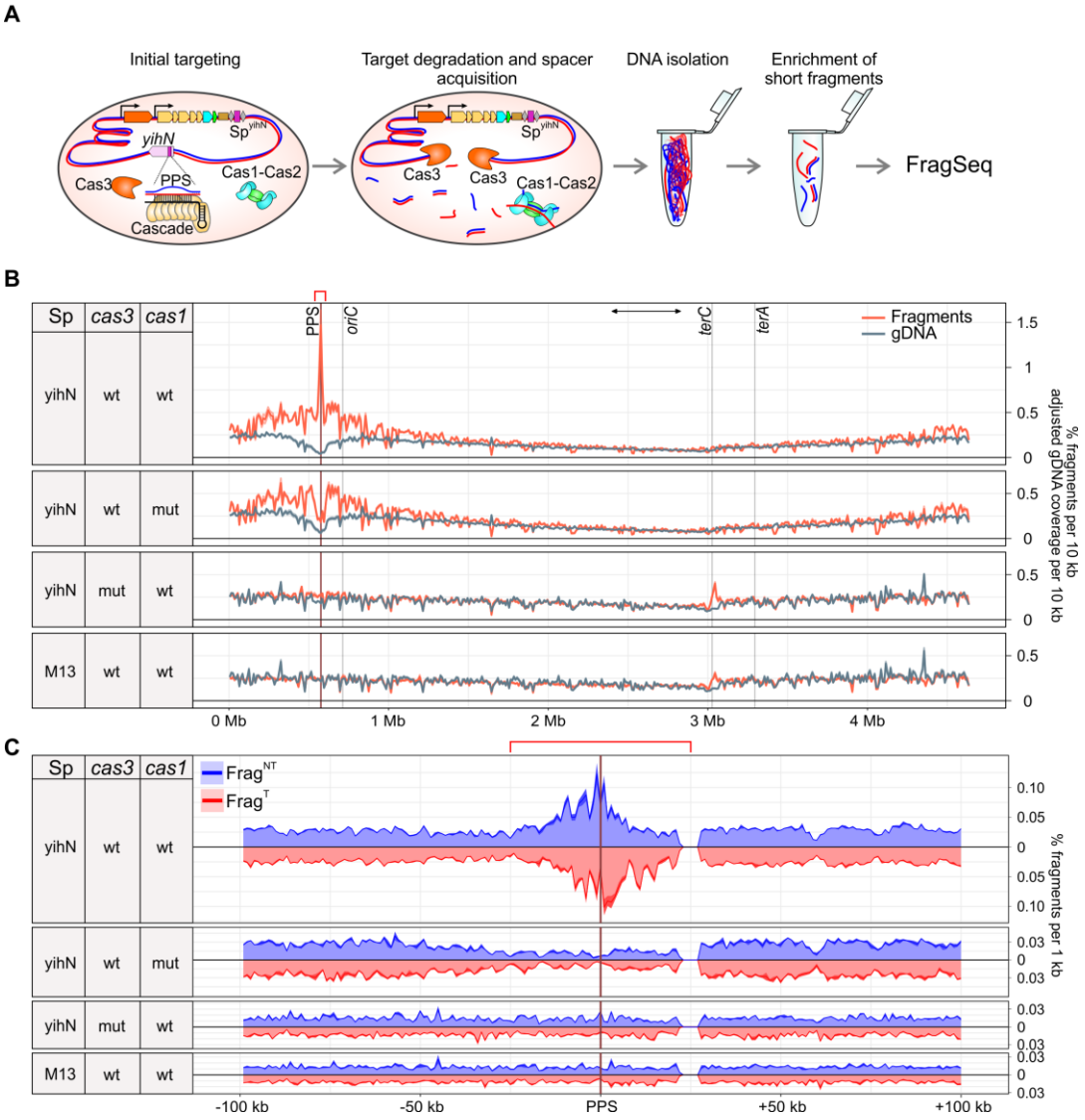


Figure 16. HTS of fragments purified from self-targeting and nontargeting cultures. **A.** A schematic showing the events occurring in *wt* self-targeting cells upon induction of *cas* gene expression and key steps in short DNA fragments purification. **B.** FragSeq results for the type I-E self-targeting system: comparisons of fragment coverage plots and total genomic DNA (gDNA) coverage plots are shown. Percentage of DNA fragments and gDNA coverage per each 10 kb of the *E. coli* genome was calculated (100% equals to either all fragments or the total coverage with all gDNA mapping reads). The coverage values for gDNA were adjusted to superimpose plots in a control region located far away from the PPS (this region is shown by a black line with opposing arrows at the top of the figure). Coordinates on the X-axis represent the location on the *E. coli* chromosome. **C.** Fragment distributions by strands in a 200-kb region centered at the PPS. The 100% value on the Y-axis corresponds to the total number of fragments mapped to the genome for indicated strains. Coordinates on the X-axis represent the distance from the PPS. Blue, nontarget-strand-derived fragments ($Frag^{NT}$); red, target-strand-derived fragments ($Frag^{T}$). Mean±SEM values obtained from three biological replicates are shown. A red bracket above the plots shown in **B** and **C** indicates the enrichment of fragments revealed in the *wt* strain.

To reveal if any specific motifs were associated with the ends of fragments generated in the PPS region, we analyzed sequences of 10 terminal positions on the fragments' 5' and 3' ends as well as the 10-nt flanking chromosomal sequences adjacent to them (40 positions in total). The background probabilities to obtain each of four nucleotides in each of the 40 positions were calculated for fragments mapped to a control region, which was not degraded during self-targeting (positions 2400000-2800000 of the reference genome). The log-odds approximation of the binomial probability for each nucleotide in the region 25 kbp upstream or 25 kbp downstream of the PPS was then calculated and presented as pLogo (Figure 17) (O'Shea et al., 2013). The relative abundance of A, T, C, or G in each position for fragments from the PPS region compared with the control region is also shown as fold-enrichment in Figure 18. The results demonstrate that several significantly overrepresented residues were present in the *cas1* mutant, the *cas3* mutant, and even in the control non-targeting strain (Figure 17). However, the magnitude of these differences was negligible (Figure 18). Much larger differences were observed near the ends of fragments produced around the PPS in the *wt* cells (Figure 18). The 5'-end terminal sequences of fragments mapped to the NT-strand upstream of the PPS and the T-strand downstream of the PPS as well as chromosomal sequences adjacent to them were enriched with A and G nucleotides (Figure 17, Figure 18). The 3'-end terminal sequences of fragments mapped to the T-strand upstream of the PPS and the NT-strand downstream of the PPS were highly enriched with T and C nucleotides (Figure 17, Figure 18). These results indicate that fragments in *wt* are excised from 5'-AAG-3'/3'-TTC-5' PAM-containing sequences oriented with respect to the PPS similarly to PAMs of protospacers selected during primed adaptation. Interestingly, almost all other analyzed positions were also significantly enriched with one or another nucleotide (Figure 17). The total number of fragments mapped to the PPS region was ≈2.6-fold higher than the number of unique genomic positions where these fragments were mapped to. When for each set of fragments with identical sequences only one fragment was included in the analysis, the number of significantly enriched positions decreased, while the enrichment with the consensus PAM remained (Figure 19).

Altogether, these data suggest that the presence of the consensus PAM stimulates the generation of fragments, while the frequency at which each particular fragment is produced might be influenced by its nucleotide composition.

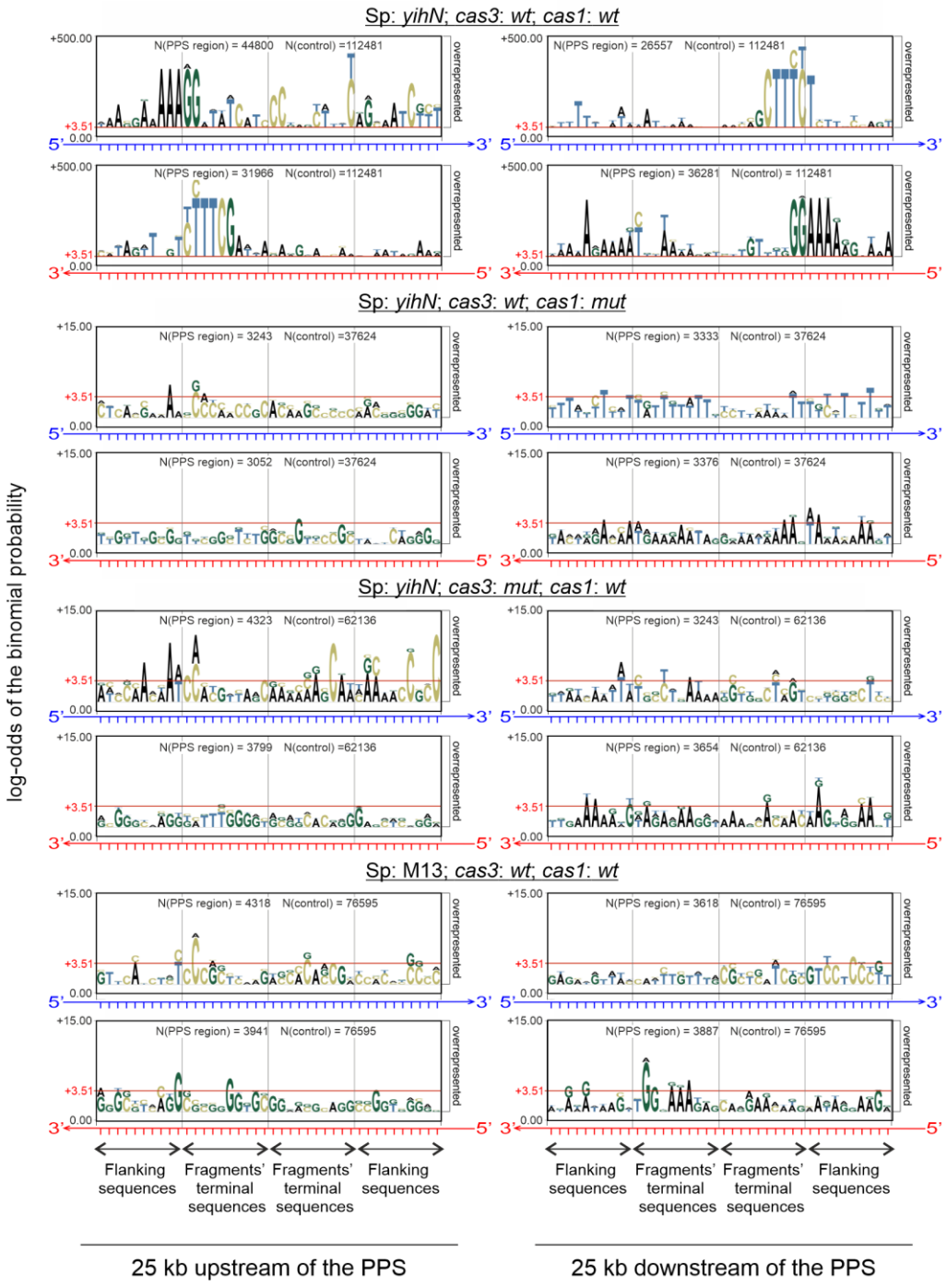


Figure 17. Significantly enriched nucleotides in terminal or flanking sequences of 16-100-nt fragments mapped to the PPS region. The PPS-region sequences were compared with the sequences of 16-100-nt fragments mapped to a control region. The figures were generated using pLogo (O'Shea et al., 2013). A region within self-targeting genome coordinates 2400000-2800000 was taken as a control. Ten positions of chromosomal sequences flanking each fragment's end and ten positions of fragments' terminal 5' or 3' sequences are shown along the X-axis. The logarithm of odds ratio of getting a frequency not higher than in the control region to the frequency not less than in the control region is shown along the Y-axis. The red line shows a statistical significance value (logarithm of the odds (1- α')/α' where α' is Bonferroni corrected α=0.05 for 4 possible nucleotides in 40 positions (160 comparisons in total, α'=0.0003125)). Nucleotides with the height higher than the significance value are significantly overrepresented.

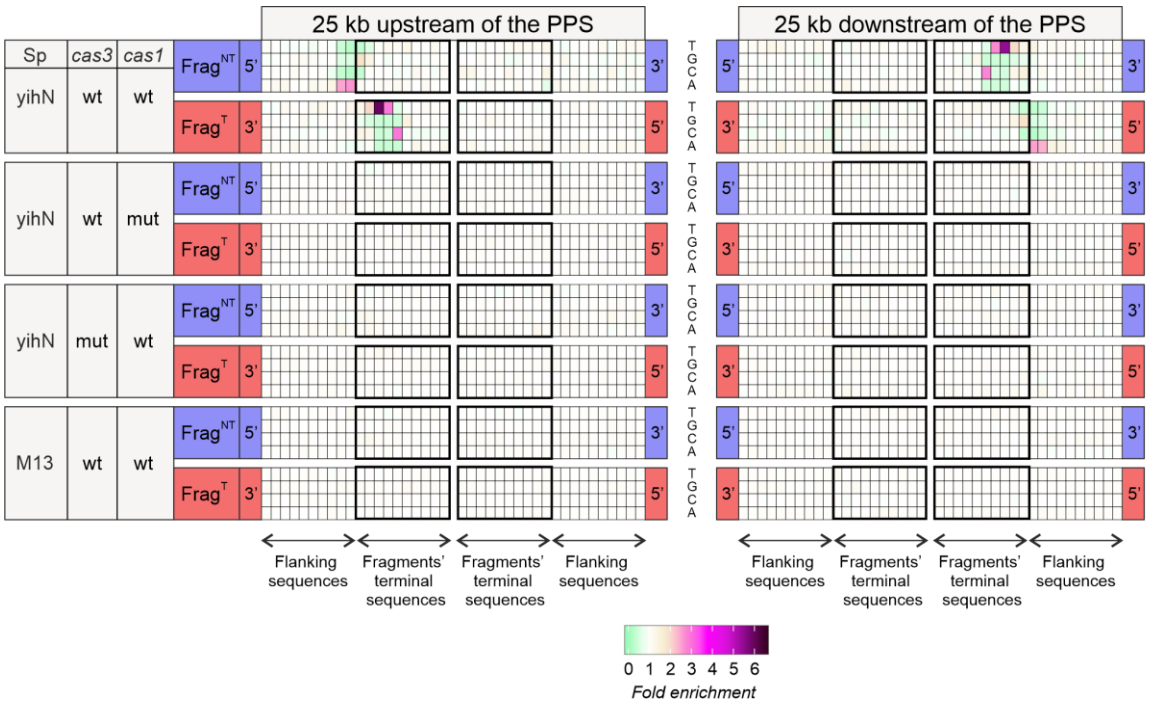


Figure 18. Sequence analysis of 16-100-nt fragments mapped to the region 25 kbp up- or 25 kbp downstream of the PPS. Heat maps of relative abundance of A, T, C, or G for the indicated fragments' 5' or 3' ends are shown. Ten positions of sequences that are detected in fragments' 5' or 3' ends are shown in black rectangles. Shading represents enrichment (>1) or depletion (<1) of each nucleotide for sequences associated with PPS-region-derived fragments vs. sequences associated with non-PPS-region-derived fragments (genome coordinates 2400000-2800000). Blue, nontarget-strand-derived fragments ($Frag^{NT}$); red, target-strand-derived fragments ($Frag^{T}$). This figure is published in (Shiriaeva et al., 2019).

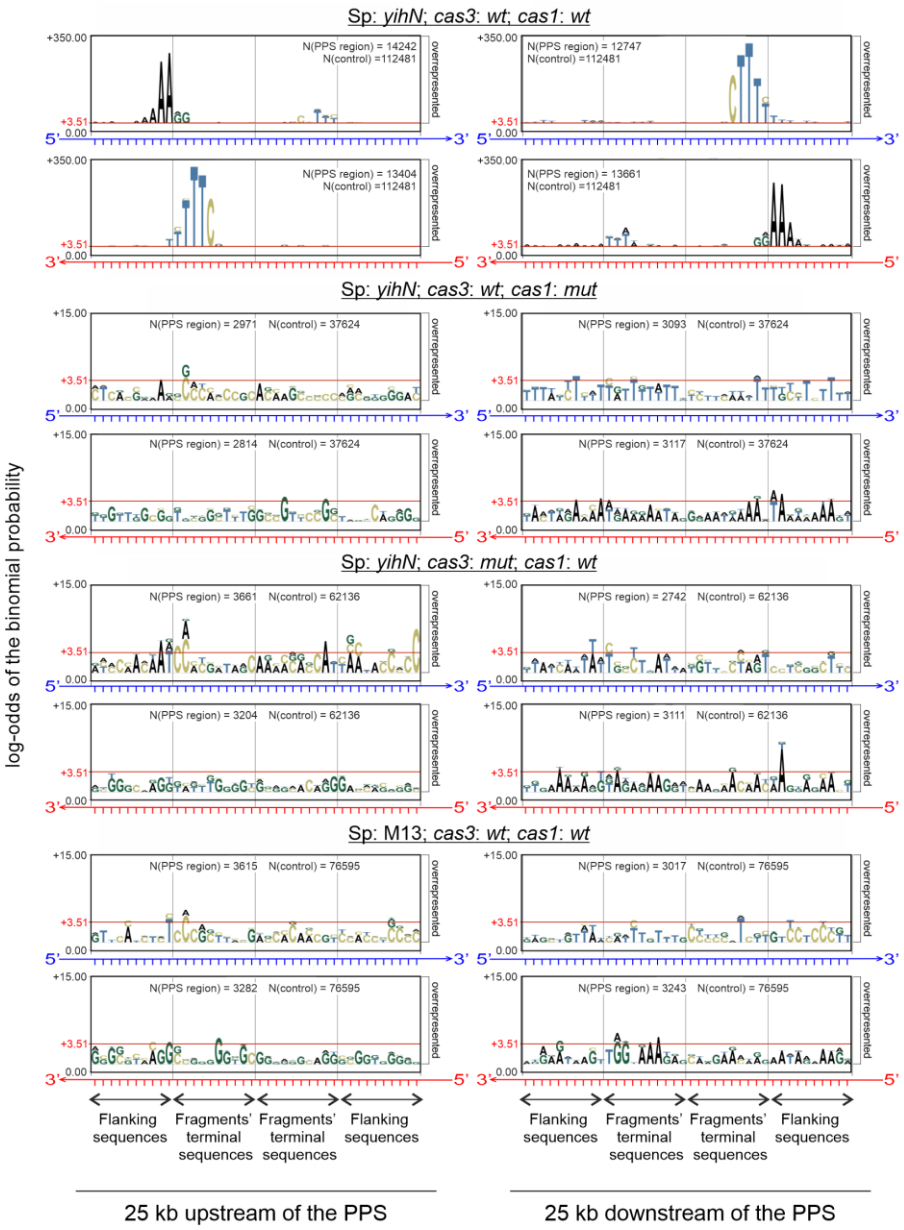


Figure 19. Significantly enriched nucleotides in terminal or flanking sequences of *unique* 16-100-nt fragments mapped to the PPS region. The PPS-region sequences were compared with the sequences of *unique* 16-100-nt fragments mapped to a control region. The figures were generated using pLogo (O'Shea et al., 2013). A region within genome coordinates 2400000-2800000 was taken as a control. The duplicates of sequences represented in more than 1 copy were removed from the analysis. Ten positions of chromosomal sequences flanking each fragment's end and ten positions of fragments' terminal 5' or 3' sequences are shown along the X-axis. The logarithm of odds ratio of getting a frequency not higher than in the control region to the frequency not less than in the control region is shown along the Y-axis. The red line shows the statistical significance value (logarithm of the odds (1- α')/α' where α' is Bonferroni corrected α=0.05 for 4 possible nucleotides in 40 positions (160 comparisons in total, α'=0.0003125)). Nucleotides whose height is higher than the significance value are significantly overrepresented.

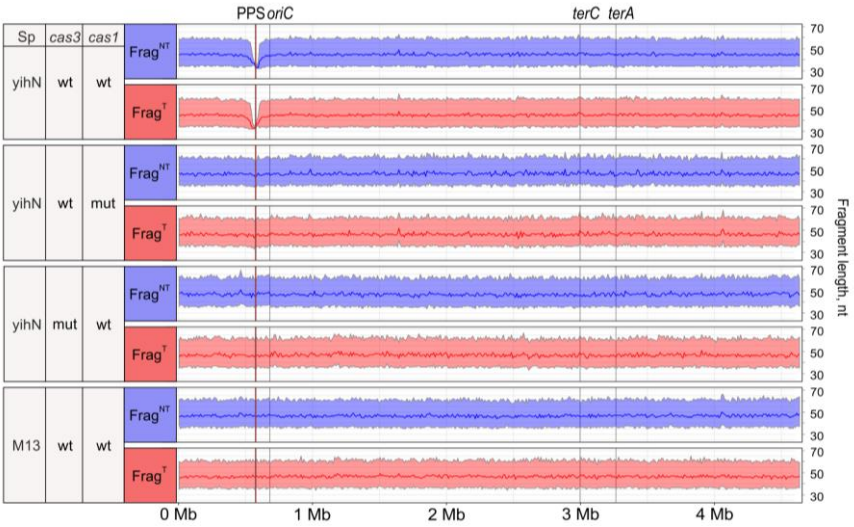


Figure 20. FragSeq results for the type I-E self-targeting system: length distributions. Length distributions of genome-derived fragments in the indicated strains. Coordinates on the X-axis represent the location on the *E. coli* chromosome. Solid lines represent the median fragment length (for fragments aligned to every 10 kb of the genome), shaded areas represent fragment lengths between the first and third quartiles.

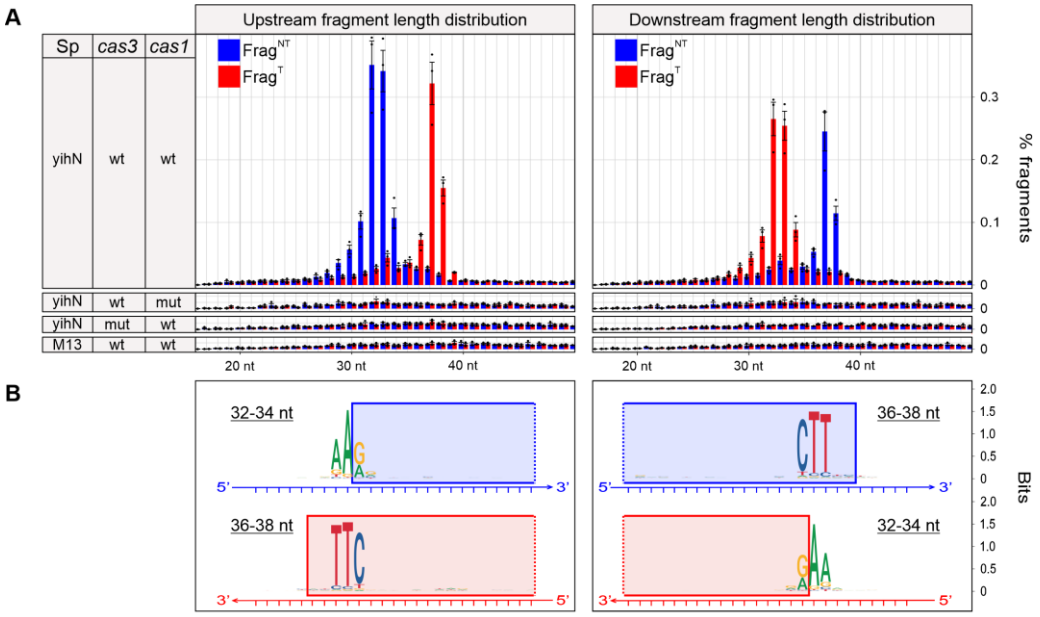


Figure 21. Spacer-size fragments associated with the 5'-AAG-3'/3'-TTC-5' PAM are produced in the *wt* self-targeting strain. **A.** Length distributions of PPS-region-derived fragments (mean±SEM values obtained from three biological replicates; the 100% value on the Y-axis corresponds to all fragments mapped to the genome). **B.** Logos of fragments' terminal sequences (shown inside colored rectangles) and chromosomal flanking sequences of genomic DNA from which PPS-region fragments are derived. Blue rectangles, sequences present in $Frag^{NT}$; red rectangles, sequences present in $Frag^{T}$. Upstream, 25-kbp region upstream of the PPS; downstream, 25-kbp region downstream of the PPS. The sequence logos were generated using ggseqlogo (Wagih, 2017). The figure is published in (Shiriaeva et al., 2019).

In all strains, median lengths of selected (16-100-nt) fragments mapping outside the PPS-containing region were 45-47 nt (Figure 20). In comparison, fragments mapping to the PPS-containing region of the *wt* self-targeting strain were shorter with the median length of 35 nt (Figure 20). When the lengths of fragments were examined separately for regions up- and downstream of the PPS, two strand-dependent peaks at 32-34 and 36-38 nt were observed (Figure 21A). Upstream of the PPS, the most abundant fragments mapping to the nontarget ($Frag^{NT}$) and target ($Frag^{T}$) strands were 32-34 and 36-38 nt, respectively, while downstream of the PPS the reversed strand bias was observed (Figure 21A). Sequence analysis of fragments and chromosomal regions flanking fragments' ends revealed that 86.8±1.1% of the 36-38 nt fragments from the target strand up- and the nontarget strand downstream of the PPS had a 3'-NNTTC-5' motif on fragments' 3' ends (Figure 21B). Among the 32-34 nt fragments mapped to the nontarget strand up- and the target strand downstream of the PPS, 91.1±0.6% had 5'-AA/G-3' or 5'-A/AG-3' motif associated with fragments' 5' ends (slash indicates the boundary between fragments' 5' ends and the adjacent chromosomal sequences) (Figure 21B). Thus, the results of FragSeq suggest cells undergoing primed adaptation accumulate spacer-size 32-34-bp double-stranded DNA fragments containing a 4-nt 3'-NNTT-5' or a 3-nt 3'-NNT-5' overhang on the PAM-derived 3'-end (Figure 21B). Furthermore, the relative abundance of these fragments and spacers that had an identical sequence and were acquired during primed adaptation showed positive correlation suggesting that the detected fragments are related to spacers (AAG-associated fragments: Pearsons' r = 0.57, 95% confidence interval 0.54-0.59, p-value < 2.2e-16; TTC-associated fragments: Pearsons' r = 0.5, 95% confidence interval 0.48-0.53, p-value < 2.2e-16).

Similar to results presented for 16-100-nt fragments (Figure 17, Figure 19), several positions not related to the PAM were significantly enriched with various nucleotides in spacer-size fragments (Figure 22). However, the number of positions significantly enriched with one or another nucleotide was decreased when only unique fragments were considered, suggesting that these positions might influence frequencies of occurrence of different fragments (Figure 23).

