**VILNIUS UNIVERSITY**

**TAUTVYDAS KARVELIS**

# TYPE II CRISPR-Cas SYSTEMS: FROM BASIC STUDIES TOWARDS GENOME EDITING

Doctoral dissertation

Physical science, biochemistry (04 P)

**Vilnius, 2016**

The work presented in this doctoral dissertation has been carried out at the Institute of Biotechnology, Vilnius University during 2011-2015.

**Supervisor**

Prof. dr. **Virginijus Šikšnys** (Vilnius University, physical sciences, biochemistry - 04 P).

**VILNIAUS UNIVERSITETAS**

**TAUTVYDAS KARVELIS**

# II TIPO CRISPR-Cas SISTEMOS: NUO FUNDAMENTINIŲ TYRIMŲ LINK GENOMŲ REDAGAVIMO

Daktaro disertacija

Fiziniai mokslai, biochemija (04 P)

**Vilnius, 2016**

Disertacija rengta 2011-2015 m. Vilniaus universiteto Biotechnologijos institute.

**Mokslinis vadovas**

Prof. dr. **Virginijus Šikšnys** (Vilniaus universitetas, fiziniai mokslai, biochemija - 04 P).

# CONTENTS

# LIST OF ABBREVIATIONS

| | |
|---|---|
| Ap | ampicillin |
| Blat | *Brevibacillus laterosporus* |
| BLESS | breaks labeling, enrichment on streptavidin and next-generation sequencing |
| bp | base pair |
| BSA | bovine serum albumin |
| Cas | CRISPR associated |
| Cas9n | Cas9 nickase |
| Cascade | CRISPR-associated complex for antiviral defense |
| Cm | chloramphenicol |
| CRISPR | clustered regularly interspaced short palindromic repeats |
| crRNA | CRISPR ribonucleic acid |
| dCas9 | dead (catalytically inactive) Cas9 |
| Digenome-seq | digested genome sequencing |
| ds | double-stranded |
| DTT | 1,4-dithiothreitol |
| EDTA | ethylenediaminetetraacetic acid |
| EMSA | electrophoretic mobility shift assay |
| GMO | genetically modified organism |
| GUIDE-seq | genome wide unbiased identification of DSBs enabled by sequencing |
| HDR | homology directed repair |
| HTGTS | high-throughput genome-wide translocation sequencing |
| IDLV | integrase-defective lentiviral vector |
| IME | immature maize embryos |
| INDEL | insertion and/or deletion |
| IPTG | isopropyl β-D-1-thiogalactopyranoside |
| NHEJ | non-homologous end joining |
| nt | nucleotide |
| PAA | polyacrylamide |
| PAGE | polyacrylamide gel electrophoresis |
| PAM | protospacer adjacent motif |
| PCR | polymerase chain reaction |
| PDB | protein data bank |
| PNK | T4 polynucleotide kinase |
| pre-crRNA | precursor crRNA |
| RNP | ribonucleoprotein |

| | |
|---|---|
| RVD | repeat variable di-residue |
| SD | standart deviation |
| SDS | sodium dodecyl sulfate |
| SEM | standard error of the mean |
| Spy | *Streptococcus pyogenes* |
| ss | single-stranded |
| Sth1 | *Streptococcus thermophilus* CRISPR1 |
| Sth3 | *Streptococcus thermophilus* CRISPR3 |
| TALE | transcription activator-like effector |
| TALEN | transcription activator-like effector nuclease |
| TEMED | N,N,N′,N′-tetramethylethylenediamine |
| tracrRNA | trans-activating crRNA |
| Tris | 2-amino-2-hydroxymethyl-1,3-propanediol |
| WGS | whole genome sequencing |
| WT | wild type |
| ZF | zinc finger |
| ZFN | zinc finger nuclease |

# INTRODUCTION

Nowadays, DNA sequencing technologies enable rapid sequencing of complete genomes within a few days. However, our abilities to manipulate DNA sequences *in vivo* in order to understand or change biological traits or functions are lagging behind due to lack of available molecular tools for genome manipulation. Naturally occurring or engineered homing endonucleases, ZFN or TALEN meganucleases that introduce double stranded breaks (DSBs) in DNA and trigger repair mechanisms that may result in gene knock-outs/knock-ins appeared as first promising genome editing tools. However, development of these proteins for precise genome manipulation requires extensive protein engineering for reprogramming of the amino acid code for a specific DNA sequence and often results in tools with low reproducibility, efficiency and specificity (Carroll, 2014; Porteus, 2016).