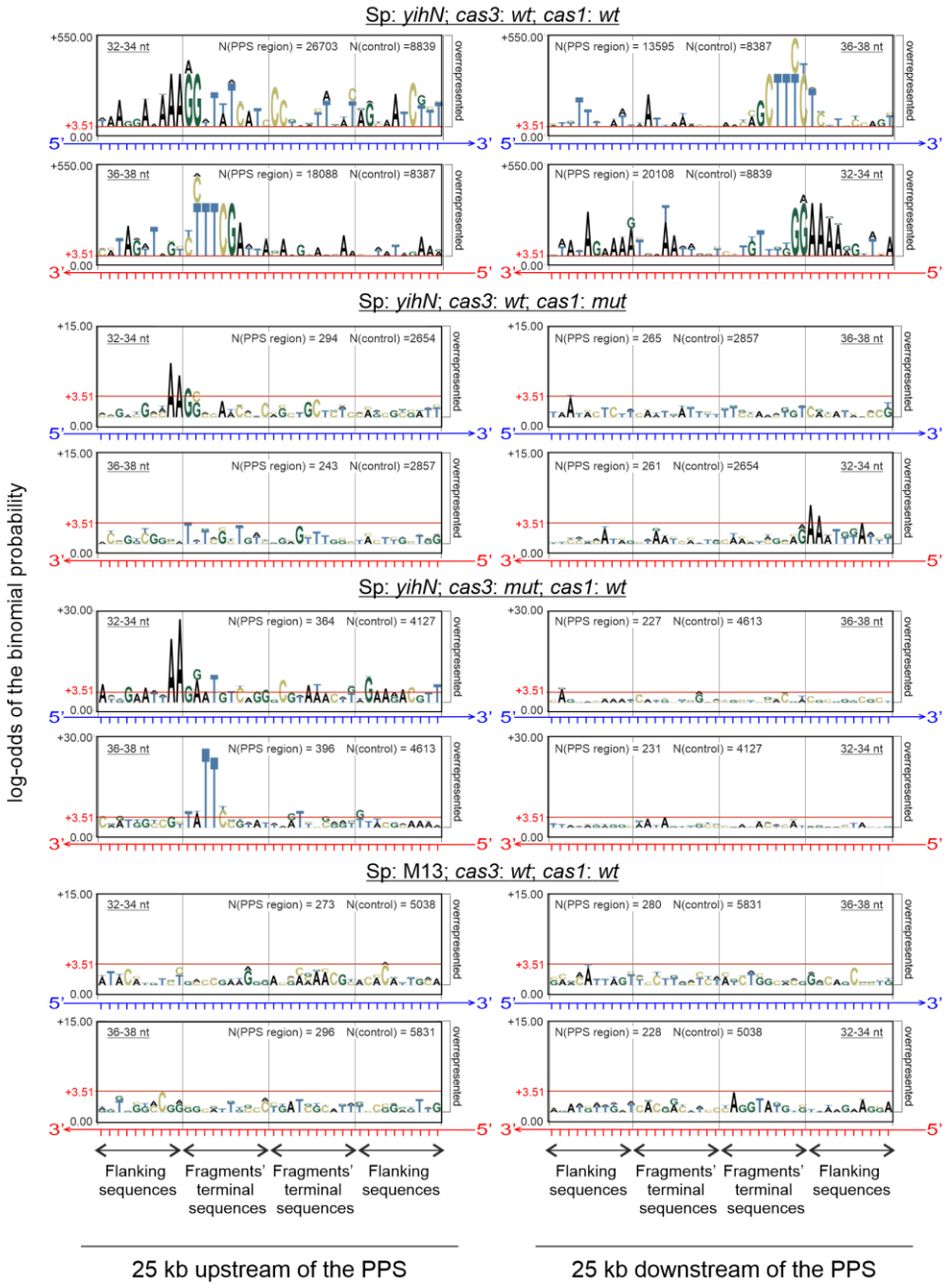


Figure 22. Significantly enriched nucleotides in terminal or flanking sequences of spacer-size fragments mapped to the PPS region. The PPS-region sequences were compared with the sequences of spacer-size fragments mapped to a control region. The figures were generated using pLogo (O'Shea et al., 2013). A region with genome coordinates 2400000-2800000 was taken as a control. Ten positions of chromosomal sequences flanking each fragment's end and ten positions of fragments' terminal 5' or 3' sequences are shown along the X-axis. The logarithm of odds ratio of getting a frequency not higher than in the control region to the frequency not less than in the control region is shown along the Y-axis. The red line shows the statistical significance value (logarithm of the odds (1- α')/α' where α' is Bonferroni corrected α=0.05 for 4 possible nucleotides in 40 positions (160 comparisons in total, α'=0.0003125)). Nucleotides whose height is over the significance value are significantly overrepresented.

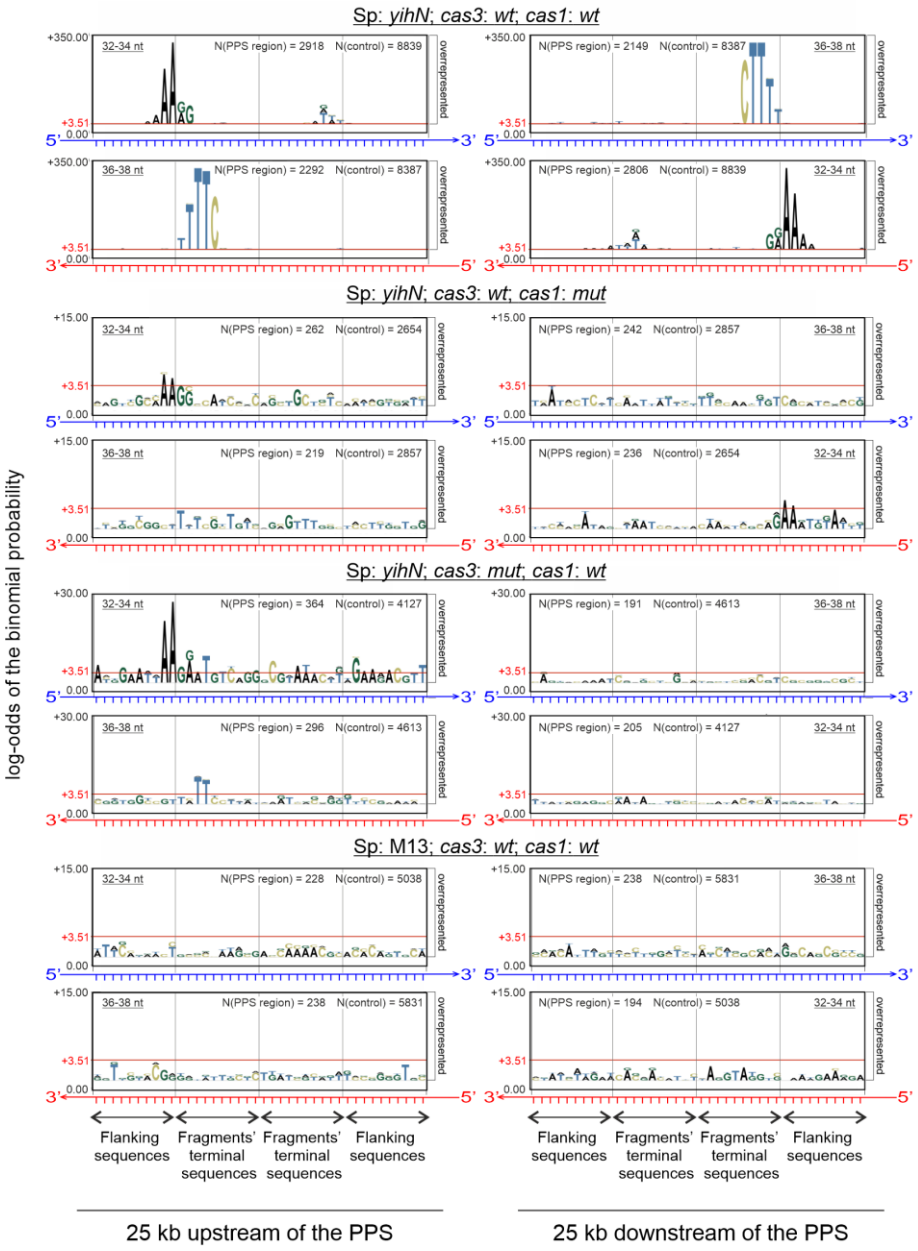


Figure 23. Significantly enriched nucleotides in terminal or flanking sequences of *unique* spacer-size fragments mapped to the PPS region. The PPS-region sequences were compared with the sequences of *unique* spacer-size fragments mapped to a control region. The figures were generated using pLogo (O'Shea et al., 2013). A region within self-targeting genome coordinates 2400000-2800000 was taken as a control. Only one copy of sequences present in more than one copy was included in the analysis. Ten positions of chromosomal sequences flanking each fragment's end and ten positions of fragments' terminal 5' or 3' sequences are shown along the X-axis. The logarithm of odds ratio of getting a frequency not higher than in the control region to the frequency not less than in the control region is shown along the Y-axis. The red line shows the statistical significance value (logarithm of the odds (1- α')/α' where α' is Bonferroni corrected α=0.05 for 4 possible nucleotides in 40 positions (160 comparisons in total, α'=0.0003125)). Nucleotides whose height is over the significance value are significantly overrepresented.

It is worth noting that in the *cas1* mutant, statistically significant enrichment with 32-34-nt fragments associated with the 5'-AA/G-3' motif was detected in fragments mapped to the NT-strand upstream and T-strand downstream of the PPS (Figure 22). At the same time, no enrichment of the 3'-NNTTC-5' motif was revealed in complementary strands. In combination with the overall low amounts of spacer-size fragments in the *cas1* mutant (Figure 21), these results suggest that consensus PAM recognition occurs, albeit at very low efficiency, in cells expressing Cas1 H208A. However, in those rare cases when the PAM is recognized, the generation of PAM-derived 3' ends is prevented. Alternatively, PAM-derived 3' ends may be more susceptible to degradation in this strain.

It should be noted that CRISPR interference is active in the *cas1* self-targeting mutant and, therefore, products of DNA degradation by Cas3 were expected in the PPS-region. Based on the coverage profile around the PPS, one could expect that Cas3 degradation products would be most enriched immediately close to the PPS. Moreover, if Cas3-produced fragments were fuelling adaptation, their amount should be greater than the amount of prespacers, since several hundred kbp of gDNA are degraded during CRISPR interference while only one or very few 33-bp spacers are integrated into the CRISPR array in only about 20% of cells (Figure 14B). In FragSeq, all DNA fragments purified from cells are analyzed, most of them likely being nonspecific products of high-molecular-weight DNA degradation during purification. The distribution of such nonspecific fragments along the genome should be similar to the overall distribution of gDNA reads, while prespacers or products of CRISPR interference should produce a signal above the background in the PPS-containing area. Indeed, due to prespacer accumulation, a 34±2-fold increase of the overall amount of fragments within 10 kb of PPS is observed in the *wt* compared to the gDNA background (Figure 16B). Surprisingly, only a 3±0.2-fold difference is observed in the *cas1* mutant (Figure 16B). With the exception of the small amount of the already mentioned AAG-associated 32-34-nt fragments found in the *cas1* mutant, we could not find any other specific fragments in the 50-kbp region surrounding the PPS that either had lengths distinct from those detected in

other genomic regions, or contained PAM at certain positions (Figure 18, Figure 20, Figure 21A). These observations do not allow us to draw a clear conclusion about the Cas3 function in primed adaptation, since no Cas3-specific products can be detected. One possibility is that Cas3, indeed, produces fragments that are further bound by Cas1-Cas2 and protected from degradation. In a mutant *cas1* strain, Cas3-produced fragments are unprotected and, therefore, degraded fast. Another scenario is that in most *wt* cells, Cas3 degrades DNA to fragments shorter than 16 nt that are not detected by our method. However, in some cells, Cas3 in cooperation with Cas1-Cas2 ensure prespacer excision through some unknown mechanism. The role of Cas3 in this scenario may be the delivery of the Cas1-Cas2 complex to protospacers while the Cas3 nuclease activity may thus be unimportant for the overall process.

In fact, we found that both the 32-34-nt fragments derived from the NT-strand and associated with the 5'-AA/G-3' motif and the 36-38-nt fragments derived from the T-strand and containing the 3'-NNTTC-5' motif on their 3' ends were significantly enriched upstream of the PPS in the *cas3* nuclease mutant (Figure 22). We assumed that adaptation could still happen in the *cas3* nuclease mutant but at a very low level since no products corresponding to extended CRISPR arrays were observed after 30 cycles of PCR (data not shown). We cut out pieces of the gel where amplicons of extended arrays were supposed to migrate based on their molecular weights and subjected the purified DNA to additional 35 cycles of PCR. Products corresponding to extended CRISPR arrays were visible on a gel after this additional amplification (data not shown). HTS of the extended CRISPR arrays revealed that newly acquired spacers were selected from the PPS-containing region in a process of primed adaptation since both a high percentage of donor protospacers with the AAG PAM and their characteristic orientation bias were observed (Figure 24). However, the set of acquired spacers and their relative abundance was dramatically different from those observed in the *wt* strain. The first difference was that spacers were selected mainly from the region upstream of the PPS in the mutant (76.6±0.1% of spacers originated from the 100-kbp region upstream of the PPS while only 0.29±0.02% originated from the 100-kbp region downstream). Moreover, in contrast

to a gradual decrease in the efficiency of protospacer selection with ~100 kbp distance from the PPS characteristic of *wt*, a sharp drop in protospacer selection efficiency occured ~1.5 kbp upstream of the PPS in the *cas3* mutant. These results suggest that the Cas3 nuclease activity *per se* is not required for prespacer generation but ensures high efficiency of protospacer selection at large distances upstream and downstream of the PPS. However, this observation should be considered as preliminary since we cannot rule out a possibility that Cas3^{H74A} has residual nuclease activity. Further studies will be required to determine the mechanism underlying this phenomenon.

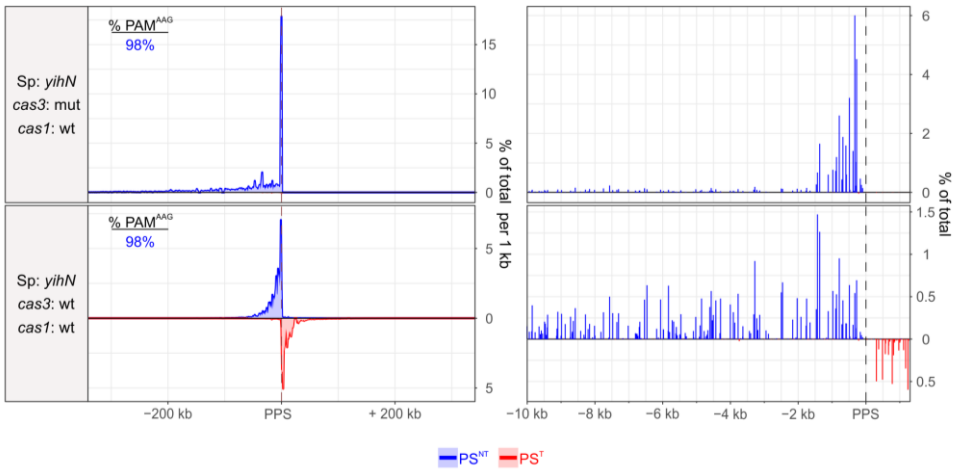


Figure 24. Comparison of spacers acquired during self-targeting in the *wt* and the *cas3* nuclease mutant strains. Plots on the left show the percentage of spacers per 1 kb of the 600-kb region centered at the PPS. The widths of blue and red lines reflect the mean±SEM values obtained in three biological replicates. Plots on the right show the percentage of individual spacers mapped within 10 kb upstream of the PPS. The 100% values on the Y-axis correspond to all spacers mapped to the genome. Spacers derived from PS^{NT} and PS^{T} are shown in blue and red, respectively.

Overall, our FragSeq results demonstrate that approximately spacer-sized PAM-associated DNA fragments accumulate in self-targeting cells from a ~50-kbp region around the PPS. These fragments may be adaptation intermediates from a step between protospacer selection and spacer integration. The presence of catalytically active Cas1 and Cas3 is required for their efficient excision from both DNA strands in the regions up- and downstream of the PPS. We also observed enrichment with fragments beyond the 50-kbp PPS-region in *wt* and *cas1* mutant strains (Figure 16B). We suggest that these fragments are produced due to the cleavage of gDNA by non-Cas cellular nucleases because the enriched regions contribute relatively few spacers.

4.1.4 Double-stranded oligonucleotides mimicking the structure of fragments detected in the self-targeting strain are efficiently integrated into the CRISPR array

To directly test whether fragments detected by FraqSeq could be integrated into CRISPR arrays, we performed a prespacer efficiency assay (Figure 25A) (Shipman et al., 2016). According to the published data, only double-stranded oligonucleotides can be integrated into the CRISPR array when electroporated into cells expressing *cas1* and *cas2* (Shipman et al., 2016). Four pairs of oligonucleotides mimicking the most abundant fragment types were used for transformation (Figure 25B). All pairs had a 3-4 nt 3'-overhang on the PAM-derived end and a blunt or nearly blunt PAM-distal end (Figure 25B). As a positive control, we used a 35 bp fully double-stranded oligo starting with the 5'-AAG-3'/3'-TTC-5' PAM which was integrated into ≈5% of CRISPR arrays in the experiments of Shipman et al (Shipman et al., 2016). As a negative control, we used a 33-bp double-stranded oligonucleotide starting with the G/C pair which was previously demonstrated to be integrated less efficiently and in both orientations (Shipman et al., 2016).

As has been previously demonstrated, transformation with the 35 bp control oligo led to its efficient integration as a properly processed spacer starting with the PAM-derived G in 10.2±1.2% of CRISPR arrays (Figure 25B). Similarly to the 35 bp control oligo, double-stranded PPS-region mimics having a blunt PAM-distal end and the PAM-derived end with the 3'-NNTT-5' or 3'-NNT-5' overhang were properly processed within the PAM-complementary sequence and integrated in the "correct" direct orientation into 9.2±0.1% and 8.6±0.1% of CRISPR arrays, respectively (Figure 25B). When the 3' terminus of the PAM-distal end was recessed by 1 nt, the efficiency of integration of the properly processed oligo dropped ≈50-fold. Remarkably, although one strand of the duplex was shortened by 1 nt, more than 85% of oligo-derived spacers retained the canonical length of 33 bp. This observation suggests that there should be a repair mechanism restoring the sequence of the shortened spacer strand by using the full-length strand as a template.

Overall, the results of prespacer efficiency assay confirm that fragments detected in the self-targeting strain undergoing primed adaptation can be efficiently integrated into the CRISPR array when they form duplex spacer precursors with a blunt PAM-distal end and a 3-4-nt 3’ overhang on the PAM-derived end.

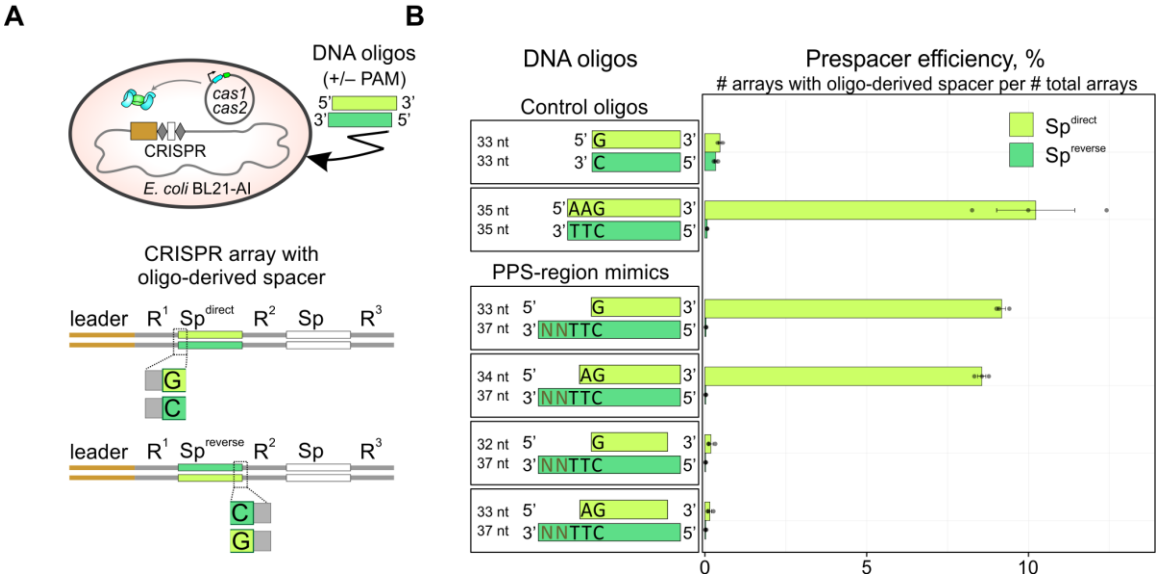


Figure 25. Double-stranded oligonucleotides mimicking the structure of fragments detected in the self-targeting strain are efficiently integrated into the CRISPR array. **A.** Prespacer efficiency assay. Top, introduction of synthetic DNA into cells containing a CRISPR array and a plasmid that directs expression of *cas1* and *cas2*. Bottom, integration of synthetic DNA into the CRISPR array occurs in either a direct (Sp^{direct}) or reverse ($Sp^{reverse}$) orientation. **B.** Results. Left, oligonucleotides analyzed. Right, percentage of arrays containing oligo-derived spacers having a direct (light green) or reverse (dark green) orientation (mean ± SEM of three biological replicates). This figure is published in (Shiriaeva et al., 2019).

4.1.5 Prespacers with a 3’ overhang on the PAM-derived end are formed by the type I-F self-targeting system

Previously Vorontsova et al. constructed an *E. coli* strain capable of self-targeting through the action of the type I-F CRISPR-Cas system from *Pseudomonas aeruginosa* (Figure 26A) (Vorontsova et al., 2015). Targeting of the genome by this system led to primed adaptation but, in contrast to the type I-E, there was a reversed strand bias: most spacers corresponded to PS^{T} upstream of the PPS and PS^{NT} downstream of the PPS. 2-bp 5’-CC-3’ PAM preceding protospacers was detected in the type I-F system. We set out to

test using FragSeq if prespacers with the structure similar to the structure of prespacers in the type I-E system can be detected in the type I-F system.

Similarly to the type I-E, we detected two major strand-specific types of short fragments accumulating in self-targeting cells around the PPS: 31-32 and 37-38 nt in length (Figure 26B). The shorter fragments mostly started with A and were generated as a result of incision between CC and the following A (65.2% and 51.8% in two biological replicates) (Figure 26B). The longer fragments contained the complementary 3'-GGT-5' motif plus three additional random nucleotides on the 3' end (34.6% and 32.1% in two biological replicates) (Figure 26B).

The presence of 5'-CC/A-3' and 3'-NNNGGT-5' motifs made us revise the PAM sequence in the type I-F. 95% of protospacers were preceded by 5'-CC-3' sequence. 63% of spacers started with A and 21% - with T. 5'-CCG-3' and 5'-CCC-3' were underrepresented (9% and 2%, respectively). Therefore, we conclude that *P. aeruginosa* PAM is 5'-CCW-3' rather than the previously reported 5'-CC-3'. A similar preference for A or T in the first position of spacers was described for the type I-F system of *Pectobacterium atrosepticum* (Fagerlund et al., 2017; Staals et al., 2016).

Taken together, our results demonstrate that despite the opposite orientation bias in spacer acquisition, processing of prespacers in type I-E and type I-F involves intermediates of similar structure with 3' overhangs on PAM-derived ends (Figure 26C).

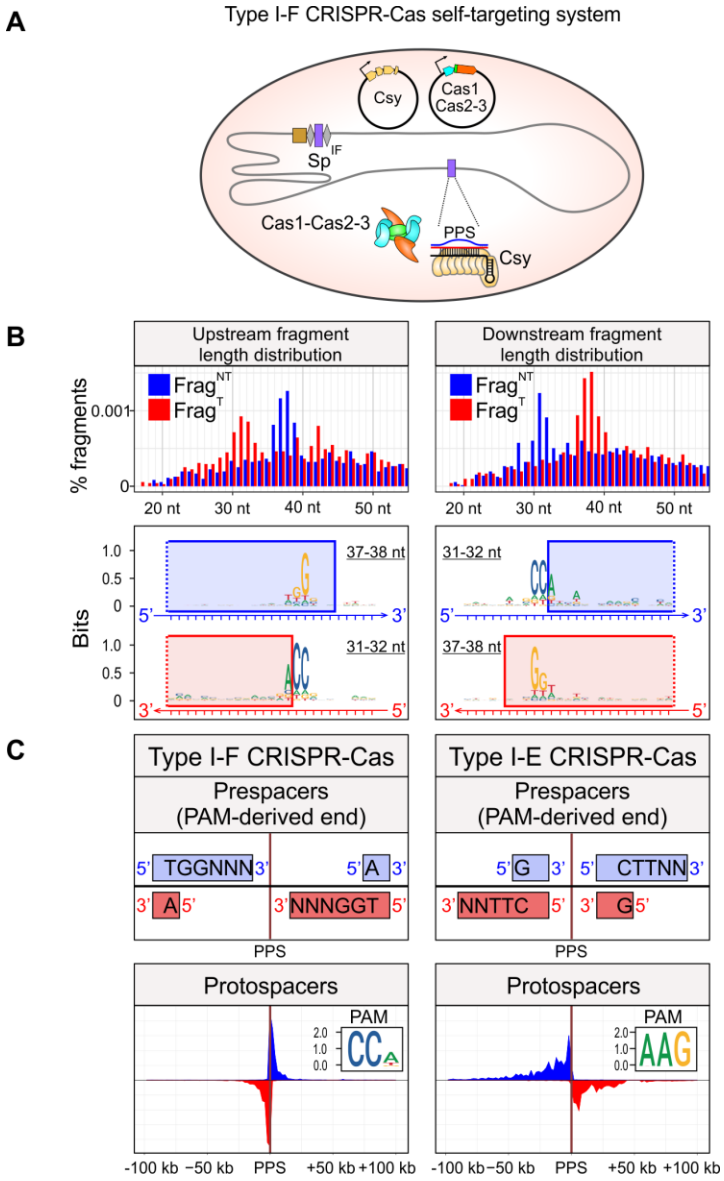


Figure 26. Comparison of prespacers and spacers formed by the type I-F and I-E CRISPR-Cas self-targeting systems. **A.** Components of type I-F CRISPR-Cas self-targeting system. Shaded oval, an *E. coli* cell; grey line, chromosomal DNA; black line, plasmid DNA; orange, tan, blue, and green pentagons, *cas* and *csy* genes; brown rectangle, array leader sequence; grey diamonds, array repeat sequences; purple rectangles, spacer and chromosomal PPS targeted by spacer-derived crRNA; Csy, type I-F effector complex; Cas1-Cas2-3, complex of Cas1 and Cas2-3 proteins. **B.** FragSeq results: length distributions of fragments (top, the 100% value on the Y-axis corresponds to all fragments mapped to the genome) and sequence features of PPS-region sequences from which fragments are derived (bottom). Logos for 31-32-nt fragments were generated by aligning sequences 10-nt upstream to 15-nt downstream of the fragment 5' end. Logos for 37-38-nt fragments were generated by aligning sequences 20-nt upstream to 5-nt downstream of the fragment 3' end. Blue rectangles, sequences present in $Frag^{NT}$; red rectangles, sequences present in $Frag^{T}$. **C.** Comparison of PPS-region-derived fragments and PPS-region protospacers in type I-F and type I-E self-targeting systems. Inset, logo derived from alignment of PPS-region PAMs. This figure is published in (Shiriaeva et al., 2019).

4.2 DNA repair enzymes are involved in CRISPR interference and primed adaptation in the type I-E CRISPR-Cas system

Some results presented in section 4.2 are published in:

Kurilovich, E., **Shiriaeva, A.**, Metlitskaya, A., Morozova, N., Ivancic-Bace, I., Severinov, K., and Savitskaya, E. (2019). Genome Maintenance Proteins Modulate Autoimmunity Mediated Primed Adaptation by the Escherichia coli Type I-E CRISPR-Cas System. Genes *10*, 872.

Self-targeting strains with deletions of DNA repair genes were constructed by A. Metlitskaya and Elena Kurilovich. The author performed all experiments with bacterial cultures presented in this chapter. The gel presented in Figure 27A was prepared by Elena Kurilovich.