The recent development of the Cas9 technology revolutionized the genome editing field. As with the most revolutionary discoveries, Cas9 nuclease genome editing technology started by trying to answer fundamental biological questions that were not directly related to genome manipulation. In 2007, CRISPR (clustered regularly interspaced short palindromic repeats)-Cas (CRISPR associated) systems emerged as a novel bacterial defense system against viruses and plasmids (Barrangou et al., 2007). Soon it was revealed that CRISPR-Cas systems use small RNA molecules called CRISPR RNA (crRNA) to guide ribonucleoprotein complexes to foreign nucleic acids and trigger their degradation (Brouns et al., 2008). RNA-based recognition of foreign nucleic acids through the Watson-Crick type base pairing, offered unprecedented opportunity for adaptation of such systems for genome editing. Type II CRISPR-Cas systems emerged as most promising candidates, because of the relatively simple CRISPR-Cas locus composition (Makarova et al., 2015). It turned out that Type II systems require a single Cas9 protein bound to a dual crRNA:tracrRNA (trans-activating crRNA) molecules for destruction of foreign DNA (Deltcheva et al., 2011; Sapranauskas et al., 2011). The crRNA is

used as guide for DNA targeting while tracrRNA is involved in bacterial RNase III dependent pre-crRNA maturation.

DNA target recognized by the Cas9 complex is composite and is comprised of DNA sequence (protospacer) complementary to the spacer component of crRNA and short adjacent sequence termed PAM (protospacer adjacent motif) that is absolutely required for efficient DSB formation (Garneau et al., 2010; Sapranauskas et al., 2011). PAM sequence is recognized by the Cas9 protein and is uniquely associated with each Cas9 protein.

These fundamental studies of Type II CRISPR-Cas systems in bacteria set the stage for the genome editing experiments in eukaryotes. This coincides not only with the start and focus of the project (with the continuously adjusting aims) described in this PhD thesis but also with the dawn of Cas9-initiated CRISPR craze that resulted in the one of the biggest discoveries in molecular biology.

**The aims of this study were:**

1. To identify molecular components required for the activity of Cas9 complex from *Streptococcus thermophilus* DGCC7710 CRISPR3-Cas system;
2. To probe whether the Cas9 complex from *S. thermophilus* DGCC7710 CRISPR3-Cas system can be used for targeted DNA cleavage *in vitro* and *in vivo*;
3. To establish the mechanism for Cas9 DNA binding and target sequence recognition;
4. To develop an experimental assay for rapid PAM characterization of newly identified Cas9 variants;
5. To validate the PAM identification assay for characterization of a novel Cas9 protein identified in *Brevibacillus laterosporus* SSP360D4.

**Scientific novelty.**

In this work, for the first time we identified all molecular components required for an assembly of the functional Cas9 complex of *S. thermophilus* CRISPR3-Cas system and demonstrated targeted DNA cleavage by the reconstituted Cas9 complex *in vitro* and *in vivo*. More specifically, we showed for the first time that Cas9 complex can be used as a tool for DNA cloning and targeted genome editing by chemical transfection of the *in vitro* assembled Cas9 ribonucleoprotein complex into the cells. Aiming to understand the molecular mechanism governing DNA recognition and cleavage by the Cas9 complex, we analyzed mechanism of R-loop formation and provided first direct evidence for directional R-loop formation, starting from PAM recognition and expanding toward the distal protospacer end. Realizing the importance of PAM sequence for Cas9 function, we developed an assay for rapid PAM identification for newly identified Cas9 proteins and characterized Type II-C Cas9 protein of *Brevibacillus laterosporus,* expanding Cas9 toolbox for genome editing applications.

**Practical value.**

The Cas9-based CRISPR-Cas system recently emerged as a versatile molecular tool for genome manipulation. In contrast to previously used platforms (homing endonucleases, ZFNs and TALENs) where DNA recognition is governed by protein-DNA interactions, Cas9 RNA-guided nucleases offer more robust and easy reprogrammable tool that recognizes the DNA target through the Watson-Crick base pairing of crRNA and DNA. We show here that basic studies of Cas9 protein provided in this thesis can be translated to the genome editing applications. Furthermore, we provide a universal *in vitro* method for the simultaneous examination of guide RNA and protospacer adjacent motif (PAM) requirements, allowing expansion of Cas9 toolbox for genome targeting by utilizing orthogonal Cas9 proteins. These results contribute to the recent genome targeting technological breakthrough based on Cas9 technology.