The author purified total genomic DNA and short DNA fragments for high-throughput sequencing; the libraries were prepared and sequenced by Waksman Genomics Core Facility, Rutgers University, USA. The author performed the analysis of high-throughput sequencing data for all data presented in Figure 28, Figure 29, Figure 31 - Figure 35. The analysis presented in Figure 30 was done by S. Medvedeva.

The author's work described in this chapter was performed in Skoltech Research Center in the center of nano- and biotechnologies of Peter the Great St. Petersburg Polytechnic University, Russia and in Konstantin Severinov laboratory at Waksman Institute of Microbiology, Rutgers University, USA.

4.2.1 The RecBC helicase and RecJ nuclease participate in the processing of type I-E prespacer 5’ ends

Our FragSeq results demonstrate that prespacers in the type I-E system of *E. coli* are asymmetric with two 5’ ends and the PAM-distal 3’ end being trimmed to the length of a mature spacer and the 3’ end on the PAM-derived side containing four additional nucleotides: 3’-NNTT-5’. The enzymes trimming prespacer 3’ ends have been studied *in vitro* by two research groups but no *in vivo* results have been reported (Kim et al., 2020; Ramachandran et al., 2020). Nothing is known about trimming of prespacer 5’ ends except that Cas1 is unlikely to be responsible for this cleavage since 5’ ends do not enter the Cas1 active sites (Nuñez et al., 2015b; Wang et al., 2015).

A set of 9 self-targeting strains with deletions of genes involved in DSB repair was obtained in our laboratory (Kurilovich et al., 2019). We demonstrated that primed adaptation efficiency is decreased in *ΔrecJ*, *ΔrecB ΔrecJ*, and *ΔrecB ΔsbcD* mutants (Figure 27A). In all strains, CRISPR interference occurred as efficiently as in the *wt* strain and the number of CFUs was decreased by approximately 3 orders of magnitude upon activation of *cas* gene expression (Figure 27B). This means that the decrease in adaptation is either caused by the involvement of RecBCD, RecJ, and SbcCD in the spacer acquisition process or by indirect effects of DNA repair deficiency on cell fitness. The latter is consistent with the decreased cell viability observed in the *ΔrecB ΔrecJ* and *ΔrecB ΔsbcD* strains in the absence of *cas* genes inducers (Figure 27B). However, the *recJ* deletion did not have any impact on cell survival suggesting that there might be direct involvement of at least RecJ nuclease in spacer acquisition.

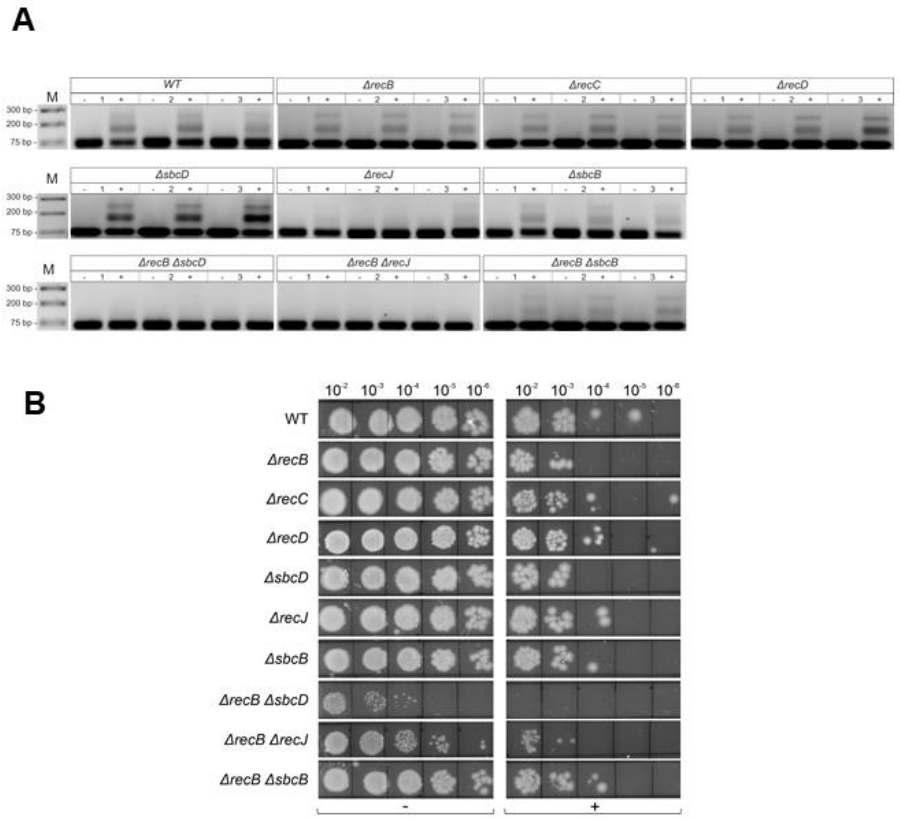


Figure 27. Primed adaptation and CRISPR interference in DNA repair mutant derivatives of the self-targeting *wt* strain. **A**. Products of amplification of CRISPR arrays with 0 (lower bands), 1 or 2 additional spacers (upper bands) obtained in three independent experiments (1, 2, 3) with (+) or without (-) activation of self-targeting. **B.** Growth of *wt* and mutant strains. The cultures were grown to the same optical density, incubated for 5 hours with (+) or without (-) *cas* genes inducers and then aliquots of cultures serial dilutions (indicated at the top) were spotted on the surface of LB agar plates. This figure is published in (Kurilovich et al., 2019).

To test if decreased production of prespacers or a difference in their lengths could be the reason for reduced adaptation in the mutants, we sequenced short DNA fragments present in each of the mutant strains and compared them with the fragments present in the *wt*. Mapping of 30-45-nt fragments to the genome showed a peak of coverage centered at the PPS in all mutant strains except for the double *ΔrecB ΔrecJ* mutant (Figure 28). Length distributions of the fragments mapped within 50 kbp of the PPS revealed two groups of fragments centered at 32-34 nt or 36-38 nt depending on the strand and position relative to the PPS (Figure 29A). Sequence analysis of the two groups of fragments separately for the upstream and downstream regions demonstrated the presence of the PAM and the 4-nt overhang on the PAM-derived 3’ end (Figure 29B). No difference in

the lengths of fragments mapped to the opposite strands and no PAM sequences were found in the *ΔrecB ΔrecJ* mutant (Figure 29A,B). These results highlight that at least one of two enzymes, RecBCD or RecJ should be present in a cell for prespacer generation.

It should be noted that higher amounts of 36-38-nt fragments were revealed in the experiment involving DNA repair mutants compared with our initial experiments analyzing fragment length distributions in *wt*, *cas1*, and *cas3* mutants where the shorter 32-34-nt fragments were more abundant (Figure 21A, Figure 29A). The libraries of fragments presented in Figure 28 and Figure 29 were prepared using a modified FragSeq protocol where short DNA fragments were purified as usual, but the libraries were prepared with a commercial kit for strand-specific DNA sequencing "Accel-NGS 1S DNA Library Kit" (Swift Biosciences). The advantage of this approach is that library construction takes only 2 hours but the drawback is that a tail of random nucleotides (mostly C and T) is attached to fragments' 3' ends prior to adapter ligation making the precise mapping of the 3' ends not possible. We assume that the differences in the ratio of the shorter and longer fragments revealed by the two library construction approaches are caused by the biases introduced at the stage of adapter ligation. In fact, we discovered that one of the ligases used in our original version of the FragSeq protocol, the 5´ App DNA/RNA Ligase, prefers to ligate 3' ends with a 3'-NGC-5' motif. The C nucleotide of this motif is in the same position (the third nucleotide from the 3' end) as the T of the 3'-NNTTC-5' motif found in prespacers. This could have likely led to the reduction of 3'-NNTTC-5'-containing prespacers detected in our initial experiments (Figure 21A). Alternatively, it could be that the adapter ligation by the Accel-NGS 1S kit is more efficient for CT-rich 3' ends. These possibilities have not been tested and will be explored in our subsequent studies. In any case, we do not make any conclusions about the relative abundance of the shorter 32-34-nt and longer 36-38-nt fragments and compare the results only obtained within a single experiment where all libraries were prepared using the same method.

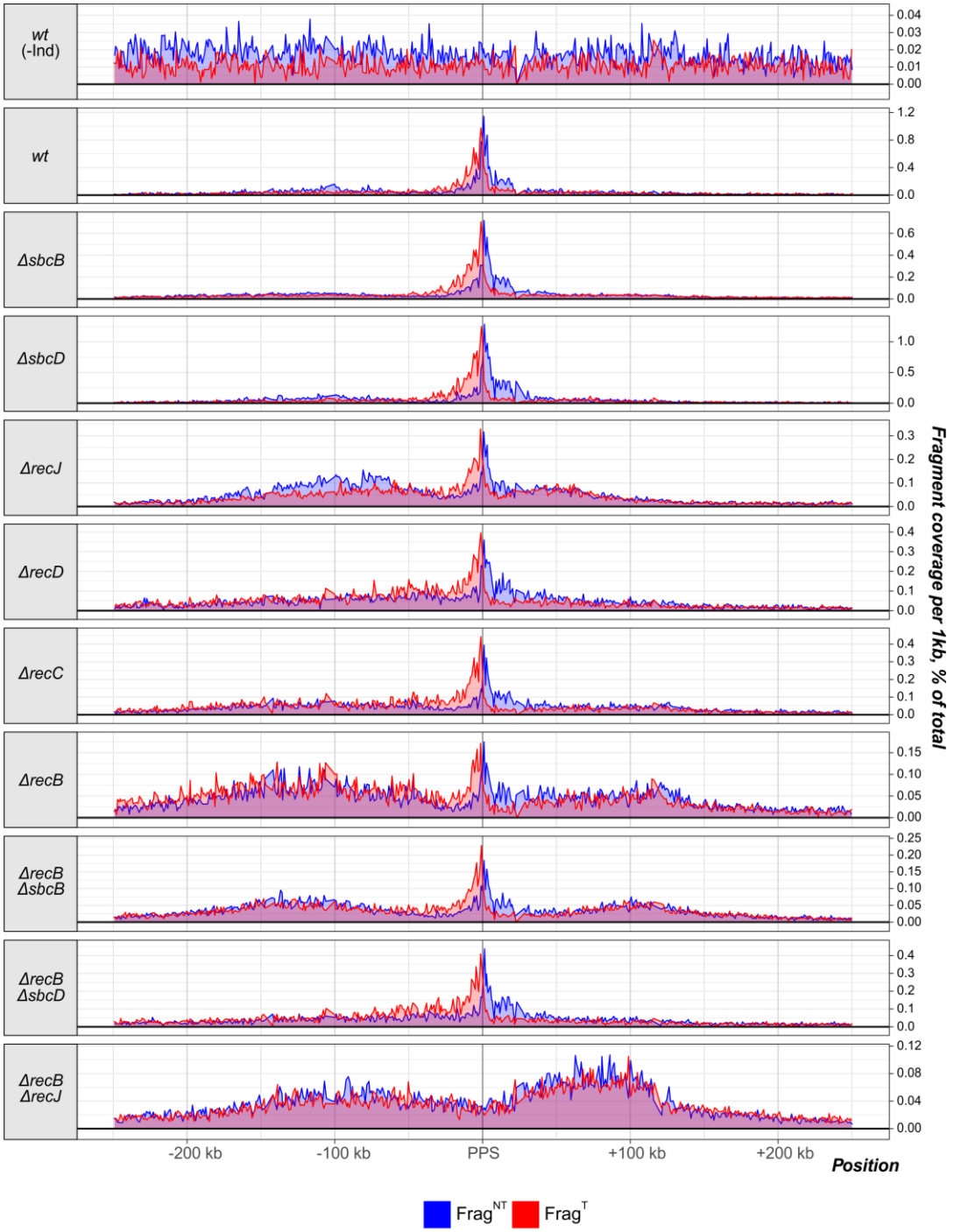


Figure 28. Coverage plots for 30-45-nt fragments purified from the *wt* self-targeting strain or its DNA repair mutant derivatives. The percentage of total coverage per 1 kb for the indicated strains is shown (the 100% value on the Y-axis corresponds to the total coverage with all fragments mapped to the genome). Coordinates on the X-axis represent positions on the chromosome with respect to the PPS. Blue, nontarget-strand-derived fragments ($Frag^{NT}$); red, target-strand-derived fragments ($Frag^{T}$).

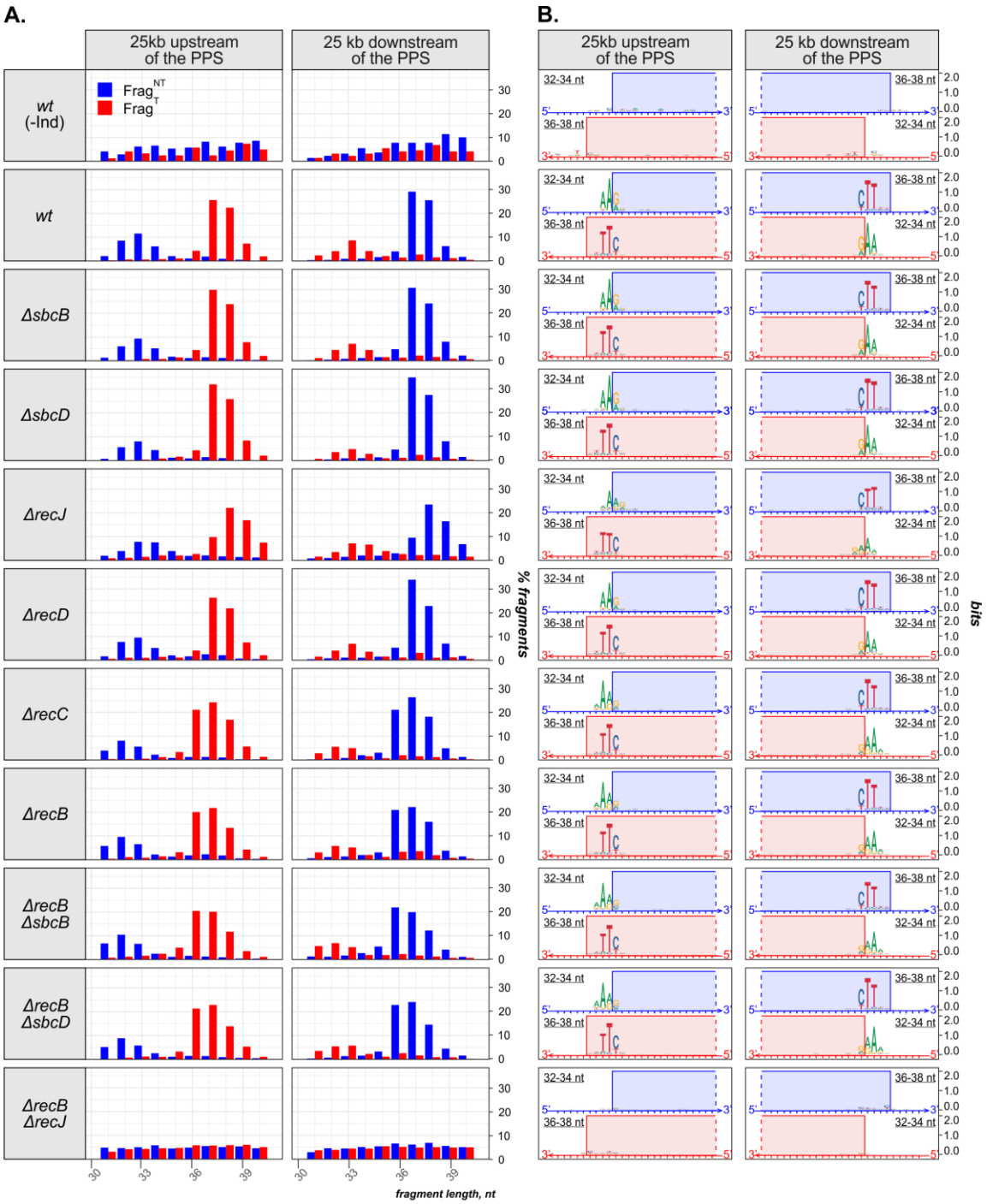


Figure 29. Analysis of lengths and sequences of fragments purified from the *wt* self-targeting strain or its DNA repair mutant derivatives. **A.** Length distributions of 31-40-nt fragments mapped within 25 kbp upstream or 25 kbp downstream of the PPS. Fragments mapped to the NT-strand ($Frag^{NT}$) and the T-strand ($Frag^{T}$) are shown in blue and red, respectively. **B.** Sequence logos of genomic DNA from which PPS-region fragments are derived. Blue rectangles, sequences present in $Frag^{NT}$; red rectangles, sequences present in $Frag^{T}$.

Plotting the lengths of fragments detected in the DNA repair mutants suggested that the distributions are shifted in some of them compared with the *wt* strain (Figure 29A). In particular, while 37-nt fragments constituted the major fraction mapped to the T-strand upstream and the NT-strand downstream of the PPS in *wt*, 38-nt fragments were predominant in the *ΔrecJ* mutant (Figure 29A). On the contrary, higher percentages of 36-nt fragments were observed in libraries prepared from the *ΔrecB*, *ΔrecC*, *ΔrecB ΔsbcD*, and *ΔrecB ΔsbcB* strains (Figure 29A). To quantitatively compare prespacer length distributions further in different strains, we calculated the fractions for each fragment length within smaller, 31-35-nt and 36-40-nt subgroups. Using the Pearson correlation (1-cor) as a distance metric, we built dendrograms of strains by hierarchical complete-linkage clustering. To test if the lengths of fragments in a sample of interest are greater or lower than in the *wt* strain, one-sided Mann–Whitney U tests were applied (Figure 30).

The results univocally demonstrate that single deletions of the *sbcB* or *sbcD* genes do not change fragment length distributions. A single deletion of *recJ* leads to the generation of longer spacer precursors, and this applies to both fragment subgroups (Figure 30). The deletions of the *recB* or *recC* gene, on the contrary, decrease the average length of both subgroups of the fragments (Figure 30). Interestingly, the *ΔrecD* mutant is clustered together with the *wt* but not the *ΔrecB* and *ΔrecC* strains. This observation suggests that the presence of the RecBC helicase activity is sufficient to ensure the proper processing of prespacers, while the RecBCD nuclease activity is not required.

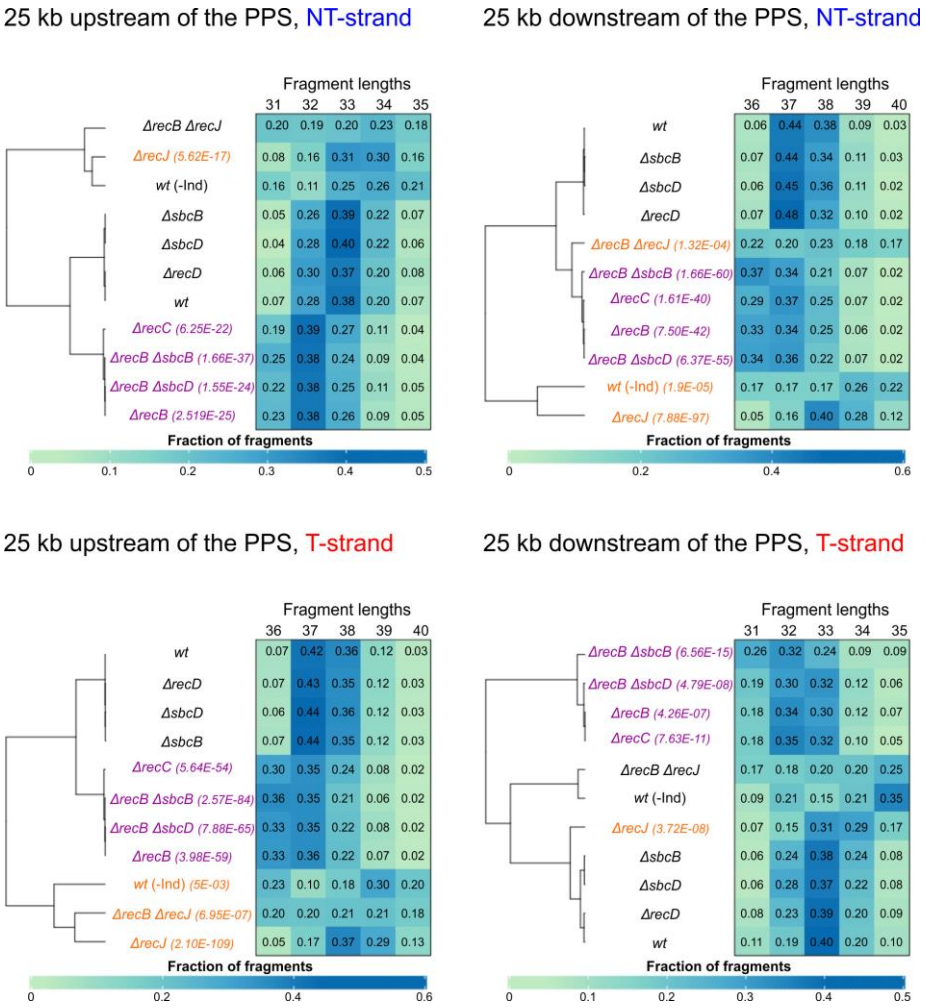


Figure 30. Clusterization of the *wt* self-targeting strain and its DNA repair mutant derivatives by hierarchical complete-linkage with the Pearson correlation (1-cor) between the fragment length distributions used as a distance metric. Shading represents the fraction of fragments where the total is a sum of 31-35-nt or 36-40-nt fragments mapped to one strand up- or downstream of the PPS (the values for each length are also indicated). In samples whose names are shown in purple, fragments lengths were significantly shorter than in the *wt*. In samples whose names are shown in orange, fragment lengths were significantly longer than in the *wt*. In brackets, p-values of one-sided Mann Whitney U tests computed in comparisons with the *wt* are shown. Only the p-values less than 0.05 after the Bonferroni correction for multiple comparisons are shown.

Changes in fragment length distributions can be caused by greater or less extensive trimming of 5’ ends, 3’ ends, or both types of ends in the mutants. In addition, the PAM-distal and PAM-derived ends can be processed in a different manner. To determine the possible source of differences in fragment lengths in the *recB*, *recC*, and *recJ* mutants, we analyzed fragments, which could have been produced during prespacer excision. To do that, we determined genomic positions of all possible “ideal” 33-bp

protospacers with an adjacent consensus 5'-AAG-3' PAM located in the NT-strand within the 25-kbp region upstream of the PPS or in the T-strand within the 25-kbp region downstream of the PPS. Such "ideal" 33-bp protospacers serve as spacer donors during primed adaptation and therefore prespacers mapping to approximately the same genomic positions are expected. However, prespacer ends do not have to necessarily coincide with the boundaries of "ideal" protospacers: greater or less extensive trimming may occur. To account for greater trimming, we selected for analysis those fragments that spanned at least the central 23-nt parts of possible "ideal" protospacers. In the crystal structure of Cas1-Cas2 bound to a 33-bp oligonucleotide, the central 23-bp region is in the form of dsDNA, while the terminal 5-bp regions on both protospacer sides are unwound and are likely to be more exposed for degradation (Nuñez et al., 2015b; Wang et al., 2015). Only fragments of 31-40 nt were analyzed. We calculated the distances from the 5' and 3' ends of selected fragments to the boundaries of "ideal" protospacers and plotted the distributions of these distances (Figure 31).

The distributions of the distances revealed that both 5' ends in the *wt*, *ΔsbcB*, *ΔsbcD,* and *ΔrecD* strains coincided with the "ideal" 5' ends (distance=0) in most fragments. The second largest fraction of fragments had one additional nucleotide on the 5' ends (distance = +1) (Figure 31). In the *ΔrecJ* mutant, most fragments had both 5' ends longer by one nucleotide compared with the "ideal" ends. One-sided Mann Whitney U test results suggest that the distributions of the distances calculated for the 5' ends in the *wt* and *ΔrecJ* mutants are significantly different. Significant differences were also revealed for the *ΔrecB, ΔrecC, ΔrecB ΔsbcB*, and *ΔrecB ΔsbcD* mutants but in this case fragments' 5' ends were trimmed more than in *wt*: the fraction of fragments with the 5' ends truncated by 1 nucleotide was comparable or even greater than the fraction corresponding to the "ideal" positions. It should be noted that the shape of the distributions calculated for the PAM-distal and PAM-derived 5' ends within each of four RecBC-deficient strains was different suggesting that the recognition of PAM influences the trimming of the 5' ends but only when no RecBC helicase is present in a cell.

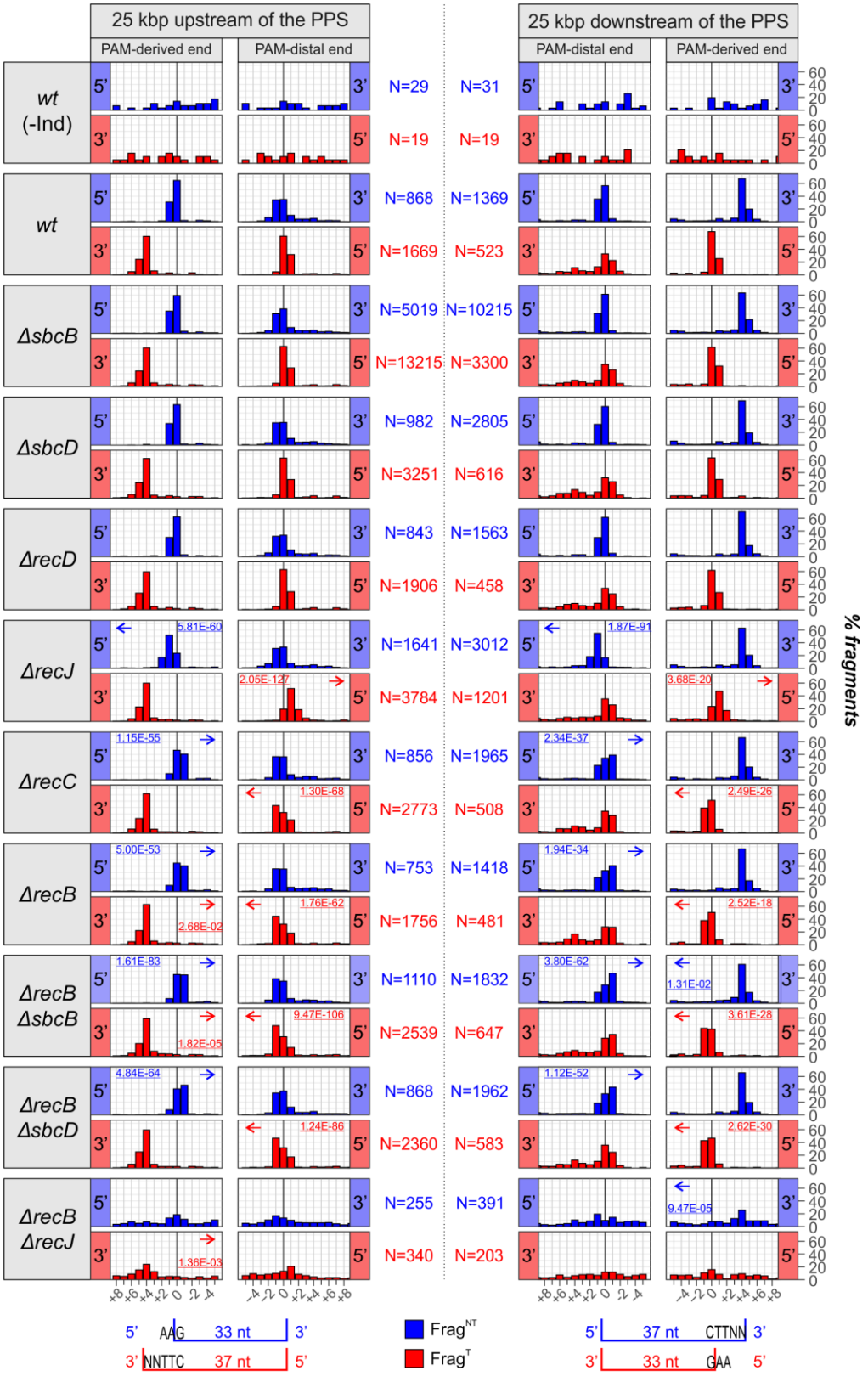

25 kbp upstream of the PPS
PAM-derived end
PAM-distal end
25 kbp downstream of the PPS
PAM-distal end
PAM-derived end
wt (-Ind)
wt
ΔsbcB
ΔsbcD
ΔrecD
ΔrecJ
ΔrecC
ΔrecB
ΔrecB ΔsbcB
ΔrecB ΔsbcD
ΔrecB ΔrecJ
N=29
N=19
N=868
N=1669
N=5019
N=13215
N=982
N=3251
N=843
N=1906
N=1641
N=3784
N=856
N=2773
N=753
N=1756
N=1110
N=2539
N=868
N=2360
N=255
N=340
N=31
N=19
N=1369
N=523
N=10215
N=3300
N=2805
N=616
N=1563
N=458
N=3012
N=1201
N=1965
N=508
N=1418
N=481
N=1832
N=647
N=1962
N=583
N=391
N=203
5.81E-60
2.05E-127
1.87E-91
3.68E-20
1.15E-55
1.30E-68
2.34E-37
2.49E-26
5.00E-53
2.68E-02
1.76E-62
1.94E-34
2.52E-18
1.61E-83
1.82E-05
9.47E-106
3.80E-62
1.31E-02
3.61E-28
4.84E-64
1.24E-86
1.12E-52
2.62E-30
1.36E-03
9.47E-05
% fragments
5' AAG 33 nt 3'
3' NNTTC 37 nt 5'
Frag^NT
Frag^T
5' 37 nt CTTNN 3'
3' 33 nt GAA 5'

Figure 31. Distributions of distances from "ideal" ends of protospacers to the ends of experimentally observed fragments in various *E. coli* strains. Zero values on the X-axis indicate that fragments' ends coincide with the "ideal" protospacer ends. Positive values on the X-axis indicate the number of nucleotides by which observed fragments' ends are longer than "ideal" protospacer ends. Negative values on the X-axis indicate the number of nucleotides by which observed fragments' ends are shorter than "ideal" protospacer ends. N, number of 31-40-nt fragments containing protospacer sequences mapped within 25 kbp upstream of the PPS (left) or 25 kbp downstream of the PPS (right). Shown in blue are fragments mapped to the NT-strand ($Frag^{NT}$); shown in red are fragments mapped to the T-strand ($Frag^{T}$). Underlined values within individual plots are p-values of one-sided Mann Whitney U tests computed in comparisons with the *wt*. Only the p-values less than 0.05 after the Bonferroni correction for multiple comparisons are shown. The arrows indicate the direction of the shift along the X-axis in samples compared with the *wt*.