**The major findings presented for defense in this thesis**:

1. crRNA and tracrRNA are required for Cas9 complex from *S. thermophilus* CRISPR3-Cas system activity;
2. *In vitro* assembled Cas9 complex is capable to cleave target DNA *in vitro* and *in vivo*;
3. PAM sequence is used as a binding and R-loop priming site by Cas9 complex followed by unidirectional crRNA hybridization to target DNA;
4. Rapid PAM characterization assay was developed;
5. Cas9 protein from *B. laterosporus* recognizes a novel NNNNCNDD PAM sequence and can be used as a tool for genome editing.

# LITERATURE OVERVIEW

## 1.1. Development of genome editing technologies

A recent technological breakthrough in the field of DNA sequencing allowed the determination of complete genome sequences for a wide variety of organisms. This was the first necessary step for search of thorough understanding of complex genetic functions that define living organisms. To be able to do extensive characterization and analysis of biochemical processes taking place in living cells it is required to link genetic sequence with encoded information and function. This depth of understanding would be greatly facilitated by technologies that allow to easily manipulate genomes in a truly precise fashion.

Early attempts to introduce exogenous DNA into organisms were carried out in the 1970s. Methods developed to manipulate DNA molecules *in vitro* and *in vivo* using restriction enzymes and basic cloning procedures led to the first genetically modified organism (GMO) (Cohen et al., 1973). Construction and transformation of the plasmid encoding kanamycin resistance gene to *E. coli* cells, allowed bacteria to survive in the presence of antibiotic. Soon thereafter, ribosomal RNA coding gene from *Xenopus laevis* was transferred to *E. coli* in the form of an engineered plasmid (Morrow et al., 1974). This was the first GMO expressing a gene of another organism belonging to a different kingdom. The first genetically modified animal was created by injecting Simian virus SV40 viral DNA into mouse blastocysts. Analysis of DNA extracted from animals derived from these cells showed that SV40 specific DNA was present in some organs in 40% of the adult survivors (Jaenisch and Mintz, 1974). However, in these trials no exogenous DNA was passed to offsprings showing that no DNA integration occurred in germ line cells. During the following few years, this issue was resolved by introducing exogenous genetic material using Moloney murine leukemia viruses (Jaenisch, 1976, 1977).

All the above described cases were the first attempts to introduce DNA into cells in a basically uncontrollable fashion. The first attempt of precise genome modification was in 1985 using human cells (Smithies et al., 1985). Authors were able to replace *HBB* gene in somatic human cells in the targeted fashion, using endogenously existing homology directed repair (HDR) mechanism. However this approach was very inefficient. Only small fraction of cells ($10^{-6}$–$10^{-7}$) contained desired mutation (Capecchi, 1989) complicating wider adaptation of this method.

The breakthrough to exogenous DNA transfer became realization, that targeted double strand breaks (DSBs) greatly facilitated DNA repair mechanisms in yeast (Rudin et al., 1989) and mammalian cells (Rouet et al., 1994). Briefly, cells contain two DSB repair pathways: non-homologous end joining (NHEJ) and HDR (Wyman and Kanaar, 2006) (Figure 1). NHEJ simply joins broken DNA strands, often creating small insertions and deletions (INDELs). HDR, on the other hand, precisely repairs DSB, by replacing the broken region using homologous template. So induction of DSBs in cells triggers both targeted mutagenesis and gene repair if homologous template is present. Proof of principle was shown by an experiment, where defective reporter gene containing 18 bp long recognition sequence of homing endonuclease I-SceI from *Saccharomyces cerevisiae* was integrated into a mouse cell genome (Choulika et al., 1995; Rouet et al., 1994). After that, introduction of I-SceI into cells with a donor DNA plasmid containing intact gene resulted in cells with corrected gene due to I-SceI induced homologous recombination between donor plasmid and defective chromosomal locus. Even though this approach to edit genome using naturally occurring endonucleases with long DNA recognition sequences was a very powerful method, but the limitation to use such enzymes to cleave any desired locus restricted broader practical applications. Instantly the efforts were shifted to develop the tools allowing to introduce DSBs precisely at the natural sites in complex genomic DNA, so a new era of programmable nucleases development began. So far, there are developed four major scaffolds of programmable nucleases: homing