The PAM-derived and PAM-distal 3' ends were processed differently within strains in which prespacers were produced (*wt*, all single deletion mutants, double mutants *ΔrecB ΔsbcB* and *ΔrecB ΔsbcD*). The PAM-derived end contained the 3'-NNTTC-5' sequence on the 3' end of most fragments or 3'-NNNTTC-5' in the second largest fraction, which corresponds to positions +4 and +5 from "ideal" protospacer ends, respectively (Figure 31). PAM-distal 3' ends either coincided with the "ideal" positions (distance=0) or were truncated by 1 nucleotide (distance = -1) in all analyzed strains. Interestingly, no differences in the processing of 3' ends compared with the *wt* were revealed by Mann Whitney U test for most mutants. Statistically significant differences were detected only for the PAM-derived 3' ends of the *ΔrecB* strain in the upstream region (p=2.68E-02), the *ΔrecB ΔsbcB* mutant in both the upstream (p=1.82E-05) and the downstream (p=1.31E-02) regions, and the *ΔrecB ΔrecJ* mutant also in the upstream (p=1.36E-03) and the downstream (p=9.47E-05) regions. The inconsistency between the PAM-derived 3' ends up- and downstream of the PPS in the *ΔrecB* mutant, a relatively high p-value and the absence of any effect of the *ΔrecC* mutation (which usually has the same phenotype as the *ΔrecB* mutation) suggest that the detected differences might be random. We can not rule out that the simultaneous deletions of *recB* and *sbcB* influence 3' end trimming (which is in line with the SbcB 3'→5' exonuclease activity) but if so, the effect is very marginal since ≈80% of all fragments were trimmed after positions +4 or +5 in both *wt* and *ΔrecB ΔsbcB* regardless of the region. The differences detected for the *ΔrecB ΔrecJ* could have been caused by the much lower efficiency of prespacer generation and higher background of non-prespacer fragments. An approach with higher

specificity to prespacers like ChIP-seq with anti-Cas1 antibodies will be required in future experiments to test if prespacers are produced in the *ΔrecB ΔrecJ* mutant.

Summarizing the FragSeq results for prespacers generated in various DNA repair mutants we conclude that the RecBC helicase and RecJ exonuclease are critical for the proper prespacer processing. Inactivation of either of these two enzymes results in changed fragment lengths due to more or less trimming of 5' ends. Simultaneous inactivation of both enzymes dramatically reduces prespacer generation abolishing spacer acquisition.

4.2.2 Degradation of DNA regions adjacent to the PPS is initiated by CRISPR interference machinery but continued by RecBCD and SbcCD nucleases

The targeting of the genome by the type I-E CRISPR-Cas system results in extensive loss of DNA surrounding the PPS (Figure 12B, Figure 13). As was discussed in chapter 4.1.1, this degradation can be visualized in *wt* as a gradual decrease of total genomic DNA coverage that starts approximately 200 kb upstream and 100 kb downstream of the PPS and reaches the minimal value in the immediate vicinity of the PPS (Figure 12B, Figure 13A). Free DNA ends are substrates for cellular non-Cas nucleases like RecBCD, SbcB, SbcCD, and RecJ. It can thus be hypothesized that once Cas3 dissociates from DNA, a free DNA end gets bound by other nucleases that continue the degradation.

In a previous work performed in our laboratory, Elena Kurilovich demonstrated that a gap in genomic coverage of the *ΔrecC, ΔrecB*, or *ΔrecD* mutant is narrower than in the *wt* strain suggesting that RecBCD broadens the gap produced by Cas3 (unpublished data). An opposite effect was demonstrated for the *ΔsbcB* mutant. The *ΔsbcD* mutation had no impact on coverage. The genome coverage profiles of other mutants undergoing self-targeting analyzed in this thesis research have not been explored so far. Therefore, we sequenced genomic DNA of nine DNA repair mutants from which we isolated prespacers discussed in chapter 4.2.1 and used the *wt* strain with and without *cas* genes induction as controls.

A gap in coverage near the PPS was observed in all induced cultures, indicative of active CRISPR interference (Figure 32). The overall shapes of the genome coverage profiles were different in the mutants making it difficult to directly compare the sizes of the gaps formed in the PPS-region between the strains. To account for the differences, we multiplied the normalized coverage in each strain by a coefficient to make the mean coverage in a region from 250 kbp to 200 kbp upstream of the PPS equal to the mean coverage of the same region in the induced *wt* sample. The shapes of the curves in the 200-kbp region upstream of the PPS were next compared pairwise (Figure 33).

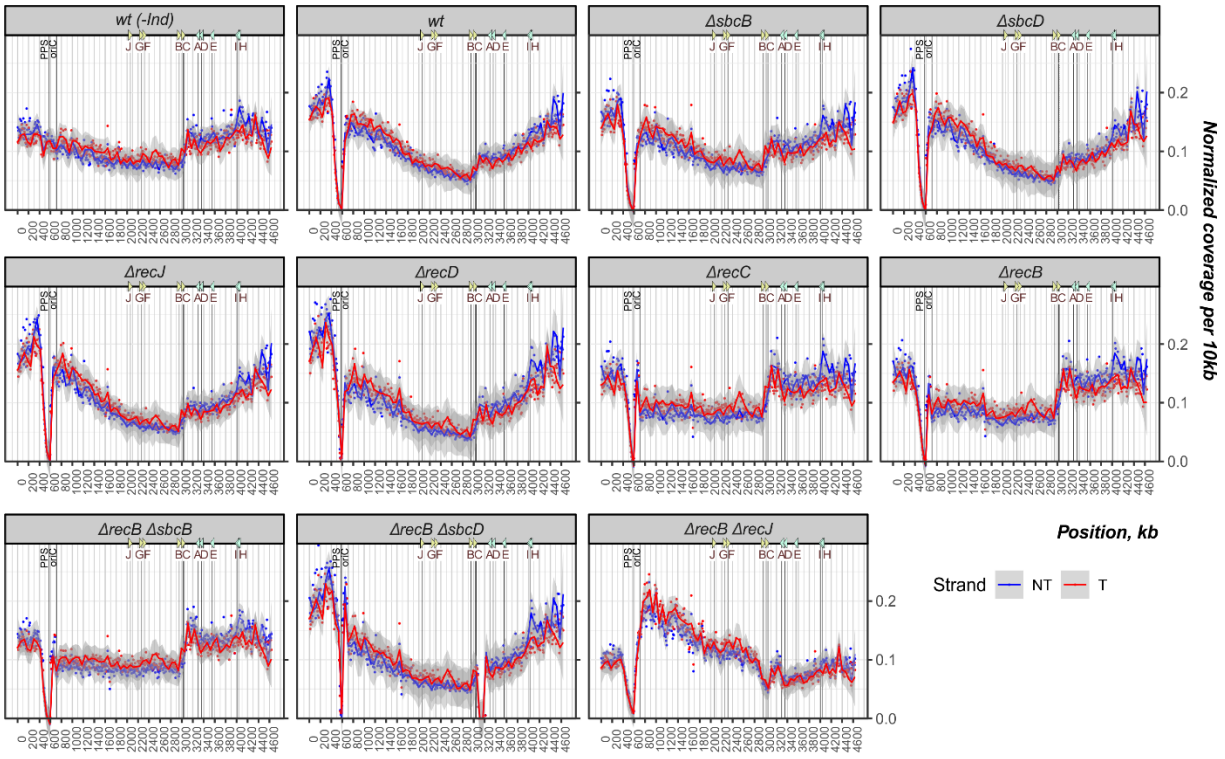


Figure 32. Effects of deletions of DNA repair genes on genomic DNA content. Graph of sequence coverage per 10 kb in the indicated strains. *oriC*, site of replication origin; PPS, priming protospacer; A-H, sites of replication termination *terA* – *terH* (*ter* sites stopping the replication fork progressing from left to right are shown in light blue; *ter* sites stopping the replication fork progressing from right to left are shown in yellow). Red and blue lines, Loess smoothing of normalized coverage per 10 kb for the NT and T strands, respectively; grey shading, 95% confidence interval. The 100% value on the Y-axis corresponds to the total coverage with all genomic DNA reads mapped to the genome.

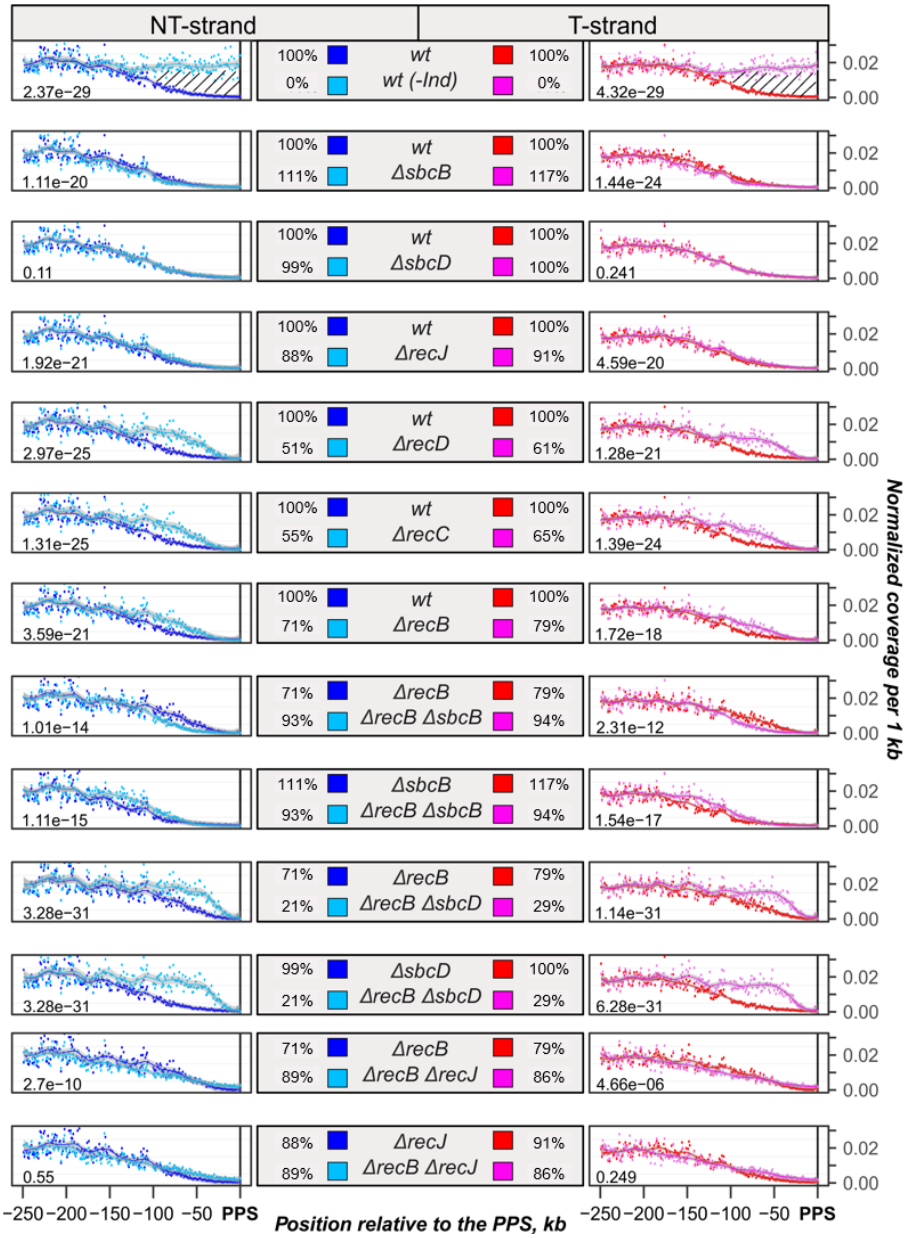


Figure 33. Effects of deletions of DNA repair genes on degradation of DNA upstream of the PPS. Pairwise comparisons of the coverage in two samples indicated in the central panel are shown. Left, coverage on the NT-strand; right, coverage on the T-strand. To compare the shapes of the curves, the coverage in the 250-kbp region upstream of the PPS was adjusted to make the mean coverage in the 50-kbp region from 250 kbp to 200 kbp upstream of the PPS identical to the mean coverage of the induced *wt* sample in the same region. The adjusted coverage was compared between every two-sample set in the 200-kbp region immediately upstream of the PPS. The numbers in the lower left corners indicate the Benjamini & Hochberg adjusted p-values (Benjamini et al., 2009) for paired two-sided Wilcoxon tests. The differences were considered significant if the BH-adjusted p-values were less than 0.01. The values in the middle panels show the percentage of DNA degraded in a given sample compared with the gap in the induced *wt* set at 100% and shown as a shaded area in the upper panel. The size of the gap was calculated as the difference in the total coverage of the given sample with inducers and *wt* without inducers in the 200-kbp region immediately upstream of the PPS.

Our results confirm that RecBCD enhances target degradation since the amount of DNA degraded near the PPS is reduced by 21-49% in the absence of the RecBCD nuclease activity (Figure 33). Very limited degradation was observed in the *ΔrecB ΔsbcD* mutant (21-29% of the degradation in *wt*), suggesting that in the absence of RecBCD, the SbcCD complex facilitates degradation of the regions extending beyond the gap produced by the CRISPR interference machinery (Figure 33).

In line with the results of Kurilovich et al., the *ΔsbcD* deletion did not influence degradation, while the *ΔsbcB* mutation enlarged the gap by 11-17% (Figure 33). The comparison of the genome coverage profiles in the double mutant *ΔrecB ΔsbcB* with the single mutant *ΔrecB* showed that the *ΔsbcB* mutation enlarged the gap of the *ΔrecB* strain by 15-22%. Since the effect of the *ΔsbcB* mutation on the *wt* and *ΔrecB* was similar, it suggests that the inhibition of degradation by SbcB is independent of RecBCD. A result similar to ours was obtained by the Bikard laboratory where an enlarged gap in genomic DNA coverage was detected in an *E. coli ΔsbcB* strain with induced CRISPR-Cas9 cleavage compared with a similar strain expressing wild-type *sbcB* (Gutierrez et al., 2018).

Controversial results were obtained for the *ΔrecJ* mutation, which seems to reduce degradation by 9-12% when introduced into the *wt* but enlarges the gap when introduced into the *ΔrecB* mutant by 7-18%. Our data are preliminary and further studies will be required to test whether these effects are reproducible and what the mechanisms could be.

4.2.3 Fragments generated by RecBCD and an unknown nuclease are detected in the regions flanking the primary area of DNA degradation by the CRISPR interference machinery

RecBCD is a nuclease/helicase that produces DNA fragments of a length varying between several nucleotides – several thousand nucleotides depending on *in vitro* reaction conditions (Goldmark and Linn, 1972; Karu et al., 1973; MacKay and Linn, 1974; Wright et al., 1971). Though it has been long known that RecBCD degrades linear

dsDNA, the products of *in vivo* cleavage have not been characterized by high-throughput sequencing.

FragSeq analysis of the *wt* and DNA repair mutant strains allowed us to detect not only spacer-size intracellular fragments but also longer fragments from 46 to 500 nt (Figure 34). To account for the background that could be produced due to mechanical breakage of high-molecular-weight genomic DNA during its purification we superimposed the profiles of fragment coverage with the profiles of total genomic DNA purified from the same sample and subtracted the latter from the former. Figure 35A shows the results of this normalization for the region surrounding the PPS. If DNA fragments are produced from some genomic regions with high efficiency, the coverage with fragments should be higher than the coverage with genomic DNA in these regions.

We detected two regions highly enriched with fragments, one upstream ("up") and another downstream ("dw") of the ≈50-kbp region centered at the targeted protospacer in the *wt*, *ΔsbcB*, *ΔsbcD*, and *ΔrecJ* strains (Figure 35A). The peaks disappeared in the *ΔrecD* mutant suggesting that RecBCD nuclease activity is required for the generation of these fragments. Length distributions were different for the fragments mapped to the T- and NT-strands. In all four strains, fragments mapped to the T-strand were on average longer than the fragments mapped to the NT-strand in the "upstream" peak (Figure 35B). The NT-strand corresponds to the strand with a free 3' terminus produced after the initial cleavage at the PPS and further degradation upstream. More frequent cutting of the 3'-terminated strand compared with the 5'-terminated strand is consistent with *in vitro* results of DNA degradation by RecBCD before encountering a Chi site (Taylor and Smith, 1995b). Following this line of reasoning, an opposite strand bias should be observed in the "downstream" peak where the 3' terminated strand corresponds to the T-strand. Indeed, longer fragments were produced from the NT-strand in the *wt*, *ΔsbcD*, and *ΔrecJ* strains (Figure 35A,B). However, longer fragments corresponding to the T-strand were observed in the "downstream" peak of the *ΔsbcB* mutant (Figure 35A,B).

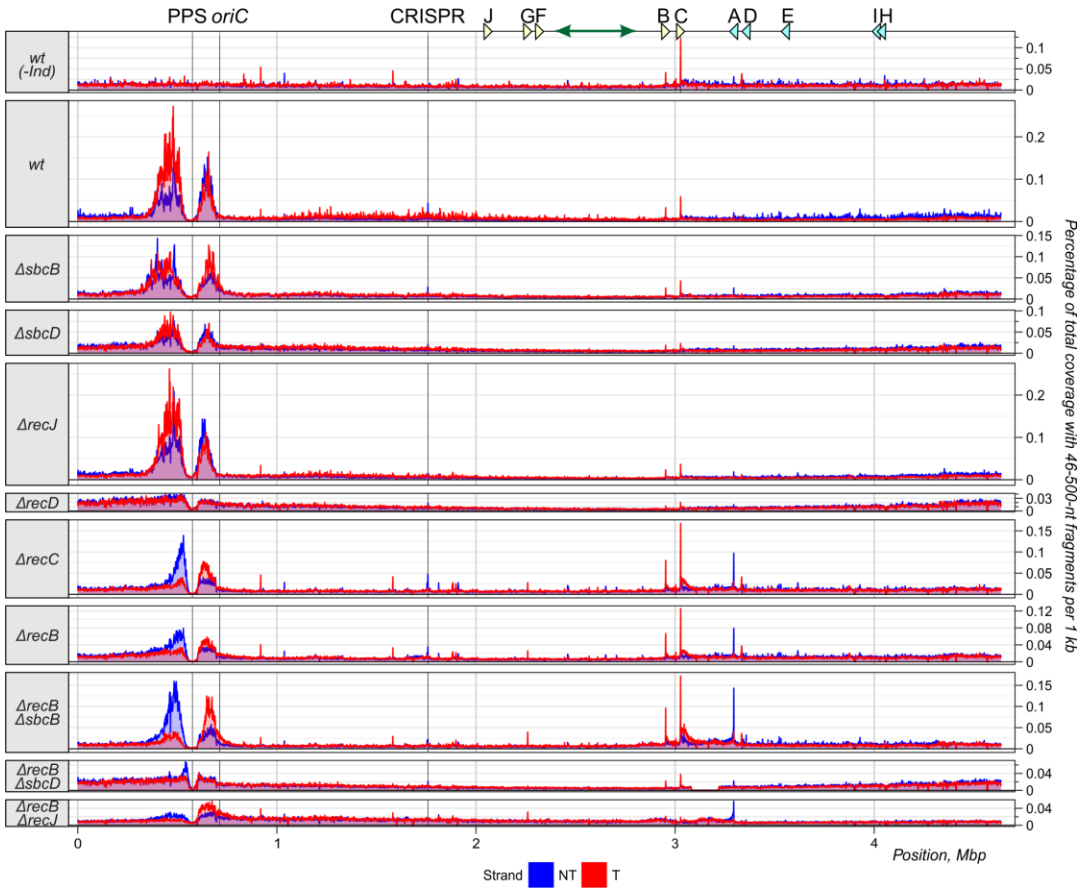


Figure 34. Fragment coverage plot for fragments 46-500 nt purified from the *wt* self-targeting strain and its DNA repair mutant derivatives. *oriC*, site of replication origin; PPS, priming protospacer; A-H, sites of replication termination *terA* – *terH* (*ter* sites stopping the replication fork progressing from left to right are shown in light blue; *ter* sites stopping the replication fork progressing from right to left are shown in yellow). The green arrow at the top shows the region used for adjustment of the fragment coverage profiles with total genomic DNA coverage profiles and background removal. The 100% value on the Y-axis corresponds to the total coverage with all fragments mapped to the genome.

The activity of RecBCD is regulated by Chi sites. We noticed that the median length of fragments produced from the NT-strand in the "upstream" peak reached local maxima within ≈3.9±1 kbp to the 5' side of the appropriately oriented Chi sites (5'-GCTGGTGG-3' sequences in the NT-strand). No such dependence was observed for the "downstream" peak where the closest local maxima of median fragment lengths were found both to the 5' and 3' side of Chi. This difference is likely caused by a higher

density of the Chi sites located in the T-strand of the "dw" region (1 Chi per ≈4 kbp) compared to the "up" region of the NT-strand (1 Chi per ≈20 kbp).

The existence of a specific subset of fragments generated from the 3' terminated strand after the recognition of a Chi is surprising and requires further studies since, based on *in vitro* studies, the degradation of the 3' terminated strand should stop at Chi (Dixon and Kowalczykowski, 1993). Interestingly, increased coverage on the NT-strand relative to the T-strand was observed to the 5' side of the Chi sites located in the "upstream" peak of the *ΔsbcB* mutant. This suggests that SbcB might be involved in the degradation of the corresponding 3' ends in the *wt* strain. In line with this, higher coverage was observed for the T-strand relative to the NT-strand in the "downstream" peak of the *ΔsbcB* mutant. We suggest that increased production of longer Chi-associated fragments from the T-strand in combination with an increased occurrence of appropriately oriented Chi sites downstream of the PPS may account for the overall higher length of fragments produced from the T-strand relative to the NT-strand in the "downstream" peak of *ΔsbcB*, which distinguishes this mutant from other RecBCD$^+$ strains used in this study.

Another pattern of fragments is produced in *ΔrecC* and *ΔrecB* cells devoid of RecBCD nuclease/helicase activities (Figure 35A,B). In this case, more extensive degradation of the 5' terminated strands was observed both up- and downstream of the PPS. Higher coverage and increased fragment lengths were observed for the 3' terminated strands. A similar strand bias was revealed in the double mutants *ΔrecB ΔsbcB, ΔrecB ΔsbcD, ΔrecB ΔrecJ* (Figure 35A,B). Further studies will be required to determine the nucleases involved in the production of fragments in the absence of RecBCD. Interestingly, the presence of the RecBC helicase (*ΔrecD* strain) inhibits the production of fragments observed in *ΔrecC* and *ΔrecB* cells (Figure 35A).