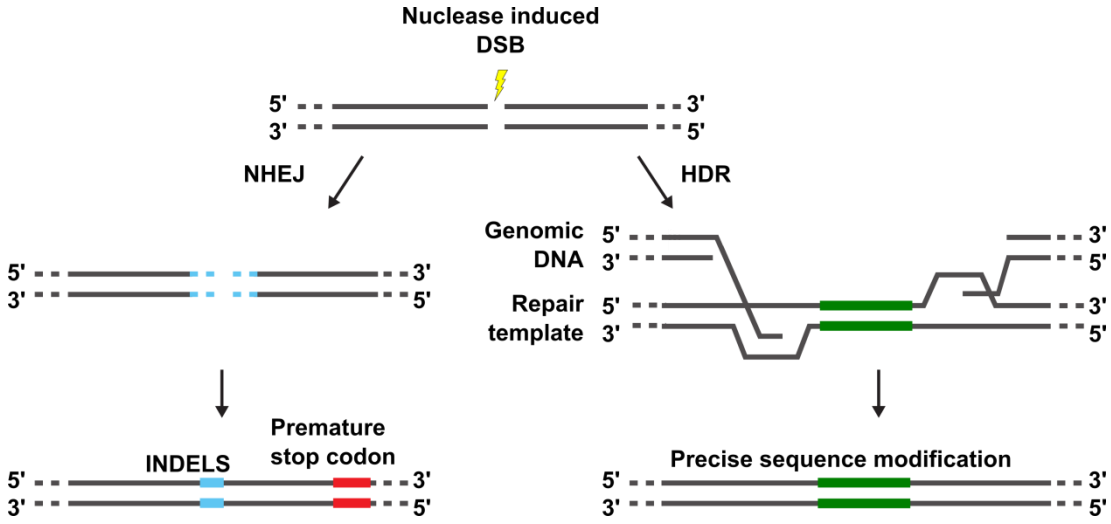


**Figure 1. A schematic representation of DNA double strand break (DSB) repair.** DSBs induced by nucleases can be repaired using two DSB repair pathways. In the error-prone NHEJ pathway, the ends of a DSB are processed by endogenous DNA repair machinery and rejoined, resulting in random INDEL mutations. INDELs introduced within the coding region of a gene can lead to frameshifts resulting in a premature stop codon. Alternatively, in the HDR pathway DSB can be repaired by replacing the broken region using homologous repair template. Therefore, precise point mutations or insertions from a single-stranded or double-stranded DNA donor template can be introduced.

endonucleases from bacterial mobile genetic elements (Smith et al., 2006), zinc finger (ZF) nucleases (ZFNs), derived from eukaryotic transcription factors (Bibikova et al., 2001, 2002, 2003; Kim et al., 1996), transcription activator-like effector (TALE) nucleases (TALENs), from *Xanthomonas* bacteria (Boch et al., 2009; Christian et al., 2010; Miller et al., 2011; Moscou and Bogdanove, 2009) and the most recent the CRISPR-Cas RNA-guided nucleases from bacterial adaptive immune system (Cong et al., 2013; Gasiunas et al., 2012; Jinek et al., 2012, 2013; Mali et al., 2013a; Zetsche et al., 2015a).

### 1.1.1. Homing endonucleases

Homing endonucleases (meganucleases) are sequence-specific endonucleases recognizing large (>12 bp) sequence sites. These enzymes are derived from a large class of mobile genetic elements (introns and inteins) that catalyze their own propagation using induced DSBs at a specific genomic locus in which the mobile element is absent, followed by duplication after HDR (Chevalier and Stoddard, 2001; Jurica and Stoddard, 1999).