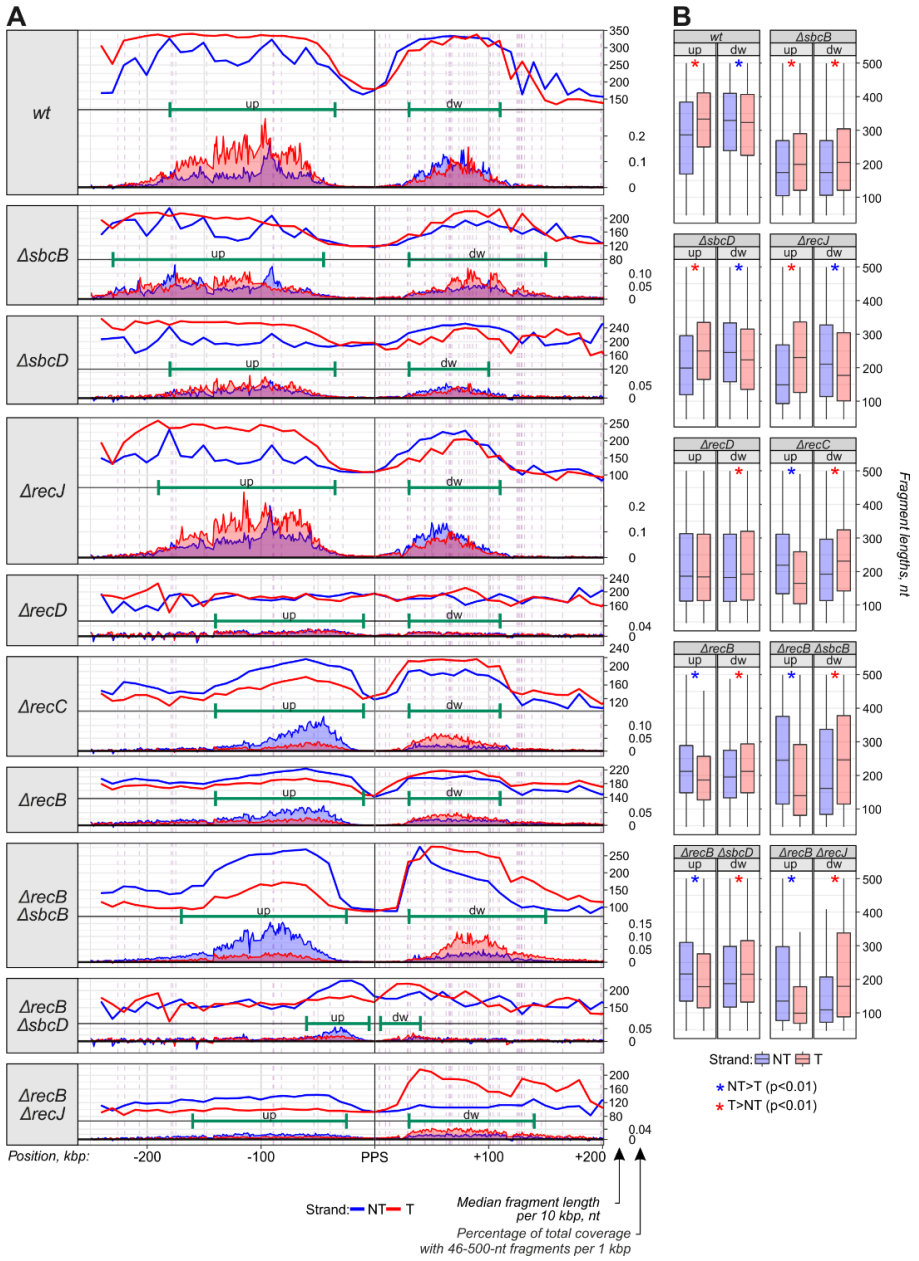


Figure 35. Analysis of 46-500-nt fragments originating from the PPS-flanking regions of *wt* self-targeting strain or its DNA repair mutant derivatives **A.** Fragment coverage plot (after background removal) for fragments 46-500 nt mapped to the PPS-region (bottom parts of the panels). Median fragment lengths in 10-kbp bins (top parts of the panels). Dashed lines indicate the positions of Chi sites appropriately oriented with respect to the position of the DSB at the PPS (Chi sites in the NT strand upstream of the PPS and Chi sites in the T strand downstream of the PPS are shown). The regions "up" and "dw" shown in green are the regions used for the comparison of fragment length distributions between the two strands shown in **B**.
B. Boxplots obtained for the lengths of fragments mapped to the "up" and "dw" regions indicated in **A.** The central line, median; hinges, the first and third quartiles; whiskers, 1.5 x IQR. The asterisks indicate p-values less than 0.01 obtained in one-sided Mann Whitney U tests computed for fragment lengths on the T- and NT-strand (after the Bonferroni correction for multiple comparisons).

DISCUSSION

Prespacer generation remains the least understood step in building CRISPR-Cas immunity. During prespacer generation by the *E. coli* type I-E system, a sequence downstream of the 5'-AAG-3' PAM should be recognized in precursor DNA and bound by the Cas1-Cas2 complex. Eventually, the 33-bp region starting with the PAM-derived G/C base pair should be excised and integrated into the CRISPR array. At the time of planning the work described in this Doctoral Thesis, it was known that the recognition of the PAM is carried out by the Cas1-Cas2 complex through the binding of one of four Cas1 subunits to the PAM-complementary 3'-TTC-5' sequence (Wang et al., 2015). It was also suggested that the Cas1-Cas2 complex generates the 3' ends of mature prespacers by cleaving one strand of the precursor DNA between the C and T nucleotides of the PAM-complementary sequence, and another strand – at a 33-bp distance from the cut site within the PAM (Wang et al., 2015). This early model did not answer several important questions.

1) What is the structure of prespacers *in vivo*?
2) Are there any intermediate forms of prespacers with a length exceeding the length of a spacer?
3) How are the 5' ends of prespacers generated?
4) Given that many nucleases are present in *E. coli* cells, can some of them participate in prespacer generation by cutting off the nucleotides not covered by Cas1-Cas2?
5) Why does the integration of prespacers into the CRISPR array yield spacers inserted in a specific orientation where the PAM-derived G/C pair becomes the last base pair of the first repeat?

The work from the Church laboratory demonstrated that the presence of the full PAM sequence in a spacer precursor (provided as a double-stranded oligonucleotide electroporated into cells expressing *cas1* and *cas2*) is crucial for the integration of the prespacer into the CRISPR array in the correct orientation (Shipman et al., 2016). This result suggests that prespacers longer than 33 bp do exist in cells but the exact length of

each strand could not be determined by the methods used in the experiments of Shipman et al.

In this Thesis, we developed an original high-throughput sequencing approach, FragSeq, to studying prespacers *in vivo*. We sequenced short DNAs purified from *E. coli* cells undergoing primed adaptation in the type I-E CRISPR-Cas system and discovered prespacers associated with the 5'-AAG-3'/3'-TTC-5' motif. We found that the AAG-associated strand of prespacers is trimmed to the length of mature spacers (≈32-34 nt) and the PAM is cleaved either between AA and G, or between A and AG. The TTC-associated strand of the detected prespacers is longer (≈36-38 nt) and has the 3'-NNTTC-5' sequence on the 3' end. Our experiments on the electroporation of oligonucleotides mimicking the detected fragments into cells expressing *cas1* and *cas2* demonstrated the high efficiency of integration of prespacers that have a blunt PAM-distal end, 33- or 34-bp double-stranded region with the 3'-NNTT-5' or 3'-NNT-5' overhang on the PAM-derived end, respectively. Most oligo-derived spacers are integrated in the "correct" orientation and have a length of 33 bp with the 3'-NNTT-5' sequence removed prior to integration. Thus, our results reveal the asymmetric structure of spacer precursors *in vivo* and demonstrate that the trimming of the PAM-complementary 3'-TTC-5' sequence between the T and C is the last processing step before the integration.

We also demonstrate that asymmetrically processed prespacers are generated by the type I-F CRISPR-Cas system of *Pseudomonas aeruginosa* suggesting that it is a characteristic of type I CRISPR-Cas systems lacking Cas4. It is tempting to speculate that the asymmetry in cleavage of the PAM-distal and PAM-proximal 3' ends contributes to the integration of prespacers into the CRISPR array in the specific orientation since the PAM-distal 3' end is processed earlier and can be engaged into the leader-side integration right away.

Shortly after our work had been published (Shiriaeva et al., 2019), two papers describing the asymmetric cleavage of the prespacer 3' ends *in vitro* came out (Kim et al., 2020; Ramachandran et al., 2020). In both works, the binding of the Cas1-Cas2 complex to a double-stranded 23-bp oligonucleotide with two long 3' overhangs, one of which

contained the 3'-TTC-5' PAM-complementary sequence, did not lead to prespacer cleavage. However, when ExoT or DnaQ exonucleases were added, the PAM-distal 3' end was trimmed to the protospacer boundary. The PAM-derived side was less processed leaving either the 3'-NNTTC-5' or 3'-NNNTTC-5' sequence on the 3' end, which is in perfect agreement with our *in vivo* results (Figure 31).

A possible explanation for the observed asymmetric cleavage can be found in crystal structures of Cas1-Cas2 (Figure 36). The Cas1 C-terminal "tail" lies on the surface of the Cas2 dimer in the Cas1-Cas2 complex not bound to a prespacer (Nuñez et al., 2014). The tail is disordered in the Cas1-Cas2 complex bound to a prespacer with long 3' overhangs solely composed of T nucleotides but covers the Cas1 catalytic pocket in the Cas1-Cas2 complex bound to a prespacer containing the 3'-TTC-5' PAM-complementary sequence within the 3' overhang (Wang et al., 2015). It was suggested that the PAM-derived 3' end is protected by the Cas1 C-terminal "tail" and is therefore trimmed by the ExoT and DnaQ exonucleases less extensively (Kim et al., 2020; Ramachandran et al., 2020). In line with this hypothesis, the loss of PAM-specificity in the selection of new spacers during naïve adaptation was observed in cells expressing a mutant *cas1* gene with a deletion of the sequence corresponding to the C-terminal "tail" (Yoganand et al., 2019).

It is not known if ExoT or DnaQ is required for processing of prespacer 3' ends *in vivo*. In a previous work from our laboratory, Musharova et al. analyzed genomic DNA from cells with active primed adaptation using primer extension reactions with primers annealed close to the protospacers that are frequently used as spacer donors (Musharova et al., 2017). Two products were formed in the first reaction with a primer annealed downstream of the PAM-distal protospacer boundary. One product corresponded to the cleavage within the 5'-AAG-3' sequence and the other one corresponded to the cleavage at the PAM-distal boundary of the protospacer. This observation is in perfect agreement with the generation of 32-34-nt AAG-associated fragments detected in the presented Thesis research. Strikingly, no products were observed by Musharova et al. in a second primer extension reaction with a primer annealed to the complementary TTC-associated

strand upstream of the protospacer PAM. The difference in the products of the two reactions can be explained if we propose that during primed adaptation the PAM-distal 3' end is generated via endonucleolytic cleavage while the PAM-derived 3' end is produced due to exonucleolytic trimming leaving no sequence for the annealing of the primer used in the second reaction by Musharova et al. Whether this hypothesis is correct and which endo- or/and exonucleases are involved in the generation of prespacer 3' ends *in vivo* is yet to be determined.

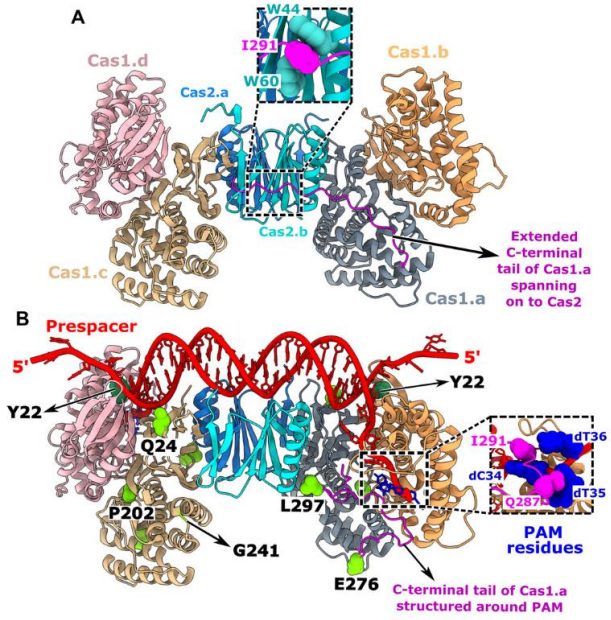


Figure 36. Conformational change in a C-terminal tail of a catalytic Cas1 subunit upon recognition of the PAM-complementary 3'-TTC-5' sequence. **A.** Structure of apo-Cas1-2. **B.** Structure of the Cas1-Cas2 complex bound to a prespacer-like oligonucleotide with the PAM-complementary sequence near one of two 3' ends. The picture is taken from Yoganand et al., 2019 (no permission is required).

The mechanism of processing of prespacer 5' ends is also enigmatic. The two 5'-end regions flanking the central 23-bp part bound to the Cas2 dimer are dislodged from the Cas1-Cas2 complex and are likely available for degradation by cellular nucleases (Figure 36B) (Nuñez et al., 2015b; Wang et al., 2015). Our results demonstrate that RecJ and RecBCD are essential for prespacer generation. When both enzymes are absent, prespacer generation is inhibited and no visible adaptation is observed. When one of the two enzymes is inactivated, prespacers are produced but their 5' ends are processed differently. In the *wt* cells, the 5' ends either coincide with the boundaries of "ideal" 33-

bp prespacers corresponding to spacers starting with the PAM-derived G/C pair (65% of 5’ ends) or are shortened by 1 nt (31% of 5’ ends). In the absence of RecJ, prespacers are trimmed less extensively and only 24% of the prespacer 5’ ends coincide with the boundaries of “ideal” prespacers, while 52% and 17% have 1 and 2 additional nucleotides, respectively. The opposite situation is observed in *ΔrecB* and *ΔrecC* strains where “ideal” 5’ ends constitute the majority (45%) but the 5’ ends shortened by 1 nt also become prominent (40%). Prespacer 5’ ends with 1 additional nucleotide constitute 10% of 5’ ends in the *ΔrecB* and *ΔrecC* mutants.

By analogy with DSBR where the RecFOR pathway (which requires RecJ) and the RecBCD pathway substitute each other, it can be proposed that two pathways leading to the generation of 5’ ends exist. However, if the two pathways were mutually independent, and one pathway operated on some prespacers while the other pathway produced the rest of prespacer 5’ ends, then the combined effect would be different from what is observed in *wt*. The proportion of ideally processed 5’ ends in *wt* is higher than in any of the mutants. Almost no prespacer 5’ ends extended by 2 nt or shortened by 1 nt are observed in *wt* while these events are frequent in *ΔrecJ* and *ΔrecB* (or *ΔrecC*), respectively. Altogether, our results better fit a hypothesis, according to which RecBCD and RecJ cooperate to ensure the generation of prespacers whose 5’ ends coincide with the ends of 33-bp spacers.

The mechanism of this cooperation is yet to be discovered but, based on the observation that the RecD subunit is dispensable for prespacer generation, we propose that the RecBC helicase rather than RecBCD is required for the generation of 5’ ends. Since RecBCD preferentially binds to double-stranded ends, we propose that spacer precursors formed during primed adaptation are bound by Cas1-Cas2 and flanked by long double-stranded regions. RecBC unwinds the ends of these precursors providing access of RecJ to the 5’-terminated single-stranded ends (Figure 37). When RecBC or RecJ is absent, other proteins likely substitute for the lacking helicase or 5’→3’ exonuclease activities and future experiments will be required to identify them.

Most experiments described in this work were performed in strains undergoing primed adaptation under conditions when the bacterial chromosome was targeted by the CRISPR interference machinery. During primed adaptation, new spacers are selected from the protospacers that have an adjacent 3'-TTC-5' PAM-complementary sequence in the target strand upstream of the PPS or in the non-target strand downstream of the PPS. Two models of primed adaptation were proposed. The first model assumes that the Cas3 nuclease/helicase and Cas1-Cas2 complex act independently: Cas3 cleaves DNA into fragments that are picked up by Cas1-Cas2 and used as spacer precursors (Swarts et al., 2012). This model could explain the observed strand bias in PAM selection if Cas3 recognized the 3'-TTC-5' sequence in one strand and produced some specific fragments selectively bound by Cas1-Cas2 as spacer precursors. The second model suggests that Cas3 and Cas1-Cas2 form a larger complex that slides along DNA searching for protospacers (Datsenko et al., 2012; Dillard et al., 2018; Redding et al., 2015). If such a complex existed, the observed strand bias could be explained by a specific architecture of the complex, for example, if only one of two catalytic Cas1 subunits was available for binding the 3'-TTC-5' motif and only in one DNA strand. It was recently shown that Cas1-Cas2 binds to ssDNA and facilitates the pairing of complementary strands (Kim et al., 2020). It is thus possible that after the binding to a TTC-containing ssDNA region, Cas1-Cas2 facilitates its pairing with the complementary region forming a precursor further trimmed by nucleases present in the cell.

Though our results do not provide direct evidence in favor of the second model, some observations suggest that the first model is not correct.

1. Künne et al. demonstrated that *in vitro* Cas3 cleaves both strands of a target plasmid into fragments that have a T nucleotide on the 3' end (Künne et al., 2016). Though it was not explained how the strand bias can be generated, the cleavage after T was proposed to be a mechanism enriching the 3'-TTC-5' motif on fragments 3' ends in support of the first model. Our results demonstrate that prespacers are ending with the 3'-NNTTC-5' sequence rather than the 3'-TTC-5' suggesting that either Cas3 specificity is different *in vivo* or the PAM-derived 3' ends are not produced by Cas3.

2. By analyzing the coverage of total genomic DNA, we revealed that the size of the gap near the PPS formed due to degradation of DNA is decreased in cells devoid of RecBCD. This result suggests that RecBCD continues degradation after Cas3 dissociates from DNA. Using FragSeq, we revealed enrichment of 46-500-nt fragments in the regions ≈100 kbp – 30 kbp upstream of the PPS and ≈20 kbp -100 kbp downstream of the PPS but not in the 50-kbp region encompassing the PPS. If Cas3 produced fragments of similar lengths, we would expect to see them in approximately the same quantities (given that RecBCD accounts for 21-49% of degradation near the PPS). Therefore, it is likely that the products of Cas3 are so short that cannot be detected by our methods and cannot be used as spacer precursors.

There are several observations in the literature indicating that the second model might be true. First, single-molecule experiments with Cas proteins from the type I-E system of *Thermobifida fusca* revealed the assembly of the primed acquisition complex (PAC) composed of Cascade, Cas3, and Cas1-Cas2, which moved along the target DNA (Dillard et al., 2018). Second, Cas3 is fused to Cas2 and forms a complex with Cas1 in the closely related type I-F CRISPR-Cas system (Fagerlund et al., 2017; Richter et al., 2012a; Rollins et al., 2017). Interestingly, during primed adaptation by the type I-E and I-F systems, the respective PAMs are recognized in the opposite strands leading to the generation of inversed gradients of spacers mapping to the PPS-region (Figure 26C). Differences in the architecture of the type I-E or I-F primed adaptation complexes could probably cause the recognition of the PAM sequences in the opposite strands.

Based on the published data and our results, we present the following speculative model of primed adaptation in the type I-E CRISPR-Cas system (Figure 37). The PAC composed of Cascade, Cas3 and Cas1-Cas2 is assembled on the PPS (Dillard et al., 2018). The PAC translocates in the direction upstream of the PPS due to the Cas3 helicase activity and has limited nuclease activity (Dillard et al., 2018). One of Cas1 subunits contacts the target strand and recognizes the 3’-TTC-5’ sequence. The recognition of the PAM-complementary sequence leads to the binding of Cas1-Cas2 to the adjacent protospacer in ssDNA. This makes the PAC stop and, probably, disintegrate.

Cas1-Cas2 bound to the single-stranded TTC-associated protospacer facilitates pairing with the complementary strand. Due to cleavages by Cas3 or some other nucleases, a long double-stranded precursor bound to Cas1-Cas2 is produced. RecBC complexes bind to the double-stranded ends and unwind the two strands. The RecJ 5'→3' exonuclease degrades the unwound 5'-terminated strands up to the protospacer sequence. The PAM-distal 3' end is generated due to an endonucleolytic cut at the protospacer boundary by an unknown endonuclease. The PAM-derived 3' end is trimmed by an unidentified 3'→5' exonuclease up to the 3'-NNTTC-5' sequence protected by the C-terminal tail of the corresponding Cas1 subunit. The asymmetrically processed prespacer bound to Cas1-Cas2 is recruited to the CRISPR array. The IHF protein bound to the leader stimulates the integration of the fully processed PAM-distal 3' end at the leader/repeat boundary (Nuñez et al., 2016). The formation of the half-site intermediate is followed by the removal of the 3'-NNTT-5' overhang from the PAM-derived 3' end and its integration at the repeat/spacer boundary (Kim et al., 2020; Ramachandran et al., 2020).

In addition to the characterization of spacer precursors and the identification of proteins involved in their generation *in vivo*, our work provides new insights into the first stages of double-strand break repair and its impact on CRISPR interference. We confirm previous unpublished results from our laboratory obtained by Elena Kurilovich that the RecBCD complex enlarges the gap in genomic coverage generated as result of DNA degradation near the PPS by Cas3. Our results suggest that RecBCD is responsible for ≈21-49% of the gap produced in the *wt* strain. High-throughput sequencing of short DNAs purified from cells containing RecBCD (*wt*, *ΔsbcB*, *ΔsbcD* or *ΔrecJ*) or cells lacking RecBCD nuclease activity (*ΔrecB*, *ΔrecC* or *ΔrecD*) was for the first time conducted in this work. It revealed 46-500-nt fragments produced by RecBCD during degradation of DNA regions flanking the primary area of degradation by CRISPR interference machinery. There are published data suggesting that *in vitro* RecBCD cleaves the 3'-terminated strand more extensively than the 5'-terminated strand, at least under certain experimental conditions (excess of Mg^{2+} concentration over ATP) (Dixon and Kowalczykowski, 1993; Taylor and Smith, 1995b). In general, our results *in vivo*

support the asymmetric degradation of the 5'- and 3'-terminated strands (Figure 35). The 3'-terminated strand is on average cleaved into shorter fragments than the 5'-terminated strand. In addition, the total coverage with the RecBCD-produced fragments is higher for the 5'-terminated strand than for the 3'-terminated strand suggesting that a part of 3'-terminated strands are cleaved into much shorter fragments not detected by HTS.

The activity of RecBCD is regulated by Chi sites recognized in the 3'-terminated strand. According to numerous genetic and biochemical results, the degradation of the 3'-terminated strand is inhibited after Chi recognition while the cleavage of the 5'-terminated strand is continued, RecBCD starts loading RecA protein on the 3' overhang that stimulates recombination to the 5' side of the Chi (Anderson and Kowalczykowski, 1997a; Dixon and Kowalczykowski, 1993; Faulds et al., 1979; Stahl et al., 1980). Surprisingly, we found that the lengths of fragments generated from the 3-5-kbp regions to the 5' side of Chi on the 3'-terminated strand are higher than for the fragments generated from the same strand before Chi recognition (Figure 35A). The source of these fragments produced from the 3'-terminated strand after Chi recognition remains unknown. Under conditions of self-targeting no homologous template for repair is available in most cells due to efficient cleavage of the PPS by Cas3. We hypothesize that the inability to complete repair might lead to disassembly of the RecA filament and cleavage of the 3' overhangs by some nucleases yielding the fragments originating from the 3'-terminated strand to the 5' side of Chi that we detected.

Surprisingly, in cells with a deletion of *recB* or *recC*, another type of fragments originating from the regions flanking the primary area of degradation by CRISPR interference machinery was observed. Two features describe these fragments. First, fragments mapped to the 3'-terminated strands were on average longer than fragments produced on the complementary region in the same regions. Second, higher coverage with fragments for the 3'-terminated strands was observed. Both characteristics were retained in the tested double mutants *ΔrecB ΔsbcD*, *ΔrecB ΔsbcB*, *ΔrecB ΔrecJ*. Therefore, the nuclease responsible for the generation of these fragments remains unidentified (we will refer to this nuclease as nuclease X). At the same time, the size of

the regions from which these fragments are produced is decreased in a double mutant *ΔrecB ΔsbcD*. Interestingly, the analysis of the gap in genomic DNA coverage shows that the extent of degradation in the *ΔrecB ΔsbcD* mutant constitutes only about 21-29% of what was observed in *wt* and about 30-37% of what was observed in *ΔrecB*. This result suggests that in the absence of RecBCD, SbcCD facilitates degradation probably by the nuclease X. In the future, it would be interesting to test the influence of other mutations introduced into the *ΔrecB* strain. It would be also interesting to repeat this analysis for cells with a single DSB not related to the type I CRISPR interference (for example, introduced by Cas9) to test if Cas3 could be the nuclease X.

Overall, in our work, we approached CRISPR adaptation, interference, and DNA degradation by cellular genome maintenance systems from a new perspective – through high-throughput sequencing of short DNA fragments produced as intermediates or end products of these processes *in vivo*. All our experiments were performed in *E. coli* and addressed the generation of prespacers by the type I-E and the type I-F CRISPR-Cas systems as well as the first stages of double-strand break repair. We believe that a similar approach can be applied to explore prespacer generation by different CRISPR-Cas types and various aspects of DNA replication and repair in diverse species.

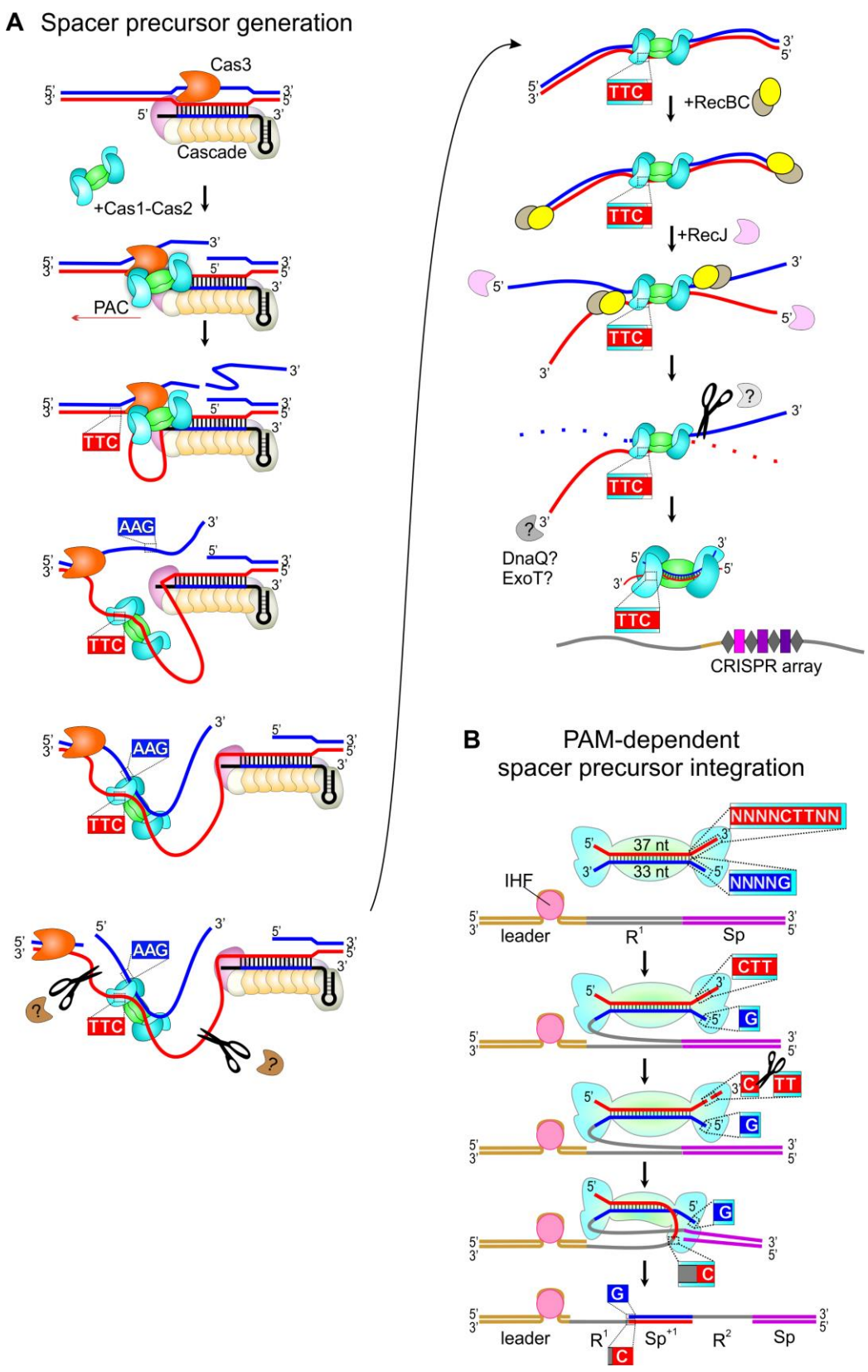

A Spacer precursor generation
Cas3
Cascade
+Cas1-Cas2
PAC
TTC
AAG
+RecBC
+RecJ
DnaQ?
ExoT?
CRISPR array
B
PAM-dependent
spacer precursor integration
NNNNCTTNN
37 nt
33 nt
NNNNG
IHF
leader
R1
Sp
CTT
G
TT
C
Sp+1
R2

Figure 37. Model of primed adaptation in the type I-E CRISPR-Cas system. **A.** Generation of spacer precursors. Primed acquisition complex (PAC) composed of Cascade, Cas3, and Cas1-Cas2 is assembled on the PPS (Dillard et al., 2018). The PAC translocates in the direction upstream of the PPS due to the Cas3 3'→5' helicase activity and has limited nuclease activity (Dillard et al., 2018). One of Cas1 subunits contacts the unwound target strand and recognizes the 3'-TTC-5' sequence. The recognition of the PAM-complementary sequence leads to the binding of Cas1-Cas2 to the adjacent protospacer in ssDNA and disintegration of the PAC. Cas1-Cas2 bound to the single-stranded TTC-associated protospacer facilitates pairing with the complementary strand (Kim et al., 2020). Due to cleavages by Cas3 or some other nucleases, a long double-stranded precursor bound to Cas1-Cas2 is produced. RecBC complexes bind to the double-strand ends and unwind the two strands. The RecJ 5'→3' exonuclease degrades the unwound 5'-terminated strands up to the protospacer sequence. The PAM-distal 3' end is generated due to an endonucleolytic cut at the protospacer boundary by an unknown endonuclease. The PAM-derived 3' end is trimmed by an unidentified 3'→5' exonuclease up to the 3'-NNTTC-5' sequence protected by the C-terminal tail of the corresponding Cas1 subunit. **B.** Prespacer integration into the CRISPR array. The asymmetrically processed prespacer bound to Cas1-Cas2 is recruited to the CRISPR array. The IHF protein bound to the leader stimulates the integration of the fully processed PAM-distal 3' end at the leader/repeat boundary (Nuñez et al., 2016). The formation of the half-site intermediate is followed by the removal of the 3'-TTNN-5' overhang from the PAM-derived 3' end and its integration at the repeat/spacer boundary (Kim et al., 2020; Ramachandran et al., 2020).

CONCLUSIONS

1. Targeting of a chromosomally located protospacer (self-targeting) by the type I-E CRISPR-Cas system of *E. coli* leads to the degradation of several hundred kbp of genomic DNA flanking the target.
2. Degradation of DNA surrounding the targeted protospacer takes place in two stages. A region comprising ≈30 kbp upstream and 20 kbp downstream of the target is degraded such that no fragments could be detected by high-throughput sequencing Degradation outside of this region of DNA is carried out by RecBCD, which degrades strands with free 3' termini more extensively than the complementary strands with free 5' termini. Longer fragments are produced from 3-5-kbp regions adjacent to the 5' side of the 5'-GCTGGTGG-3' Chi motifs located in the 3'-terminated strands.
3. In the absence of RecBCD, degradation of DNA around the targeted protospacer also proceeds in two stages. The first stage is similar to that observed in $RecBCD^+$ cells (degradation of ≈50 kbp of DNA around the target without production of fragments that can be revealed by our methods). During the second stage, strands with free 5' termini are degraded more extensively than the complementary strands with free 3' termini. The nucleases performing this cleavage have not been identified yet but they must be less processive than RecBCD since the width of the gap in total genomic DNA coverage is decreased in *ΔrecB* and *ΔrecC* mutants compared with the *wt* strain. In addition, the activity of these unidentified nucleases likely depends on SbcCD since the gap in total genomic DNA coverage is further narrowed in a double mutant *ΔrecB ΔsbcD* and so are the regions enriched with the fragments produced at the second stage of interference. The presence of the RecBC helicase prevents the generation of these fragments.
4. Degradation of genomic DNA during self-targeting inhibits cell division but cells remain alive at least for 5 hours, during which spacers derived from the regions flanking the targeted protospacer get incorporated into the CRISPR array through primed adaptation.

5. Prespacers generated during primed adaptation by the type I-E CRISPR-Cas system are double-stranded 33-34-bp fragments with a blunt end on a PAM-distal side and 4 or 3 additional nucleotides on the 3’ end of the PAM-derived side (5’-TTNN-‘3 or 5’-TNN-3’ overhang). Spacer precursors with a 3′-end overhang on the PAM-derived side are also produced during primed adaptation by the type I-F CRISPR-Cas system of *Pseudomonas aeruginosa* suggesting that the asymmetrical structure of prespacers is a common characteristic of type I CRISPR-Cas systems lacking Cas4.
6. The RecBCD helicase and RecJ 5’→3’ exonuclease activities are involved in the generation of prespacer 5’ ends in the type I-E CRISPR-Cas system. RecBCD and RecJ are redundant but prespacers are trimmed, respectively, more or less extensively when the former or the latter enzyme is not present in a cell. Inactivation of both enzymes abolishes prespacer generation.

BIBLIOGRAPHY

1. 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.

2. Amundsen, S.K., and Smith, G.R. (2003). Interchangeable parts of the Escherichia coli recombination machinery. Cell *112*, 741–744.

3. Amundsen, S.K., Taylor, A.F., Chaudhury, A.M., and Smith, G.R. (1986). recD: the gene for an essential third subunit of exonuclease V. Proc. Natl. Acad. Sci. U.S.A. *83*, 5558–5562.

4. Anderson, D.G., and Kowalczykowski, S.C. (1997a). The Translocating RecBCD Enzyme Stimulates Recombination by Directing RecA Protein onto ssDNA in a χ-Regulated Manner. Cell *90*, 77–86.

5. Anderson, D.G., and Kowalczykowski, S.C. (1997b). The recombination hot spot chi is a regulatory element that switches the polarity of DNA degradation by the RecBCD enzyme. Genes Dev. *11*, 571–581.

6. Arslan, Z., Hermanns, V., Wurm, R., Wagner, R., and Pul, Ü. (2014). Detection and characterization of spacer integration intermediates in type I-E CRISPR-Cas system. Nucleic Acids Res. *42*, 7884–7893.

7. Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K.A., Tomita, M., Wanner, B.L., and Mori, H. (2006). Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol. Syst. Biol. *2*, 2006.0008.

8. Babu, M., Beloglazova, N., Flick, R., Graham, C., Skarina, T., Nocek, B., Gagarinova, A., Pogoutse, O., Brown, G., Binkowski, A., et al. (2011). A dual function of the CRISPR-Cas system in bacterial antivirus immunity and DNA repair. Mol Microbiol *79*, 484–502.

9. Barbour, S.D., Nagaishi, H., Templin, A., and Clark, A.J. (1970). Biochemical and genetic studies of recombination proficiency in Escherichia coli. II. Rec+ revertants caused by indirect suppression of rec- mutations. Proc. Natl. Acad. Sci. U.S.A. *67*, 128–135.

10. Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A., and Horvath, P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. Science *315*, 1709–1712.

11. Bell, J.C., Plank, J.L., Dombrowski, C.C., and Kowalczykowski, S.C. (2012). Direct imaging of RecA nucleation and growth on single molecules of SSB-coated ssDNA. Nature *491*, 274–278.

12. Beloglazova, N., Brown, G., Zimmerman, M.D., Proudfoot, M., Makarova, K.S., Kudritska, M., Kochinyan, S., Wang, S., Chruszcz, M., Minor, W., et al. (2008). A novel family of sequence-specific endoribonucleases associated with the clustered regularly interspaced short palindromic repeats. J. Biol. Chem. *283*, 20361–20371.

13. Benjamini, Y., Heller, R., and Yekutieli, D. (2009). Selective inference in complex research. Philos Trans A Math Phys Eng Sci *367*, 4255–4271.

14. Bernheim, A., Bikard, D., Touchon, M., and Rocha, E.P.C. (2019). 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. *374*, 20180088.

15. Bianco, P.R., and Kowalczykowski, S.C. (1997). The recombination hotspot Chi is recognized by the translocating RecBCD enzyme as the single strand of DNA containing the sequence 5′-GCTGGTGG-3′. PNAS *94*, 6706–6711.

16. Bianco, P.R., Brewer, L.R., Corzett, M., Balhorn, R., Yeh, Y., Kowalczykowski, S.C., and Baskin, R.J. (2001). Processive translocation and DNA unwinding by individual RecBCD enzyme molecules. Nature *409*, 374–378.

17. Biek, D.P., and Cohen, S.N. (1986). Identification and characterization of recD, a gene affecting plasmid maintenance and recombination in Escherichia coli. J. Bacteriol. *167*, 594–603.

18. Blattner, F.R., Plunkett, G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., et al. (1997). The complete genome sequence of Escherichia coli K-12. Science *277*, 1453–1462.

19. Bobay, L.-M., Touchon, M., and Rocha, E.P.C. (2013). Manipulating or superseding host recombination functions: a dilemma that shapes phage evolvability. PLoS Genet. *9*, e1003825.

20. Boehmer, P.E., and Emmerson, P.T. (1992). The RecB subunit of the Escherichia coli RecBCD enzyme couples ATP hydrolysis to DNA unwinding. J. Biol. Chem. *267*, 4981–4987.

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

22. Boulos, L., Prévost, M., Barbeau, B., Coallier, J., and Desjardins, R. (1999). LIVE/DEAD®BacLight™: application of a new rapid staining method for direct enumeration of viable and total bacteria in drinking water. Journal of Microbiological Methods *37*, 77–86.

23. Braedt, G., and Smith, G.R. (1989). Strand specificity of DNA unwinding by RecBCD enzyme. Proc. Natl. Acad. Sci. U.S.A. *86*, 871–875.

24. Brouns, S.J.J., Jore, M.M., Lundgren, M., Westra, E.R., Slijkhuis, R.J.H., Snijders, A.P.L., Dickman, M.J., Makarova, K.S., Koonin, E.V., and van der Oost, J. (2008). Small CRISPR RNAs guide antiviral defense in prokaryotes. Science *321*, 960–964.

25. Cady, K.C., Bondy-Denomy, J., Heussler, G.E., Davidson, A.R., and O'Toole, G.A. (2012). The CRISPR/Cas adaptive immune system of Pseudomonas aeruginosa mediates resistance to naturally occurring and engineered phages. J. Bacteriol. *194*, 5728–5738.

26. 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.

27. Cassuto, E., West, S.C., Mursalim, J., Conlon, S., and Howard-Flanders, P. (1980). Initiation of genetic recombination: homologous pairing between duplex DNA molecules promoted by recA protein. Proc. Natl. Acad. Sci. U.S.A. *77*, 3962–3966.

28. Chase, J.W., and Richardson, C.C. (1974a). Exonuclease VII of Escherichia coli. Mechanism of action. J. Biol. Chem. *249*, 4553–4561.

29. Chase, J.W., and Richardson, C.C. (1974b). Exonuclease VII of Escherichia coli. Purification and properties. J. Biol. Chem. *249*, 4545–4552.

30. Chaudhury, A.M., and Smith, G.R. (1984). A new class of Escherichia coli recBC mutants: implications for the role of RecBC enzyme in homologous recombination. Proc. Natl. Acad. Sci. U.S.A. *81*, 7850–7854.

31. Cheng, K., Wilkinson, M., Chaban, Y., and Wigley, D.B. (2020). A conformational switch in response to Chi converts RecBCD from phage destruction to DNA repair. Nat. Struct. Mol. Biol. *27*, 71–77.

32. Churchill, J.J., Anderson, D.G., and Kowalczykowski, S.C. (1999). The RecBC enzyme loads RecA protein onto ssDNA asymmetrically and independently of chi, resulting in constitutive recombination activation. Genes Dev. *13*, 901–911.

33. Clark, A.J., and Margulies, A.D. (1965). ISOLATION AND CHARACTERIZATION OF RECOMBINATION-DEFICIENT MUTANTS OF ESCHERICHIA COLI K12. Proc. Natl. Acad. Sci. U.S.A. *53*, 451–459.

34. Connelly, J.C., and Leach, D.R. (1996). The sbcC and sbcD genes of Escherichia coli encode a nuclease involved in palindrome inviability and genetic recombination. Genes Cells *1*, 285–291.

35. Connelly, J.C., de Leau, E.S., Okely, E.A., and Leach, D.R. (1997). Overexpression, purification, and characterization of the SbcCD protein from Escherichia coli. J. Biol. Chem. *272*, 19819–19826.

36. Connelly, J.C., Kirkham, L.A., and Leach, D.R. (1998). The SbcCD nuclease of Escherichia coli is a structural maintenance of chromosomes (SMC) family protein that cleaves hairpin DNA. Proc. Natl. Acad. Sci. U.S.A. *95*, 7969–7974.

37. Connelly, J.C., de Leau, E.S., and Leach, D.R. (1999). DNA cleavage and degradation by the SbcCD protein complex from Escherichia coli. Nucleic Acids Res. *27*, 1039–1046.

38. Connelly, J.C., de Leau, E.S., and Leach, D.R.F. (2003). Nucleolytic processing of a protein-bound DNA end by the E. coli SbcCD (MR) complex. DNA Repair (Amst.) *2*, 795–807.

39. Cox, M.M., and Lehman, I.R. (1981). Directionality and polarity in recA protein-promoted branch migration. Proc. Natl. Acad. Sci. U.S.A. *78*, 6018–6022.

40. Craig, N.L., and Nash, H.A. (1984). E. coli integration host factor binds to specific sites in DNA. Cell *39*, 707–716.

41. Dabert, P., Ehrlich, S.D., and Gruss, A. (1992). Chi sequence protects against RecBCD degradation of DNA in vivo. Proc. Natl. Acad. Sci. U.S.A. *89*, 12073–12077.

42. Datsenko, K.A., and Wanner, B.L. (2000). One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. Proc. Natl. Acad. Sci. U.S.A. *97*, 6640–6645.

43. 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.

44. Deltcheva, E., Chylinski, K., Sharma, C.M., Gonzales, K., Chao, Y., Pirzada, Z.A., Eckert, M.R., Vogel, J., and Charpentier, E. (2011). CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. Nature *471*, 602–607.

45. Dermić, D. (2006). Functions of multiple exonucleases are essential for cell viability, DNA repair and homologous recombination in recD mutants of Escherichia coli. Genetics *172*, 2057–2069.

46. Dermić, D., Zahradka, D., and Petranović, M. (2006). Exonuclease requirements for recombination of lambda-phage in recD mutants of Escherichia coli. Genetics *173*, 2399–2402.

47. Deveau, H., Barrangou, R., Garneau, J.E., Labonté, J., Fremaux, C., Boyaval, P., Romero, D.A., Horvath, P., and Moineau, S. (2008). Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. J. Bacteriol. *190*, 1390–1400.

48. Díez-Villaseñor, C., Almendros, C., García-Martínez, J., and Mojica, F.J.M. (2010). Diversity of CRISPR loci in Escherichia coli. Microbiology (Reading, Engl.) *156*, 1351–1361.

49. 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.

50. Dillard, K.E., Brown, M.W., Johnson, N.V., Xiao, Y., Dolan, A., Hernandez, E., Dahlhauser, S.D., Kim, Y., Myler, L.R., Anslyn, E.V., et al. (2018). Assembly and Translocation of a CRISPR-Cas Primed Acquisition Complex. Cell *175*, 934-946.e15.

51. Dillingham, M.S., and Kowalczykowski, S.C. (2008). RecBCD enzyme and the repair of double-stranded DNA breaks. Microbiol. Mol. Biol. Rev. *72*, 642–671, Table of Contents.

52. Dillingham, M.S., Spies, M., and Kowalczykowski, S.C. (2003). RecBCD enzyme is a bipolar DNA helicase. Nature *423*, 893–897.

53. Dillingham, M.S., Webb, M.R., and Kowalczykowski, S.C. (2005). Bipolar DNA translocation contributes to highly processive DNA unwinding by RecBCD enzyme. J. Biol. Chem. *280*, 37069–37077.

54. Dixit, B., Ghosh, K.K., Fernandes, G., Kumar, P., Gogoi, P., and Kumar, M. (2016). Dual nuclease activity of a Cas2 protein in CRISPR-Cas subtype I-B of Leptospira interrogans. FEBS Lett. *590*, 1002–1016.

55. Dixon, D.A., and Kowalczykowski, S.C. (1993). The recombination hotspot χ is a regulatory sequence that acts by attenuating the nuclease activity of the E. coli RecBCD enzyme. Cell *73*, 87–96.

56. 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*.

57. van Duijn, E., Barbu, I.M., Barendregt, A., Jore, M.M., Wiedenheft, B., Lundgren, M., Westra, E.R., Brouns, S.J.J., Doudna, J.A., van der Oost, J., et al. (2012).

Native tandem and ion mobility mass spectrometry highlight structural and modular similarities in clustered-regularly-interspaced shot-palindromic-repeats (CRISPR)-associated protein complexes from Escherichia coli and Pseudomonas aeruginosa. Mol. Cell Proteomics *11*, 1430–1441.

58. Eggleston, A.K., and Kowalczykowski, S.C. (1993). Biochemical characterization of a mutant recBCD enzyme, the recB2109CD enzyme, which lacks chi-specific, but not non-specific, nuclease activity. J. Mol. Biol. *231*, 605–620.

59. El Karoui, M., Biaudet, V., Schbath, S., and Gruss, A. (1999). Characteristics of Chi distribution on different bacterial genomes. Res. Microbiol. *150*, 579–587.

60. Enquist, L.W., and Skalka, A. (1973). Replication of bacteriophage lambda DNA dependent on the function of host and viral genes. I. Interaction of red, gam and rec. J. Mol. Biol. *75*, 185–212.

61. 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.

62. van Erp, P.B.G., Jackson, R.N., Carter, J., Golden, S.M., Bailey, S., and Wiedenheft, B. (2015). Mechanism of CRISPR-RNA guided recognition of DNA targets in Escherichia coli. Nucleic Acids Res. *43*, 8381–8391.

63. van Erp, P.B.G., Patterson, A., Kant, R., Berry, L., Golden, S.M., Forsman, B.L., Carter, J., Jackson, R.N., Bothner, B., and Wiedenheft, B. (2018). Conformational Dynamics of DNA Binding and Cas3 Recruitment by the CRISPR RNA-Guided Cascade Complex. ACS Chem. Biol. *13*, 481–490.

64. Estrella, M.A., Kuo, F.-T., and Bailey, S. (2016). RNA-activated DNA cleavage by the Type III-B CRISPR-Cas effector complex. Genes Dev. *30*, 460–470.

65. Fagerlund, R.D., Wilkinson, M.E., Klykov, O., Barendregt, A., Pearce, F.G., Kieper, S.N., Maxwell, H.W.R., Capolupo, A., Heck, A.J.R., Krause, K.L., 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.

66. Faulds, D., Dower, N., Stahl, M.M., and Stahl, F.W. (1979). Orientation-dependent recombination hotspot activity in bacteriophage lambda. J. Mol. Biol. *131*, 681–695.

67. Fineran, P.C., Gerritzen, M.J.H., Suárez-Diez, M., Künne, T., Boekhorst, J., van Hijum, S.A.F.T., Staals, R.H.J., and Brouns, S.J.J. (2014). Degenerate target sites mediate rapid primed CRISPR adaptation. Proc. Natl. Acad. Sci. U.S.A. *111*, E1629-1638.

68. Forget, A.L., and Kowalczykowski, S.C. (2012). Single-molecule imaging of DNA pairing by RecA reveals a three-dimensional homology search. Nature *482*, 423–427.

69. Fu, B.X.H., Wainberg, M., Kundaje, A., and Fire, A.Z. (2017). High-Throughput Characterization of Cascade type I-E CRISPR Guide Efficacy Reveals Unexpected PAM Diversity and Target Sequence Preferences. Genetics *206*, 1727–1738.

70. Ganesan, A.K., and Seawell, P.C. (1975). The effect of lexA and recF mutations on post-replication repair and DNA synthesis in Escherichia coli K-12. Mol. Gen. Genet. *141*, 189–205.

71. Ganesan, S., and Smith, G.R. (1993). Strand-specific binding to duplex DNA ends by the subunits of the Escherichia coli RecBCD enzyme. J. Mol. Biol. *229*, 67–78.

72. Gao, P., Yang, H., Rajashankar, K.R., Huang, Z., and Patel, D.J. (2016). Type V CRISPR-Cas Cpf1 endonuclease employs a unique mechanism for crRNA-mediated target DNA recognition. Cell Res. *26*, 901–913.

73. Garrett, S., Shiimori, M., Watts, E.A., Clark, L., Graveley, B.R., and Terns, M.P. (2020). Primed CRISPR DNA uptake in Pyrococcus furiosus. Nucleic Acids Res. *48*, 6120–6135.

74. Garside, E.L., Schellenberg, M.J., Gesner, E.M., Bonanno, J.B., Sauder, J.M., Burley, S.K., Almo, S.C., Mehta, G., and MacMillan, A.M. (2012). Cas5d processes pre-crRNA and is a member of a larger family of CRISPR RNA endonucleases. RNA *18*, 2020–2028.

75. 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*, E2579-2586.

76. Gibson, F.P., Leach, D.R., and Lloyd, R.G. (1992). Identification of sbcD mutations as cosuppressors of recBC that allow propagation of DNA palindromes in Escherichia coli K-12. J. Bacteriol. *174*, 1222–1228.

77. Gillen, J.R., Willis, D.K., and Clark, A.J. (1981). Genetic analysis of the RecE pathway of genetic recombination in Escherichia coli K-12. J. Bacteriol. *145*, 521–532.

78. Goldberg, G.W., Jiang, W., Bikard, D., and Marraffini, L.A. (2014). Conditional tolerance of temperate phages via transcription-dependent CRISPR-Cas targeting. Nature *514*, 633–637.

79. Goldmark, P.J., and Linn, S. (1970). An endonuclease activity from Escherichia coli absent from certain rec- strains. Proc. Natl. Acad. Sci. U.S.A. *67*, 434–441.

80. Goldmark, P.J., and Linn, S. (1972). Purification and properties of the recBC DNase of Escherichia coli K-12. J. Biol. Chem. *247*, 1849–1860.

81. Gong, B., Shin, M., Sun, J., Jung, C.-H., Bolt, E.L., van der Oost, J., and Kim, J.-S. (2014). Molecular insights into DNA interference by CRISPR-associated nuclease-helicase Cas3. Proc Natl Acad Sci U S A *111*, 16359–16364.

82. Gorbalenya, A.E., Koonin, E.V., Donchenko, A.P., and Blinov, V.M. (1988). A novel superfamily of nucleoside triphosphate-binding motif containing proteins which are probably involved in duplex unwinding in DNA and RNA replication and recombination. FEBS Lett. *235*, 16–24.

83. Gorbalenya, A.E., Koonin, E.V., Donchenko, A.P., and Blinov, V.M. (1989). Two related superfamilies of putative helicases involved in replication, recombination, repair and expression of DNA and RNA genomes. Nucleic Acids Res. *17*, 4713–4730.

84. Gutierrez, B., Ng, J.W., Cui, L., Becavin, C., and Bikard, D. (2018). Genome-wide CRISPR-Cas9 screen in E. coli identifies design rules for efficient targeting. BioRxiv 308148.

85. Haft, D.H., Selengut, J., Mongodin, E.F., and Nelson, K.E. (2005). A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. PLoS Comput. Biol. *1*, e60.

86. Hale, C., Kleppe, K., Terns, R.M., and Terns, M.P. (2008). Prokaryotic silencing (psi)RNAs in Pyrococcus furiosus. RNA *14*, 2572–2579.

87. Hall, S.D., and Kolodner, R.D. (1994). Homologous pairing and strand exchange promoted by the Escherichia coli RecT protein. Proc. Natl. Acad. Sci. U.S.A. *91*, 3205–3209.

88. Hall, S.D., Kane, M.F., and Kolodner, R.D. (1993). Identification and characterization of the Escherichia coli RecT protein, a protein encoded by the recE region that promotes renaturation of homologous single-stranded DNA. J. Bacteriol. *175*, 277–287.

89. Halpern, D., Chiapello, H., Schbath, S., Robin, S., Hennequet-Antier, C., Gruss, A., and El Karoui, M. (2007). Identification of DNA motifs implicated in maintenance of bacterial core genomes by predictive modeling. PLoS Genet. *3*, 1614–1621.

90. Han, E.S., Cooper, D.L., Persky, N.S., Sutera, V.A., Whitaker, R.D., Montello, M.L., and Lovett, S.T. (2006). RecJ exonuclease: substrates, products and interaction with SSB. Nucleic Acids Res. *34*, 1084–1091.

91. Handa, N., Bianco, P.R., Baskin, R.J., and Kowalczykowski, S.C. (2005). Direct visualization of RecBCD movement reveals cotranslocation of the RecD motor after chi recognition. Mol. Cell *17*, 745–750.

92. Handa, N., Morimatsu, K., Lovett, S.T., and Kowalczykowski, S.C. (2009). Reconstitution of initial steps of dsDNA break repair by the RecF pathway of E. coli. Genes Dev *23*, 1234–1245.

93. Harmon, F.G., and Kowalczykowski, S.C. (2001). Biochemical characterization of the DNA helicase activity of the escherichia coli RecQ helicase. J. Biol. Chem. *276*, 232–243.

94. Haurwitz, R.E., Jinek, M., Wiedenheft, B., Zhou, K., and Doudna, J.A. (2010). Sequence- and structure-specific RNA processing by a CRISPR endonuclease. Science *329*, 1355–1358.

95. Hayes, R.P., Xiao, Y., Ding, F., van Erp, P.B.G., Rajashankar, K., Bailey, S., Wiedenheft, B., and Ke, A. (2016). Structural basis for promiscuous PAM recognition in type I-E Cascade from E. coli. Nature *530*, 499–503.

96. Heller, R.C., and Marians, K.J. (2006). Replisome assembly and the direct restart of stalled replication forks. Nat. Rev. Mol. Cell Biol. *7*, 932–943.

97. Hochstrasser, M.L., Taylor, D.W., Bhat, P., Guegler, C.K., Sternberg, S.H., Nogales, E., and Doudna, J.A. (2014). CasA mediates Cas3-catalyzed target degradation during CRISPR RNA-guided interference. Proc Natl Acad Sci U S A *111*, 6618–6623.

98. Horii, Z., and Clark, A.J. (1973). Genetic analysis of the recF pathway to genetic recombination in Escherichia coli K12: isolation and characterization of mutants. J. Mol. Biol. *80*, 327–344.

99. Horvath, P., Romero, D.A., Coûté-Monvoisin, A.-C., Richards, M., Deveau, H., Moineau, S., Boyaval, P., Fremaux, C., and Barrangou, R. (2008). Diversity, activity, and evolution of CRISPR loci in Streptococcus thermophilus. J. Bacteriol. *190*, 1401–1412.

100. 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.

101. Huo, Y., Nam, K.H., Ding, F., Lee, H., Wu, L., Xiao, Y., Farchione, F.D., Zhou, S., Rajashankar, R., Kurinov, I., et al. (2014). Structures of CRISPR Cas3 offer mechanistic insights into Cascade-activated DNA unwinding and degradation. Nat Struct Mol Biol *21*, 771–777.

102. Ishino, Y., Shinagawa, H., Makino, K., Amemura, M., and Nakata, A. (1987). Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. J. Bacteriol. *169*, 5429–5433.

103. 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.

104. Jackson, R.N., Golden, S.M., van Erp, P.B.G., Carter, J., Westra, E.R., Brouns, S.J.J., van der Oost, J., Terwilliger, T.C., Read, R.J., and Wiedenheft, B. (2014a). Structural biology. Crystal structure of the CRISPR RNA-guided surveillance complex from Escherichia coli. Science *345*, 1473–1479.

105. Jackson, R.N., Lavin, M., Carter, J., and Wiedenheft, B. (2014b). Fitting CRISPR-associated Cas3 into the helicase family tree. Curr. Opin. Struct. Biol. *24*, 106–114.