Until now, five homing endonuclease families were characterized on the basis of conserved protein motifs (Stoddard, 2006, 2011). Most widespread and best characterized is LAGLIDADG family. The name of this family corresponds to a short amino acid sequence (motif) which is found in all of the proteins of this family. LAGLIDADG endonucleases exist both as homodimers (where the two identical protein subunits are each 160 to 200 residues in size) and as monomeric proteins where a tandem repeat of two LAGLIDADG domains is connected by a variable peptide linker (their individual domains are often only 100 to 120 residues in size). While homodimeric LAGLIDADG enzymes recognize palindromic or near palindromic sequences (*e.g.* I-CreI), monomeric proteins (*e.g.* I-SceI) can recognize fully asymmetric DNA target sites (Stoddard, 2011). LAGLIDADG endonucleases contain a conserved core structure, characterized by a αββαββα fold, which is unique for these proteins. Typically, two αββαββα folds are facing each other across a twofold symmetrical or pseudo-symmetrical axis and both contribute to the active center, thus leading to tight connection between target binding and cleavage activities (Figure 2A and B). The β-sheets are organized in a saddle shaped pair

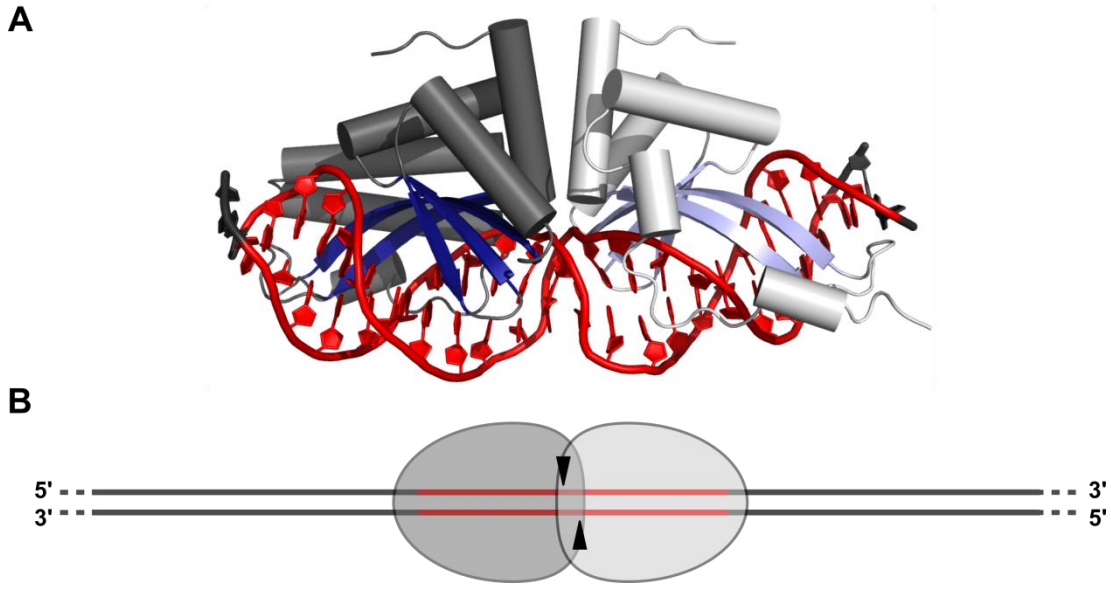


**Figure 2**. **Homing endonucleases from LAGLIDADG family** (**A**) Structure of I-CreI homodimer bound to target DNA (indicated in red) (PDB id: 1G9Y). Homing endonuclease domains are shown in grey with blue β-sheets showing DNA binding saddle. (**B**) Schematic representation of LAGLIDADG family meganuclease bound to target DNA. DNA cleavage position is indicated by black triangles

of hairpins, which are located above the DNA major groove, and are responsible for specific DNA recognition (Pâques and Duchateau, 2007).

During 1990s several hundreds of homing endonucleases had been identified, but still the range of recognizable sequences was way too limited to address a complexity of genomes. Speculations, that altering DNA recognition contacts could overcome this issue only shifted the problem to another level – protein engineering (Pâques and Duchateau, 2007). Decades of attempts to change the substrate specificity of restriction enzymes clearly demonstrated the complexity of the task (Lanio et al., 2000; Rimseliene et al., 2003; Samuelson and Xu, 2002).