106. Jansen, R., Embden, J.D.A. van, Gaastra, W., and Schouls, L.M. (2002). Identification of genes that are associated with DNA repeats in prokaryotes. Mol. Microbiol. *43*, 1565–1575.

107. 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.

108. Jore, M.M., Lundgren, M., van Duijn, E., Bultema, J.B., Westra, E.R., Waghmare, S.P., Wiedenheft, B., Pul, U., Wurm, R., Wagner, R., et al. (2011). Structural basis for CRISPR RNA-guided DNA recognition by Cascade. Nat. Struct. Mol. Biol. *18*, 529–536.

109. Joseph, J.W., and Kolodner, R. (1983a). Exonuclease VIII of Escherichia coli. I. Purification and physical properties. J. Biol. Chem. *258*, 10411–10417.

110. Joseph, J.W., and Kolodner, R. (1983b). Exonuclease VIII of Escherichia coli. II. Mechanism of action. J. Biol. Chem. *258*, 10418–10424.

111. Ka, D., Kim, D., Baek, G., and Bae, E. (2014). Structural and functional characterization of Streptococcus pyogenes Cas2 protein under different pH conditions. Biochem. Biophys. Res. Commun. *451*, 152–157.

112. Ka, D., Hong, S., Jeong, U., Jeong, M., Suh, N., Suh, J.-Y., and Bae, E. (2017). Structural and dynamic insights into the role of conformational switching in the nuclease

activity of the Xanthomonas albilineans Cas2 in CRISPR-mediated adaptive immunity. Struct Dyn *4*, 054701.

113. Kaiser, K., and Murray, N.E. (1979). Physical characterisation of the "Rac prophage" in E. coli K12. Mol. Gen. Genet. *175*, 159–174.

114. Karu, A.E., and Linn, S. (1972). Uncoupling of the recBC ATPase from DNase by DNA crosslinked with psoralen. Proc. Natl. Acad. Sci. U.S.A. *69*, 2855–2859.

115. Karu, A.E., MacKay, V., Goldmark, P.J., and Linn, S. (1973). The recBC deoxyribonuclease of Escherichia coli K-12. Substrate specificity and reaction intermediates. J. Biol. Chem. *248*, 4874–4884.

116. Karu, A.E., Sakaki, Y., Echols, H., and Linn, S. (1975). The gamma protein specified by bacteriophage gamma. Structure and inhibitory activity for the recBC enzyme of Escherichia coli. J. Biol. Chem. *250*, 7377–7387.

117. Kazlauskiene, M., Tamulaitis, G., Kostiuk, G., Venclovas, Č., and Siksnys, V. (2016). Spatiotemporal Control of Type III-A CRISPR-Cas Immunity: Coupling DNA Degradation with the Target RNA Recognition. Mol. Cell *62*, 295–306.

118. Kieper, S.N., Almendros, C., Behler, J., McKenzie, R.E., Nobrega, F.L., Haagsma, A.C., Vink, J.N.A., Hess, W.R., and Brouns, S.J.J. (2018). Cas4 Facilitates PAM-Compatible Spacer Selection during CRISPR Adaptation. Cell Rep *22*, 3377–3384.

119. Kim, S., Loeff, L., Colombo, S., Jergic, S., Brouns, S.J.J., and Joo, C. (2020). Selective loading and processing of prespacers for precise CRISPR adaptation. Nature *579*, 141–145.

120. Kobayashi, I., Murialdo, H., Crasemann, J.M., Stahl, M.M., and Stahl, F.W. (1982). Orientation of cohesive end site cos determines the active orientation of chi sequence in stimulating recA . recBC-mediated recombination in phage lambda lytic infections. Proc. Natl. Acad. Sci. U.S.A. *79*, 5981–5985.

121. Korangy, F., and Julin, D.A. (1994). Efficiency of ATP hydrolysis and DNA unwinding by the RecBC enzyme from Escherichia coli. Biochemistry *33*, 9552–9560.

122. Krasin, F., and Hutchinson, F. (1977). Repair of DNA double-strand breaks in Escherichia coli, which requires recA function and the presence of a duplicate genome. J. Mol. Biol. *116*, 81–98.

123. Künne, T., Kieper, S.N., Bannenberg, J.W., Vogel, A.I.M., Miellet, W.R., Klein, M., Depken, M., Suarez-Diez, M., and Brouns, S.J.J. (2016). Cas3-Derived Target DNA Degradation Fragments Fuel Primed CRISPR Adaptation. Mol. Cell *63*, 852–864.

124. Kurilovich, E., Shiriaeva, A., Metlitskaya, A., Morozova, N., Ivancic-Bace, I., Severinov, K., and Savitskaya, E. (2019). Genome Maintenance Proteins Modulate Autoimmunity Mediated Primed Adaptation by the Escherichia coli Type I-E CRISPR-Cas System. Genes *10*, 872.

125. Kushner, S.R., Nagaishi, H., Templin, A., and Clark, A.J. (1971). Genetic Recombination in Escherichia coli: The Role of Exonuclease I. PNAS *68*, 824–827.

126. Kushner, S.R., Nagaishi, H., and Clark, A.J. (1974). Isolation of exonuclease VIII: the enzyme associated with sbcA indirect suppressor. Proc. Natl. Acad. Sci. U.S.A. *71*, 3593–3597.

127. Kuzminov, A. (1995). Collapse and repair of replication forks in Escherichia coli. Mol. Microbiol. *16*, 373–384.

128. Kuzminov, A. (2001). Single-strand interruptions in replicating chromosomes cause double-strand breaks. Proc. Natl. Acad. Sci. U.S.A. *98*, 8241–8246.

129. Kuzminov, A., and Stahl, F.W. (1997). Stability of linear DNA in recA mutant Escherichia coli cells reflects ongoing chromosomal DNA degradation. J. Bacteriol. *179*, 880–888.

130. Kuzminov, A., Schabtach, E., and Stahl, F.W. (1994). Chi sites in combination with RecA protein increase the survival of linear DNA in Escherichia coli by inactivating exoV activity of RecBCD nuclease. EMBO J. *13*, 2764–2776.

131. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

132. 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.e5.

133. Lee, H., Dhingra, Y., and Sashital, D.G. (2019). The Cas4-Cas1-Cas2 complex mediates precise prespacer processing during CRISPR adaptation. Elife *8*.

134. Lehman, I.R. (1960). The deoxyribonucleases of Escherichia coli. I. Purification and properties of a phosphodiesterase. J. Biol. Chem. *235*, 1479–1487.

135. Lehman, I.R., and Nussbaum, A.L. (1964). THE DEOXYRIBONUCLEASES OF ESCHERICHIA COLI. V. ON THE SPECIFICITY OF EXONUCLEASE I (PHOSPHODIESTERASE). J. Biol. Chem. *239*, 2628–2636.

136. Lemak, S., Beloglazova, N., Nocek, B., Skarina, T., Flick, R., Brown, G., Popovic, A., Joachimiak, A., Savchenko, A., and Yakunin, A.F. (2013). Toroidal structure and DNA cleavage by the CRISPR-associated [4Fe-4S] cluster containing Cas4 nuclease SSO0001 from Sulfolobus solfataricus. J. Am. Chem. Soc. *135*, 17476–17487.

137. Lemak, S., Nocek, B., Beloglazova, N., Skarina, T., Flick, R., Brown, G., Joachimiak, A., Savchenko, A., and Yakunin, A.F. (2014). The CRISPR-associated Cas4 protein Pcal_0546 from Pyrobaculum calidifontis contains a [2Fe-2S] cluster: crystal structure and nuclease activity. Nucleic Acids Res. *42*, 11144–11155.

138. Lesterlin, C., Ball, G., Schermelleh, L., and Sherratt, D.J. (2014). RecA bundles mediate homology pairing between distant sisters during DNA break repair. Nature *506*, 249–253.

139. Levy, A., Goren, M.G., Yosef, I., Auster, O., Manor, M., Amitai, G., Edgar, R., Qimron, U., and Sorek, R. (2015). CRISPR adaptation biases explain preference for acquisition of foreign DNA. Nature *520*, 505–510.

140. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

141. Li, M., Wang, R., Zhao, D., and Xiang, H. (2014). Adaptation of the Haloarcula hispanica CRISPR-Cas system to a purified virus strictly requires a priming process. Nucleic Acids Res. *42*, 2483–2492.

142. Lim, C.T., Lai, P.J., Leach, D.R.F., Maki, H., and Furukohri, A. (2015). A novel mode of nuclease action is revealed by the bacterial Mre11/Rad50 complex. Nucleic Acids Res. *43*, 9804–9816.

143. Lintner, N.G., Kerou, M., Brumfield, S.K., Graham, S., Liu, H., Naismith, J.H., Sdano, M., Peng, N., She, Q., Copié, V., et al. (2011). Structural and functional characterization of an archaeal clustered regularly interspaced short palindromic repeat (CRISPR)-associated complex for antiviral defense (CASCADE). J. Biol. Chem. *286*, 21643–21656.

144. Liu, T., Li, Y., Wang, X., Ye, Q., Li, H., Liang, Y., She, Q., and Peng, N. (2015). Transcriptional regulator-mediated activation of adaptation genes triggers CRISPR de novo spacer acquisition. Nucleic Acids Res. *43*, 1044–1055.

145. Liu, T., Liu, Z., Ye, Q., Pan, S., Wang, X., Li, Y., Peng, W., Liang, Y., She, Q., and Peng, N. (2017). Coupling transcriptional activation of CRISPR-Cas system and DNA repair genes by Csa3a in Sulfolobus islandicus. Nucleic Acids Res. *45*, 8978–8992.

146. Lloyd, R.G., and Buckman, C. (1985). Identification and genetic analysis of sbcC mutations in commonly used recBC sbcB strains of Escherichia coli K-12. Journal of Bacteriology *164*, 836–844.

147. Lloyd, R.G., Porton, M.C., and Buckman, C. (1988). Effect of recF, recJ, recN, recO and ruv mutations on ultraviolet survival and genetic recombination in a recD strain of Escherichia coli K12. Mol. Gen. Genet. *212*, 317–324.

148. Loeff, L., Brouns, S.J.J., and Joo, C. (2018). Repetitive DNA Reeling by the Cascade-Cas3 Complex in Nucleotide Unwinding Steps. Mol. Cell *70*, 385-394.e3.

149. López-Amorós, R., Comas, J., and Vives-Rego, J. (1995). Flow cytometric assessment of Escherichia coli and Salmonella typhimurium starvation-survival in seawater using rhodamine 123, propidium iodide, and oxonol. Appl. Environ. Microbiol. *61*, 2521–2526.

150. Lovett, S.T., and Clark, A.J. (1984). Genetic analysis of the recJ gene of Escherichia coli K-12. J. Bacteriol. *157*, 190–196.

151. Lovett, S.T., and Kolodner, R.D. (1989). Identification and purification of a single-stranded-DNA-specific exonuclease encoded by the recJ gene of Escherichia coli. Proc. Natl. Acad. Sci. U.S.A. *86*, 2627–2631.

152. Lovett, S.T., Luisi-DeLuca, C., and Kolodner, R.D. (1988). The genetic dependence of recombination in recD mutants of Escherichia coli. Genetics *120*, 37–45.

153. Low, B. (1973). Restoration by the rac locus of recombinant forming ability in recB - and recC - merozygotes of Escherichia coli K-12. Mol. Gen. Genet. *122*, 119–130.

154. Luisi-DeLuca, C., Lovett, S.T., and Kolodner, R.D. (1989). Genetic and physical analysis of plasmid recombination in recB recC sbcB and recB recC sbcA Escherichia coli K-12 mutants. Genetics *122*, 269–278.

155. MacKay, V., and Linn, S. (1974). The mechanism of degradation of duplex deoxyribonucleic acid by the recBC enzyme of Escherichia coli K-12. J. Biol. Chem. *249*, 4286–4294.

156. Mahdi, A.A., and Lloyd, R.G. (1989). Identification of the recR locus of Escherichia coli K-12 and analysis of its role in recombination and DNA repair. Mol. Gen. Genet. *216*, 503–510.

157. Majsec, K., Bolt, E.L., and Ivančić-Baće, I. (2016). Cas3 is a limiting factor for CRISPR-Cas immunity in Escherichia coli cells lacking H-NS. BMC Microbiol. *16*, 28.

158. Majumdar, S., and Terns, M.P. (2019). CRISPR RNA-guided DNA cleavage by reconstituted Type I-A immune effector complexes. Extremophiles *23*, 19–33.

159. Majumdar, S., Zhao, P., Pfister, N.T., Compton, M., Olson, S., Glover, C.V.C., Wells, L., Graveley, B.R., Terns, R.M., and Terns, M.P. (2015). Three CRISPR-Cas immune effector complexes coexist in Pyrococcus furiosus. RNA *21*, 1147–1158.

160. Majumdar, S., Ligon, M., Skinner, W.C., Terns, R.M., and Terns, M.P. (2017). Target DNA recognition and cleavage by a reconstituted Type I-G CRISPR-Cas immune effector complex. Extremophiles *21*, 95–107.

161. 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.

162. Makarova, K.S., Haft, D.H., Barrangou, R., Brouns, S.J.J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F.J.M., Wolf, Y.I., Yakunin, A.F., et al. (2011a). Evolution and classification of the CRISPR-Cas systems. Nat. Rev. Microbiol. *9*, 467–477.

163. Makarova, K.S., Aravind, L., Wolf, Y.I., and Koonin, E.V. (2011b). Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. Biol. Direct *6*, 38.

164. Makarova, K.S., Wolf, Y.I., Alkhnbashi, O.S., Costa, F., Shah, S.A., Saunders, S.J., Barrangou, R., Brouns, S.J.J., Charpentier, E., Haft, D.H., et al. (2015). An updated evolutionary classification of CRISPR-Cas systems. Nat. Rev. Microbiol. *13*, 722–736.

165. Makarova, K.S., Wolf, Y.I., Iranzo, J., Shmakov, S.A., Alkhnbashi, O.S., Brouns, S.J.J., Charpentier, E., Cheng, D., Haft, D.H., Horvath, P., et al. (2020). Evolutionary classification of CRISPR-Cas systems: a burst of class 2 and derived variants. Nat. Rev. Microbiol. *18*, 67–83.

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

167. Masterson, C., Boehmer, P.E., McDonald, F., Chaudhuri, S., Hickson, I.D., and Emmerson, P.T. (1992). Reconstitution of the activities of the RecBCD holoenzyme of Escherichia coli from the purified subunits. J. Biol. Chem. *267*, 13564–13572.

168. McEntee, K., Weinstock, G.M., and Lehman, I.R. (1979). Initiation of general recombination catalyzed in vitro by the recA protein of Escherichia coli. Proc. Natl. Acad. Sci. U.S.A. *76*, 2615–2619.

169. Meddows, T.R., Savory, A.P., Grove, J.I., Moore, T., and Lloyd, R.G. (2005). RecN protein and transcription factor DksA combine to promote faithful recombinational repair of DNA double-strand breaks. Molecular Microbiology *57*, 97–110.

170. Menon, A.L., Poole, F.L., Cvetkovic, A., Trauger, S.A., Kalisiak, E., Scott, J.W., Shanmukh, S., Praissman, J., Jenney, F.E., Wikoff, W.R., et al. (2009). Novel multiprotein complexes identified in the hyperthermophilic archaeon Pyrococcus furiosus by non-denaturing fractionation of the native proteome. Mol. Cell Proteomics *8*, 735–751.

171. Michel, B., and Sandler, S.J. (2017). Replication Restart in Bacteria. Journal of Bacteriology *199*.

172. 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.

173. Mojica, F.J., Díez-Villaseñor, C., Soria, E., and Juez, G. (2000). Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. Mol. Microbiol. *36*, 244–246.

174. Mojica, F.J.M., Díez-Villaseñor, C., García-Martínez, J., and Soria, E. (2005). Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. J. Mol. Evol. *60*, 174–182.

175. Mojica, F.J.M., Díez-Villaseñor, C., García-Martínez, J., and Almendros, C. (2009). Short motif sequences determine the targets of the prokaryotic CRISPR defence system. Microbiology *155*, 733–740.

176. Moore, S.D. (2011). Assembling new Escherichia coli strains by transduction using phage P1. Methods Mol. Biol. *765*, 155–169.

177. Morgan, M., Anders, S., Lawrence, M., Aboyoun, P., Pagès, H., and Gentleman, R. (2009). ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. Bioinformatics *25*, 2607–2608.

178. Morimatsu, K., and Kowalczykowski, S.C. (2003). RecFOR proteins load RecA protein onto gapped DNA to accelerate DNA strand exchange: a universal step of recombinational repair. Mol. Cell *11*, 1337–1347.

179. Morimatsu, K., and Kowalczykowski, S.C. (2014). RecQ helicase and RecJ nuclease provide complementary functions to resect DNA for homologous recombination. Proc. Natl. Acad. Sci. U.S.A. *111*, E5133-5142.

180. Mulepati, S., and Bailey, S. (2013). In vitro reconstitution of an Escherichia coli RNA-guided immune system reveals unidirectional, ATP-dependent degradation of DNA target. J. Biol. Chem. *288*, 22184–22192.

181. Murphy, K.C., Fenton, A.C., and Poteete, A.R. (1987). Sequence of the bacteriophage P22 anti-recBCD (abc) genes and properties of P22 abc region deletion mutants. Virology *160*, 456–464.

182. Musharova, O., Klimuk, E., Datsenko, K.A., Metlitskaya, A., Logacheva, M., Semenova, E., Severinov, K., and Savitskaya, E. (2017). Spacer-length DNA intermediates are associated with Cas1 in cells undergoing primed CRISPR adaptation. Nucleic Acids Res. *45*, 3297–3307.

183. Musharova, O., Sitnik, V., Vlot, M., Savitskaya, E., Datsenko, K.A., Krivoy, A., Fedorov, I., Semenova, E., Brouns, S.J.J., and Severinov, K. (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.

184. Muskavitch, K.M., and Linn, S. (1982). A unified mechanism for the nuclease and unwinding activities of the recBC enzyme of Escherichia coli. J. Biol. Chem. *257*, 2641–2648.

185. Nam, K.H., Haitjema, C., Liu, X., Ding, F., Wang, H., DeLisa, M.P., and Ke, A. (2012a). Cas5d protein processes pre-crRNA and assembles into a cascade-like interference complex in subtype I-C/Dvulg CRISPR-Cas system. Structure *20*, 1574–1584.

186. Nam, K.H., Ding, F., Haitjema, C., Huang, Q., DeLisa, M.P., and Ke, A. (2012b). Double-stranded endonuclease activity in Bacillus halodurans clustered regularly interspaced short palindromic repeats (CRISPR)-associated Cas2 protein. J. Biol. Chem. *287*, 35943–35952.

187. Neylon, C., Kralicek, A.V., Hill, T.M., and Dixon, N.E. (2005). Replication termination in Escherichia coli: structure and antihelicase activity of the Tus-Ter complex. Microbiol. Mol. Biol. Rev. *69*, 501–526.

188. Nimkar, S., and Anand, B. (2020). Cas3/I-C mediated target DNA recognition and cleavage during CRISPR interference are independent of the composition and architecture of Cascade surveillance complex. Nucleic Acids Res. *48*, 2486–2501.

189. Nuñez, J.K., Kranzusch, P.J., Noeske, J., Wright, A.V., Davies, C.W., and Doudna, J.A. (2014). Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. Nat. Struct. Mol. Biol. *21*, 528–534.

190. Nuñez, J.K., Lee, A.S.Y., Engelman, A., and Doudna, J.A. (2015a). Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity. Nature *519*, 193–198.

191. Nuñez, J.K., Harrington, L.B., Kranzusch, P.J., Engelman, A.N., and Doudna, J.A. (2015b). Foreign DNA capture during CRISPR-Cas adaptive immunity. Nature *527*, 535–538.

192. Nuñez, J.K., Bai, L., Harrington, L.B., Hinder, T.L., and Doudna, J.A. (2016). CRISPR Immunological Memory Requires a Host Factor for Specificity. Mol. Cell *62*, 824–833.

193. Nussenzweig, P.M., McGinn, J., and Marraffini, L.A. (2019). Cas9 Cleavage of Viral Genomes Primes the Acquisition of New Immunological Memories. Cell Host Microbe *26*, 515-526.e6.

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

195. Oliver, D.B., and Goldberg, E.B. (1977). Protection of parental T4 DNA from a restriction exonuclease by the product of gene 2. J. Mol. Biol. *116*, 877–881.

196. O'Shea, J.P., Chou, M.F., Quader, S.A., Ryan, J.K., Church, G.M., and Schwartz, D. (2013). pLogo: a probabilistic approach to visualizing sequence motifs. Nat. Methods *10*, 1211–1212.

197. Palas, K.M., and Kushner, S.R. (1990). Biochemical and physical characterization of exonuclease V from Escherichia coli. Comparison of the catalytic activities of the RecBC and RecBCD enzymes. J. Biol. Chem. *265*, 3447–3454.

198. Phillips, R.J., Hickleton, D.C., Boehmer, P.E., and Emmerson, P.T. (1997). The RecB protein of Escherichia coli translocates along single-stranded DNA in the 3' to 5' direction: a proposed ratchet mechanism. Mol. Gen. Genet. *254*, 319–329.

199. Plagens, A., Tjaden, B., Hagemann, A., Randau, L., and Hensel, R. (2012). Characterization of the CRISPR/Cas subtype I-A system of the hyperthermophilic crenarchaeon Thermoproteus tenax. J. Bacteriol. *194*, 2491–2500.

200. Plagens, A., Tripp, V., Daume, M., Sharma, K., Klingl, A., Hrle, A., Conti, E., Urlaub, H., and Randau, L. (2014). In vitro assembly and activity of an archaeal CRISPR-Cas type I-A Cascade interference complex. Nucleic Acids Res. *42*, 5125–5138.

201. Ponticelli, A.S., Schultz, D.W., Taylor, A.F., and Smith, G.R. (1985). Chi-dependent DNA strand cleavage by RecBC enzyme. Cell *41*, 145–151.

202. Pougach, K., Semenova, E., Bogdanova, E., Datsenko, K.A., Djordjevic, M., Wanner, B.L., and Severinov, K. (2010). Transcription, processing and function of CRISPR cassettes in Escherichia coli. Mol. Microbiol. *77*, 1367–1379.

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

204. 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.

205. Radovcic, M., Killelea, T., Savitskaya, E., Wettstein, L., Bolt, E.L., and Ivancic-Bace, 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.

206. Ragunathan, K., Joo, C., and Ha, T. (2011). Real-time observation of strand exchange reaction with high spatiotemporal resolution. Structure *19*, 1064–1073.

207. Ramachandran, A., Summerville, L., Learn, B.A., DeBell, L., and Bailey, S. (2020). Processing and integration of functionally oriented prespacers in the Escherichia coli CRISPR system depends on bacterial host exonucleases. J. Biol. Chem. *295*, 3403–3414.

208. Rao, C., Guyard, C., Pelaz, C., Wasserscheid, J., Bondy-Denomy, J., Dewar, K., and Ensminger, A.W. (2016). Active and adaptive Legionella CRISPR-Cas reveals a recurrent challenge to the pathogen. Cell. Microbiol. *18*, 1319–1338.

209. 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.

210. Redding, S., Sternberg, S.H., Marshall, M., Gibb, B., Bhat, P., Guegler, C.K., Wiedenheft, B., Doudna, J.A., and Greene, E.C. (2015). Surveillance and Processing of Foreign DNA by the Escherichia coli CRISPR-Cas System. Cell *163*, 854–865.

211. Reeks, J., Naismith, J.H., and White, M.F. (2013). CRISPR interference: a structural perspective. Biochem. J. *453*, 155–166.

212. Repar, J., Briški, N., Buljubašić, M., Zahradka, K., and Zahradka, D. (2013). Exonuclease VII is involved in “reckless” DNA degradation in UV-irradiated Escherichia coli. Mutat. Res. *750*, 96–104.

213. Resnick, M.A. (1976). The repair of double-strand breaks in DNA; a model involving recombination. J. Theor. Biol. *59*, 97–106.

214. Rice, P.A., Yang, S., Mizuuchi, K., and Nash, H.A. (1996). Crystal structure of an IHF-DNA complex: a protein-induced DNA U-turn. Cell *87*, 1295–1306.

215. Richter, C., Gristwood, T., Clulow, J.S., and Fineran, P.C. (2012a). In vivo protein interactions and complex formation in the Pectobacterium atrosepticum subtype I-F CRISPR/Cas System. PLoS ONE *7*, e49549.

216. Richter, C., Dy, R.L., McKenzie, R.E., Watson, B.N.J., Taylor, C., Chang, J.T., McNeil, M.B., Staals, R.H.J., and Fineran, P.C. (2014). Priming in the Type I-F CRISPR-Cas system triggers strand-independent spacer acquisition, bi-directionally from the primed protospacer. Nucleic Acids Res. *42*, 8516–8526.

217. Richter, H., Zoephel, J., Schermuly, J., Maticzka, D., Backofen, R., and Randau, L. (2012b). Characterization of CRISPR RNA processing in Clostridium thermocellum and Methanococcus maripaludis. Nucleic Acids Res. *40*, 9887–9896.

218. Rinken, R., Thomas, B., and Wackernagel, W. (1992). Evidence that recBC-dependent degradation of duplex DNA in Escherichia coli recD mutants involves DNA unwinding. J. Bacteriol. *174*, 5424–5429.

219. 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.

220. Rollins, M.F., Schuman, J.T., Paulus, K., Bukhari, H.S.T., and Wiedenheft, B. (2015). Mechanism of foreign DNA recognition by a CRISPR RNA-guided surveillance complex from Pseudomonas aeruginosa. Nucleic Acids Res. *43*, 2216–2222.

221. Rollins, M.F., Chowdhury, S., Carter, J., Golden, S.M., Wilkinson, R.A., Bondy-Denomy, J., Lander, G.C., and Wiedenheft, B. (2017). Cas1 and the Csy complex are opposing regulators of Cas2/3 nuclease activity. Proc. Natl. Acad. Sci. U.S.A. *114*, E5113–E5121.

222. Roman, L.J., and Kowalczykowski, S.C. (1989a). Characterization of the adenosinetriphosphatase activity of the Escherichia coli RecBCD enzyme: relationship of ATP hydrolysis to the unwinding of duplex DNA. Biochemistry *28*, 2873–2881.

223. Roman, L.J., and Kowalczykowski, S.C. (1989b). Characterization of the helicase activity of the Escherichia coli RecBCD enzyme using a novel helicase assay. Biochemistry *28*, 2863–2873.

224. Roots, R., Kraft, G., and Gosschalk, E. (1985). The formation of radiation-induced DNA breaks: the ratio of double-strand breaks to single-strand breaks. Int. J. Radiat. Oncol. Biol. Phys. *11*, 259–265.

225. Rosamond, J., Telander, K.M., and Linn, S. (1979). Modulation of the action of the recBC enzyme of Escherichia coli K-12 by Ca2+. J. Biol. Chem. *254*, 8646–8652.

226. Rutkauskas, M., Sinkunas, T., Songailiene, I., Tikhomirova, M.S., Siksnys, V., and Seidel, R. (2015). Directional R-Loop Formation by the CRISPR-Cas Surveillance Complex Cascade Provides Efficient Off-Target Site Rejection. Cell Rep *10*, 1534–1543.

227. Sakai, A., and Cox, M.M. (2009). RecFOR and RecOR as distinct RecA loading pathways. J. Biol. Chem. *284*, 3264–3272.

228. Sakaki, Y., Karu, A.E., Linn, S., and Echols, H. (1973). Purification and properties of the gamma-protein specified by bacteriophage lambda: an inhibitor of the host RecBC recombination enzyme. Proc. Natl. Acad. Sci. U.S.A. *70*, 2215–2219.

229. Samai, P., Pyenson, N., Jiang, W., Goldberg, G.W., Hatoum-Aslan, A., and Marraffini, L.A. (2015). Co-transcriptional DNA and RNA Cleavage during Type III CRISPR-Cas Immunity. Cell *161*, 1164–1174.

230. Sashital, D.G., Wiedenheft, B., and Doudna, J.A. (2012). Mechanism of foreign DNA selection in a bacterial adaptive immune system. Mol. Cell *46*, 606–615.

231. 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.

232. Semenova, E., Jore, M.M., Datsenko, K.A., Semenova, A., Westra, E.R., Wanner, B., van der Oost, J., Brouns, S.J.J., and Severinov, K. (2011). Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. Proc. Natl. Acad. Sci. U.S.A. *108*, 10098–10103.

233. Semenova, E., Savitskaya, E., Musharova, O., Strotskaya, A., Vorontsova, D., Datsenko, K.A., Logacheva, M.D., and 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.

234. Shibata, T., Cunningham, R.P., DasGupta, C., and Radding, C.M. (1979). Homologous pairing in genetic recombination: complexes of recA protein and DNA. Proc. Natl. Acad. Sci. U.S.A. *76*, 5100–5104.

235. Shiimori, M., Garrett, S.C., Chambers, D.P., Glover, 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.

236. 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.e6.

237. Shipman, S.L., Nivala, J., Macklis, J.D., and Church, G.M. (2016). Molecular recordings by directed CRISPR spacer acquisition. Science *353*, aaf1175.

238. Shiriaeva, A., Semenova, E., and Severinov, K. (2018). CRISPR-Cas systems of bacteria and archaea. How did the components of prokaryotic adaptive immune systems become a tool for genome editing and controlling transcription. In Editing of Genes and Genomes., (Novosibirsk: Publishing House of Siberian Branch of the Russian Academy of Sciences), pp. 93–157.

239. Shiriaeva, A., Fedorov, I., Vyhovskyi, D., and Severinov, K. (2020). Detection of CRISPR adaptation. Biochem. Soc. Trans. *48*, 257–269.

240. Shiriaeva, A.A., Savitskaya, E., Datsenko, K.A., Vvedenskaya, I.O., Fedorova, I., Morozova, N., Metlitskaya, A., Sabantsev, A., Nickels, B.E., Severinov, K., et al. (2019).

Detection of spacer precursors formed in vivo during primed CRISPR adaptation. Nat Commun *10*, 4603.

241. 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.

242. 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*.

243. Silverstein, J.L., and Goldberg, E.B. (1976). T4 DNA injection. II. Protection of entering DNA from host exonuclease V. Virology *72*, 212–223.

244. Simmon, V.F., and Lederberg, S. (1972). Degradation of bacteriophage lambda deoxyribonucleic acid after restriction by Escherichia coli K-12. J. Bacteriol. *112*, 161–169.

245. Singleton, M.R., Dillingham, M.S., Gaudier, M., Kowalczykowski, S.C., and Wigley, D.B. (2004). Crystal structure of RecBCD enzyme reveals a machine for processing DNA breaks. Nature *432*, 187–193.

246. Sinha, A.K., Possoz, C., and Leach, D.R.F. (2020). The Roles of Bacterial DNA Double-Strand Break Repair Proteins in Chromosomal DNA Replication. FEMS Microbiol. Rev. *44*, 351–368.

247. Sinkunas, T., Gasiunas, G., Fremaux, C., Barrangou, R., Horvath, P., and Siksnys, V. (2011). Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. EMBO J. *30*, 1335–1342.

248. Sinkunas, T., Gasiunas, G., Waghmare, S.P., Dickman, M.J., Barrangou, R., Horvath, P., and Siksnys, V. (2013). In vitro reconstitution of Cascade-mediated CRISPR immunity in Streptococcus thermophilus. The EMBO Journal *32*, 385–394.

249. Smith, G.R. (1989). Homologous recombination in E. coli: multiple pathways for multiple reasons. Cell *58*, 807–809.

250. Smith, G.R., Kunes, S.M., Schultz, D.W., Taylor, A., and Triman, K.L. (1981). Structure of chi hotspots of generalized recombination. Cell *24*, 429–436.

251. Spies, M., Bianco, P.R., Dillingham, M.S., Handa, N., Baskin, R.J., and Kowalczykowski, S.C. (2003). A molecular throttle: the recombination hotspot chi controls DNA translocation by the RecBCD helicase. Cell *114*, 647–654.

252. Spies, M., Dillingham, M.S., and Kowalczykowski, S.C. (2005). Translocation by the RecB motor is an absolute requirement for {chi}-recognition and RecA protein loading by RecBCD enzyme. J. Biol. Chem. *280*, 37078–37087.

253. Spies, M., Amitani, I., Baskin, R.J., and Kowalczykowski, S.C. (2007). RecBCD enzyme switches lead motor subunits in response to chi recognition. Cell *131*, 694–705.

254. 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.

255. Stahl, F.W., and Stahl, M.M. (1977). Recombination pathway specificity of Chi. Genetics *86*, 715–725.

256. Stahl, F.W., Stahl, M.M., Malone, R.E., and Crasemann, J.M. (1980). Directionality and nonreciprocality of Chi-stimulated recombination in phage lambda. Genetics *94*, 235–248.

257. Sternberg, S.H., Redding, S., Jinek, M., Greene, E.C., and Doudna, J.A. (2014). DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. Nature *507*, 62–67.

258. Stocks, S.M. (2004). Mechanism and use of the commercially available viability stain, BacLight. Cytometry A *61*, 189–195.

259. Strotskaya, A., Savitskaya, E., Metlitskaya, A., Morozova, N., Datsenko, K.A., Semenova, E., and 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.

260. 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.

261. Szostak, J.W., Orr-Weaver, T.L., Rothstein, R.J., and Stahl, F.W. (1983). The double-strand-break repair model for recombination. Cell *33*, 25–35.

262. Taylor, A., and Smith, G.R. (1980). Unwinding and rewinding of DNA by the RecBC enzyme. Cell *22*, 447–457.

263. Taylor, A.F., and Smith, G.R. (1985). Substrate specificity of the DNA unwinding activity of the RecBC enzyme of Escherichia coli. J. Mol. Biol. *185*, 431–443.

264. Taylor, A.F., and Smith, G.R. (1995a). Monomeric RecBCD enzyme binds and unwinds DNA. J. Biol. Chem. *270*, 24451–24458.

265. Taylor, A.F., and Smith, G.R. (1995b). Strand specificity of nicking of DNA at Chi sites by RecBCD enzyme. Modulation by ATP and magnesium levels. J. Biol. Chem. *270*, 24459–24467.

266. Taylor, A.F., and Smith, G.R. (2003). RecBCD enzyme is a DNA helicase with fast and slow motors of opposite polarity. Nature *423*, 889–893.

267. Taylor, A.F., Schultz, D.W., Ponticelli, A.S., and Smith, G.R. (1985). RecBC enzyme nicking at Chi sites during DNA unwinding: location and orientation-dependence of the cutting. Cell *41*, 153–163.

268. Tseng, Y.C., Hung, J.L., and Wang, T.C. (1994). Involvement of RecF pathway recombination genes in postreplication repair in UV-irradiated Escherichia coli cells. Mutat. Res. *315*, 1–9.

269. Umezu, K., and Kolodner, R.D. (1994). Protein interactions in genetic recombination in Escherichia coli. Interactions involving RecO and RecR overcome the inhibition of RecA by single-stranded DNA-binding protein. J. Biol. Chem. *269*, 30005–30013.

270. Umezu, K., Nakayama, K., and Nakayama, H. (1990). Escherichia coli RecQ protein is a DNA helicase. Proc. Natl. Acad. Sci. U.S.A. *87*, 5363–5367.

271. Umezu, K., Chi, N.W., and Kolodner, R.D. (1993). Biochemical interaction of the Escherichia coli RecF, RecO, and RecR proteins with RecA protein and single-stranded DNA binding protein. Proc. Natl. Acad. Sci. U.S.A. *90*, 3875–3879.

272. Vischer, N.O.E., Verheul, J., Postma, M., van den Berg van Saparoea, B., Galli, E., Natale, P., Gerdes, K., Luirink, J., Vollmer, W., Vicente, M., et al. (2015). Cell age

dependent concentration of Escherichia coli divisome proteins analyzed with ImageJ and ObjectJ. Front Microbiol *6*, 586.

273. Vorontsova, D., Datsenko, K.A., Medvedeva, S., Bondy-Denomy, J., Savitskaya, E.E., Pougach, K., Logacheva, M., Wiedenheft, B., Davidson, A.R., Severinov, 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.

274. Vvedenskaya, I.O., Goldman, S.R., and Nickels, B.E. (2015). Preparation of cDNA libraries for high-throughput RNA sequencing analysis of RNA 5' ends. Methods Mol. Biol. *1276*, 211–228.

275. Wagih, O. (2017). ggseqlogo: a versatile R package for drawing sequence logos. Bioinformatics *33*, 3645–3647.

276. Wang, J., Chen, R., and Julin, D.A. (2000). A single nuclease active site of the Escherichia coli RecBCD enzyme catalyzes single-stranded DNA degradation in both directions. J. Biol. Chem. *275*, 507–513.

277. Wang, J., Li, J., Zhao, H., Sheng, G., Wang, M., Yin, M., and Wang, Y. (2015). Structural and Mechanistic Basis of PAM-Dependent Spacer Acquisition in CRISPR-Cas Systems. Cell *163*, 840–853.

278. West, S.C., Cassuto, E., Mursalim, J., and Howard-Flanders, P. (1980). Recognition of duplex DNA containing single-stranded regions by recA protein. Proc. Natl. Acad. Sci. U.S.A. *77*, 2569–2573.

279. Westra, E.R., Pul, U., Heidrich, N., Jore, M.M., Lundgren, M., Stratmann, T., Wurm, R., Raine, A., Mescher, M., Van Heereveld, L., et al. (2010). H-NS-mediated

repression of CRISPR-based immunity in Escherichia coli K12 can be relieved by the transcription activator LeuO. Mol. Microbiol. *77*, 1380–1393.

280. Westra, E.R., van Erp, P.B.G., Künne, T., Wong, S.P., Staals, R.H.J., Seegers, C.L.C., Bollen, S., Jore, M.M., Semenova, E., Severinov, K., 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.

281. Westra, E.R., Semenova, E., Datsenko, K.A., Jackson, R.N., Wiedenheft, B., Severinov, K., and Brouns, S.J.J. (2013). Type I-E CRISPR-cas systems discriminate target from non-target DNA through base pairing-independent PAM recognition. PLoS Genet. *9*, e1003742.

282. Wiedenheft, B., van Duijn, E., Bultema, J.B., Bultema, J., Waghmare, S.P., Waghmare, S., Zhou, K., Barendregt, A., Westphal, W., Heck, A.J.R., et al. (2011a). RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. Proc. Natl. Acad. Sci. U.S.A. *108*, 10092–10097.

283. Wiedenheft, B., Lander, G.C., Zhou, K., Jore, M.M., Brouns, S.J.J., van der Oost, J., Doudna, J.A., and Nogales, E. (2011b). Structures of the RNA-guided surveillance complex from a bacterial immune system. Nature *477*, 486–489.

284. Willetts, N.S., and Mount, D.W. (1969). Genetic analysis of recombination-deficient mutants of Escherichia coli K-12 carrying rec mutations cotransducible with thyA. J. Bacteriol. *100*, 923–934.

285. Wilson, J.E., and Chin, A. (1991). Chelation of divalent cations by ATP, studied by titration calorimetry. Anal. Biochem. *193*, 16–19.

286. 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.

287. Wright, M., Buttin, G., and Hurwitz, J. (1971). The Isolation and Characterization from Escherichia coli of an Adenosine Triphosphate-dependent Deoxyribonuclease Directed by rec B, C Genes. J. Biol. Chem. *246*, 6543–6555.

288. Xiao, Y., Luo, M., Hayes, R.P., Kim, J., Ng, S., Ding, F., Liao, M., and Ke, A. (2017). Structure Basis for Directional R-loop Formation and Substrate Handover Mechanisms in Type I CRISPR-Cas System. Cell *170*, 48-60.e11.

289. Xiao, Y., Luo, M., Dolan, A.E., Liao, M., and Ke, A. (2018). Structure basis for RNA-guided DNA degradation by Cascade and Cas3. Science *361*.

290. Xu, L., and Marians, K.J. (2002). A dynamic RecA filament permits DNA polymerase-catalyzed extension of the invading strand in recombination intermediates. J. Biol. Chem. *277*, 14321–14328.

291. Xue, C., Seetharam, A.S., Musharova, O., Severinov, K., Brouns, S.J.J., Severin, A.J., and Sashital, D.G. (2015). CRISPR interference and priming varies with individual spacer sequences. Nucleic Acids Res. *43*, 10831–10847.

292. Xue, C., Whitis, N.R., and Sashital, D.G. (2016). Conformational Control of Cascade Interference and Priming Activities in CRISPR Immunity. Mol. Cell *64*, 826–834.

293. Xue, C., Zhu, Y., Zhang, X., Shin, Y.-K., and Sashital, D.G. (2017). Real-Time Observation of Target Search by the CRISPR Surveillance Complex Cascade. Cell Rep *21*, 3717–3727.

294. Yang, C.-D., Chen, Y.-H., Huang, H.-Y., Huang, H.-D., and Tseng, C.-P. (2014). CRP represses the CRISPR/Cas system in Escherichia coli: evidence that endogenous CRISPR spacers impede phage P1 replication. Mol. Microbiol. *92*, 1072–1091.

295. Yoganand, K.N., Muralidharan, M., Nimkar, S., and Anand, B. (2019). Fidelity of prespacer capture and processing is governed by the PAM-mediated interactions of Cas1-2 adaptation complex in CRISPR-Cas type I-E system. J. Biol. Chem. *294*, 20039–20053.

296. Yoganand, K.N.R., Sivathanu, R., Nimkar, S., and Anand, B. (2017). 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. *45*, 367–381.

297. 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.

298. 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*, 14396–14401.

299. Yu, M., Souaya, J., and Julin, D.A. (1998a). Identification of the nuclease active site in the multifunctional RecBCD enzyme by creation of a chimeric enzyme. J. Mol. Biol. *283*, 797–808.

300. Yu, M., Souaya, J., and Julin, D.A. (1998b). The 30-kDa C-terminal domain of the RecB protein is critical for the nuclease activity, but not the helicase activity, of the RecBCD enzyme from Escherichia coli. PNAS *95*, 981–986.

301. Zhang, J., Kasciukovic, T., and White, M.F. (2012). The CRISPR Associated Protein Cas4 Is a 5′ to 3′ DNA Exonuclease with an Iron-Sulfur Cluster. PLOS ONE *7*, e47232.

302. Zhang, X.-D., Dou, S.-X., Xie, P., Hu, J.-S., Wang, P.-Y., and Xi, X.G. (2006). Escherichia coli RecQ is a rapid, efficient, and monomeric helicase. J. Biol. Chem. *281*, 12655–12663.

303. Zhang, Z., Pan, S., Liu, T., Li, Y., and Peng, N. (2019). Cas4 nucleases can effect specific integration of CRISPR spacers. J. Bacteriol.

APPENDICES

Table 1. Strains used in this study

Name	Description	Source
KD403	K-12 F^+, *lacUV5-cas3 araBp8-cse1*, CRISPR I: repeat-Sp^{yihN}-repeat, CRISPR II deleted. Sp^{yihN} (TCAAACAACCGACCTTGTTGTTCGCTATTGCC) targets chromosomal protospacer PPS (CCAAACAACCGACCTTGTTGTTCGCTATTGCC) within *yihN* gene forming a mismatch between crRNA and PPS at position +1.	This study
KD518	Like KD403, except Cas1 H208A	This study
KD753	Like KD403, except Cas3 H74A	This study
KD263	Like KD403, except CRISPR I: repeat-Sp^{M13}-repeat. Sp^{M13} (CTGTCTTTCGCTGCTGAGGGTGACGATCCCGC) targets g8 gene of M13 phage.	Shmakov et al., 2014
ΔrecB	Like KD403, except *recB::FRT*	This study
ΔrecC	Like KD403, except *recC::FRT*	This study
ΔrecD	Like KD403, except *recD::FRT*	This study
ΔrecJ	Like KD403, except *recJ::FRT-kan-FRT*	This study
ΔsbcB	Like KD403, except *sbcB::FRT*	This study
ΔsbcD	Like KD403, except *sbcD::FRT*	This study
ΔrecB ΔrecJ	Like KD403, except *recB::FRT recJ::FRT-kan-FRT*	This study
ΔrecB ΔsbcD	Like KD403, except *sbcD::FRT recB::FRT-kan-FRT*	This study
ΔrecB ΔsbcB	Like KD403, except *sbcB::FRT recB::FRT-kan-FRT*	This study

BL21-AI	F$^-$*ompT hsdSB (rB- mB-) gal dcm araB::T7RNAP-tetA*	Invitrogen
KD675	BL21-AI_ΔCRISPR carrying *Pseudomonas aeruginosa* CRISPR array with a single spacer (ACGCAGTTGCTGAGTGTGATCGATGCCATCAG) and a protospacer with a mismatch at position +1 (TCGCAGTTGCTGAGTGTGATCGATGCCATCAG) preceded by a functional GG PAM introduced into ompL/yihN intergenic region corresponding to the positions 4372171-4372261 of NC_012947	Vorontsova et al., 2015

Table 2. List of primers used for amplification of CRISPR arrays.

Name	Sequence (5' to 3')	Purpose
LDR-F2	ATGCTTTAAGAACAAATGTATACTTTTAG	Monitoring of primed adaptation in KD263 and KD403
Ec_minR	CGAAGGCGTCTTGATGGGTTTG	
LDR-F2	ATGCTTTAAGAACAAATGTATACTTTTAG	Monitoring of primed adaptation and high-throughput sequencing of spacers from KD403 and its DNA repair mutant derivatives.
autoSp2_R	AATAGCGAACAACAAGGTCGGTTG	
BLCRdir	GGTAGATTGTGACTGGCTTAAAAAATC	High-throughput sequencing of spacers acquired during prespacer efficiency assay in BL21-AI
BLCRreverse	GTTTGAGCGATGATATTTGTGCTC	

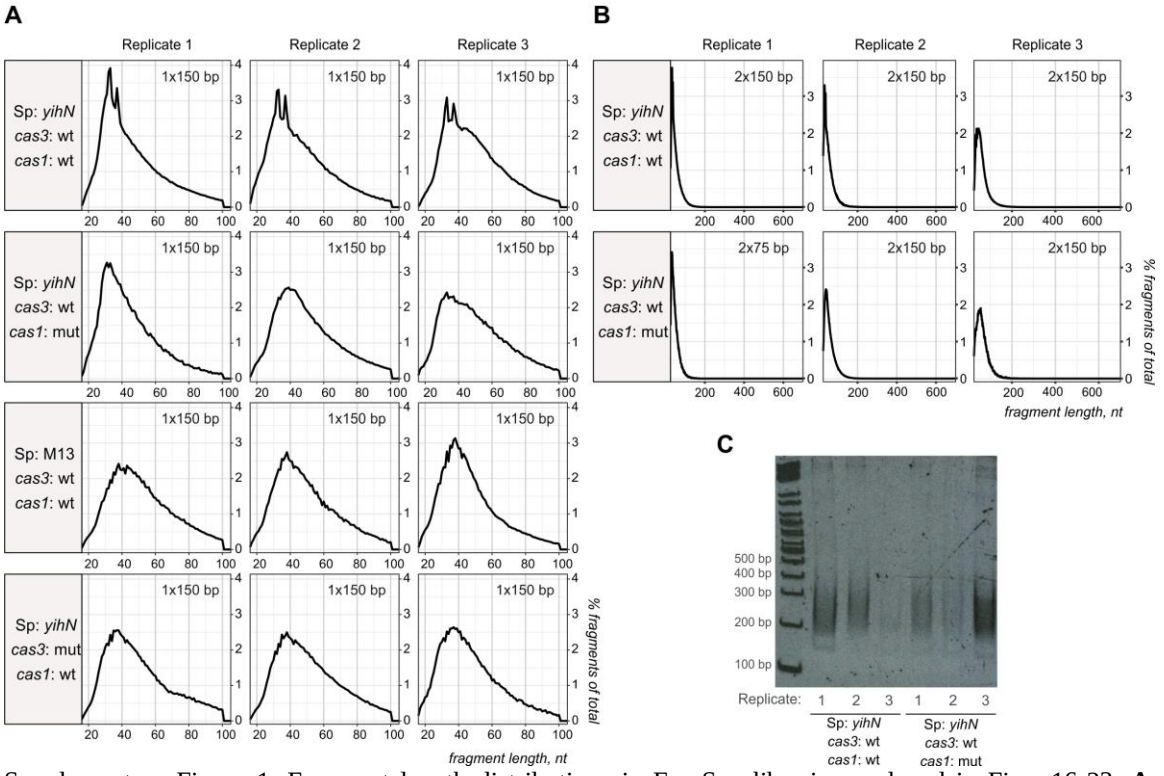
Table 3. Oligonucleotides used for prespacer efficiency assay

#	Transforming oligo names	Transforming oligo sequences
1.	G_33	5'GCCCAATTTACTACTCGTTCTGGTGTTTCTCGT 3' 3'CGGGTTAAATGATGAGCAAGACCACAAAGAGCA 5'
	C_33	
2.	AAG_35	5' AAGCCCAATTTACTACTCGTTCTGGTGTTTCTCGT 3' 3' TTCGGGTTAAATGATGAGCAAGACCACAAAGAGCA 5'
	TTC_35	
3.	G_33	5' GCCCAATTTACTACTCGTTCTGGTGTTTCTCGT 3' 3' AGTTCGGGTTAAATGATGAGCAAGACCACAAAGAGCA 5'
	AGTTC_37	
4.	AG_34	5' AGCCCAATTTACTACTCGTTCTGGTGTTTCTCGT 3' 3' AGTTCGGGTTAAATGATGAGCAAGACCACAAAGAGCA 5'
	AGTTC_37	
5.	G_32	5' GCCCAATTTACTACTCGTTCTGGTGTTTCTCG 3' 3' AGTTCGGGTTAAATGATGAGCAAGACCACAAAGAGCA 5'
	AGTTC_37	
6.	AG_33	5' AGCCCAATTTACTACTCGTTCTGGTGTTTCTCG 3' 3' AGTTCGGGTTAAATGATGAGCAAGACCACAAAGAGCA 5'
	AGTTC_37	

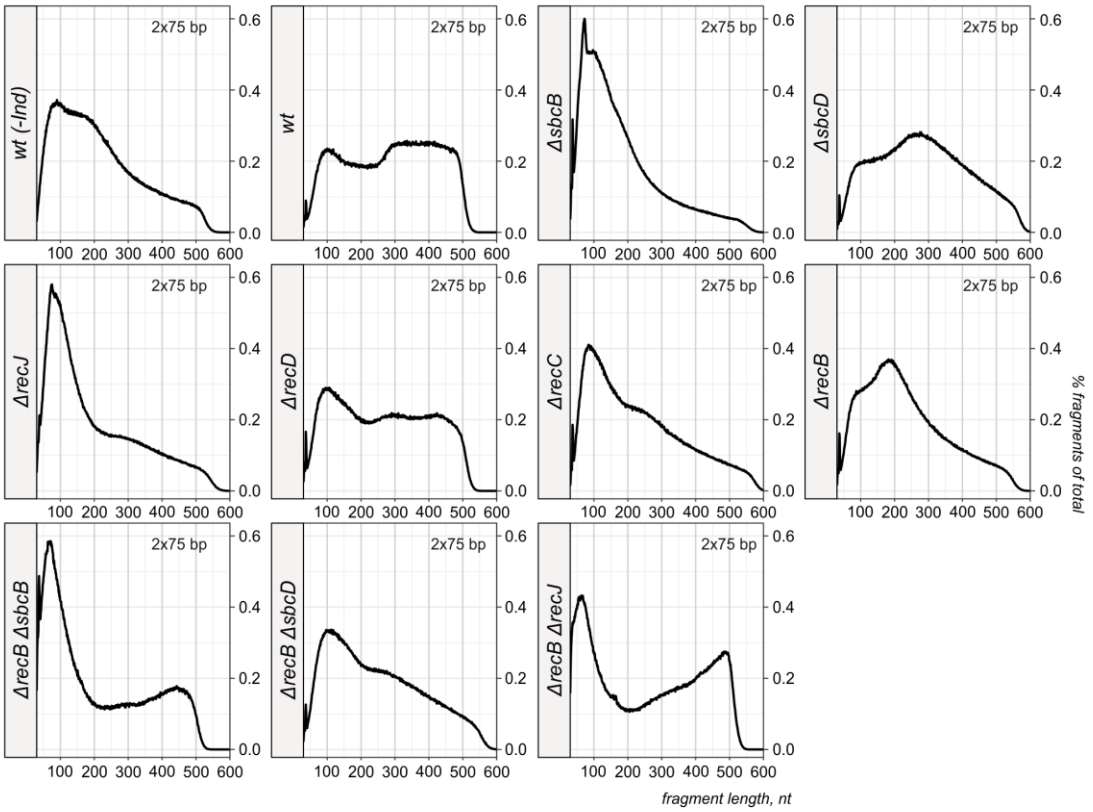
*****Nucleotides corresponding to the PAM are written in red

Table 4. List of adapters used for FragSeq of fragments purified from KD263, KD753, KD403, KD518, and KD675 strains

Name	Sequence (5' to 3')	Description
i112	GTTCAGAGTTCTACAGTCCGACGATCCTGANNNNNNNNNN	5′ adapter with CTGA barcode and 11N extension used in KD263 short DNA fragments library preparation (barcode is underlined)
i113	GTTCAGAGTTCTACAGTCCGACGATCGACTNNNNNNNNNN	5′ adapter with GACT barcode and 11N extension used in KD753 short DNA fragments library preparation (barcode is underlined)
i114	GTTCAGAGTTCTACAGTCCGACGATCAGTCNNNNNNNNNN	5′ adapter with AGTC barcode and 11N extension used in KD403 short DNA fragments library preparation (barcode is underlined)
i115	GTTCAGAGTTCTACAGTCCGACGATCTCAGNNNNNNNNNN	5′ adapter with TCAG barcode and 11N extension used in KD518 and KD675 short DNA fragments library preparation (barcode sequence is underlined)
i116	Phos/NNNNNNNNNTGGAATTCTCGGGTGCCAAGG/ddC/	3′ adapter with 9N random sequence used in short DNA fragments library preparation



Supplementary Figure 1. Fragment length distributions in FragSeq libraries analyzed in Figs. 16-23. **A.** Fragment length distributions of fragments sequenced using a single-end 1x150 bp protocol. **B.** Fragment length distributions in six libraries shown in panel **A** re-sequenced in pair-end (2x75 bp or 2x150 bp) modes. The Y-axes in both panels show the percentage of fragments with a certain length among all fragments mapped to the genome. **C.** A representative example of sequencing libraries (from the top two rows in panels **A** and **B**) run on a 10% TBE polyacrylamide gel. The size of amplified adapter dimers without an insert is 142 bp.



Supplementary Figure 2. Fragment length distributions in FragSeq libraries analyzed in Figs. 28-31 and 34-35. The libraries were sequenced in a pair-end (2x75 bp) mode. The Y-axes show the percentage of fragments with a certain length among all fragments mapped to the genome.