For reengineering homing endonucleases, three main strategies were adapted (Stoddard, 2014): altering individual contacts in the protein (Gimble et al., 2003; Sussman et al., 2004), shuffling the domains from LAGLIDADG meganucleases family members (Baxter et al., 2012; Smith et al., 2006; Steuer et al., 2004) or combining both approaches (Gao et al., 2010; Grizot et al., 2010) followed by complex and laborious *in vitro* and *in vivo* experimental selection. However, the success of these studies has been moderate. The fact that DNA recognition mechanism cannot be reduced to simple modular code and that in parallel a more attractive technology based on ZF (described below) was being developed, failed to ensure popularity of these enzymes in genome editing field, although biotechnology companies like Cellectis Bioresearch, Pacific Biosciences and academic groups continue to develop this system (Porteus, 2016).

### 1.1.2. ZFNs

ZFN are engineered proteins, but they originated from the natural components: zinc finger (ZF) proteins and FokI endonuclease. In early 1990s investigators discovered, that Type II restriction endonuclease FokI contains two distinct and separable sequence-specific DNA binding and cleavage domains (Li et al., 1992). This observation led to a hypothesis that novel

cleavage specificities could be obtained by linking nonspecific FokI DNA cleavage domain to alternative DNA binding modules. Few years later, first active chimeric nuclease was engineered by fusing Drosophila Ultrabithorax homeodomain with cleavage domain from FokI nuclease (Kim and Chandrasegaran, 1994), demonstrating the possibility to design nucleases with altered specificities.

Meanwhile, the story of another component began in early 1980s. It was shown, that immature oocytes of *Xenopus laevis* store 5S RNA molecules in the form of 7S ribonucleoprotein (RNP) particles (Picard and Wegnez, 1979). Soon, it was identified, that protein in the complex is transcription factor IIIA (TFIIIA) (Pelham and Brown, 1980). Biochemical studies of TFIIIA identified repeating 30 amino acids motif within the protein responsible for binding DNA, which was called zinc finger (ZF) motif, because it contained bound zinc (Zn) (Miller et al., 1985). This new fold offered a novel principle of DNA recognition that was distinct from helix-turn-helix motif found in the first DNA interacting proteins described. Most often, the latter bind to DNA as symmetric dimers, thus recognizing palindromic DNA sequences (Klug, 2010). In contrast, ZF proteins bind to variable length DNA in linear and polar fashion. The modular design of ZF proteins offers a vast number of combinatorial possibilities to bind specific DNA targets. Therefore, it is not surprising that these proteins are abundant in eukaryotic sequence specific transcription factors and represents the second most frequently encoded protein family in human genome (~3% of genes contain ZF motif) (Gaj et al., 2013; Klug, 2010).

Despite the extensive biochemical characterization of ZF proteins, the precise amino acids responsible for specific binding to DNA remained unknown. The breakthrough to understand how the ZF proteins recognize DNA targets was the solved crystal structure of the mouse Zif268 transcription factor bound to DNA (Pavletich and Pabo, 1991). Zif268 DNA binding domain consisting from three zinc fingers confirmed the modularity of recognition and the coordination of a single Zn atom by two histidine and two cysteine residues

($Cys_2$-$His_2$ motif) in each finger domain (Figure 3A and B). It was shown that the dominant contacts with three consecutive bases in the same DNA strand are formed with α helix, which is positioned in the DNA major groove binding to DNA by three amino acids at -1, 3 and 6 helical positions (Figure 3A). Two years later, another solved structure of ZF protein bound to DNA revealed additional minor contact made from amino acid at helical position 2 to the adjacent base of the other DNA strand (Fairall et al., 1993) (Figure 3A). The most important discoveries here were that each DNA base was recognized by single amino acid and every ZF functions as a mostly independent module that recognizes DNA triplet (Figure 3A and B). By rational design (Desjarlais and Berg, 1992) or advanced phage display library screening approaches (Choo and Klug, 1994) researchers started to engineer ZF proteins resulting in the first active ZF protein with altered substrate specificity in 1994 (Choo et al., 1994). Three finger peptide was engineered to bind to unique 9 bp region of the oncogenic p190 *bcr-abl* fusion gene, originated from translocation of 9 and 22 chromosomes tips, resulting fusion between parent *bcr* and *abl* genes (Choo et al., 1994). Mouse cells harboring *bcr-abl* oncogene transfected with vector expressing ZF peptide showed reduced expression of *bcr-abl*, due to transcriptional block imposed by specific binding of the ZF protein. In addition, authors fused the same ZF peptide with VP16 activation domain from Herpes simplex virus. Reporter gene expression from the plasmid, containing promoter with the same ZF recognition sequence, increased by 30 fold after the transfection of cells with vector encoding ZF-VP16 fusion protein (Choo et al., 1994), demonstrating that engineered ZF proteins can be used as a guide to deliver functionally active proteins to specific DNA sequences.

At this point in time, all necessary components for engineering a programmable nuclease were present: nonspecific DNA cleavage domain from FokI endonuclease and modular DNA binding ZF proteins. During the next few years, first artificial ZFN protein was described (Kim et al., 1996). By fusing a ZF

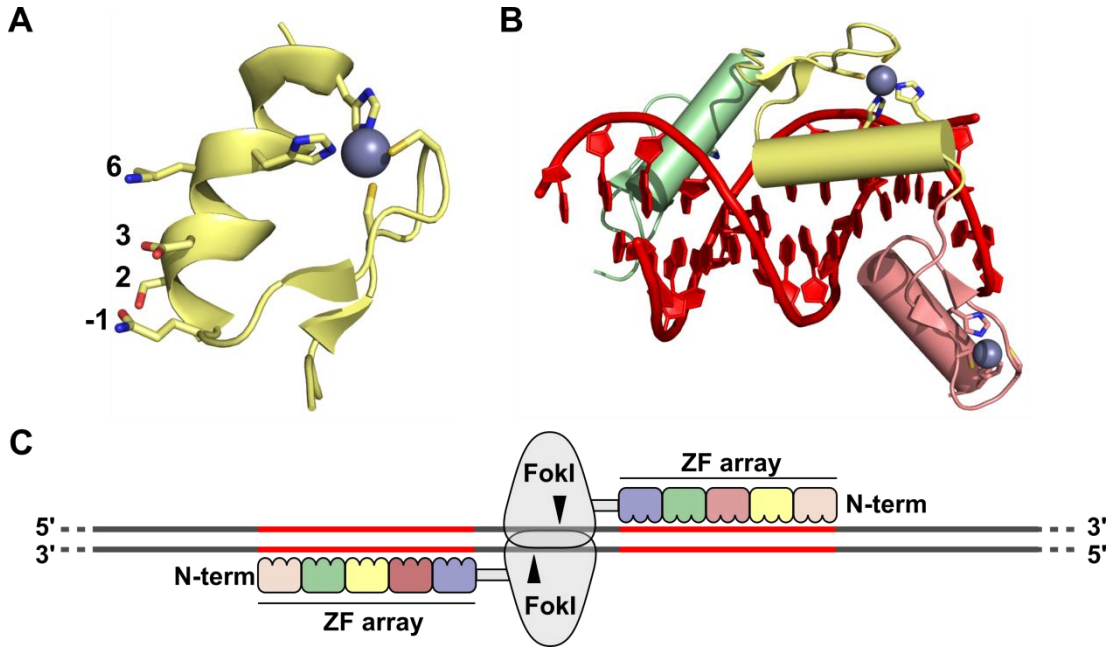


**Figure 3. Zinc finger (ZF) nucleases (ZFNs).** (**A**) Single finger motif highlighting individual residues responsible for contacting DNA bases. Positions −1, 3, and 6 are responsible for the dominant contacts with three consecutive bases. The residue in position 2 forms additional minor contact with DNA base adjacent to DNA triplet. Coordinated Zn atom by two histidine and two cysteine residues is visualized as a gray sphere. (**B**) Structure of three zinc fingers motifs bound to their target DNA in the major groove (PDB id: 1MEY). (**C**) Schematic representation of ZFNs bound to DNA target. ZFNs have multiple DNA binding motifs linked in series to create a ZF array. Each array binds a separate half-site sequence (indicated in red), and a dimeric FokI nuclease cleaves the DNA within what is referred to as the spacer region. DNA cleavage position is indicated by black triangles.

protein to FokI nuclease domain authors were able to demonstrate specific nuclease activity on the λ DNA *in vitro* and the race to apply them for genome editing applications had begun. Unfortunately, it took five years to reach this goal, because of an obstacle that was not known at the time. To efficiently generate DSB, FokI nuclease must dimerize (Bitinaite et al., 1998; Wah et al., 1998). Soon, this requirement was confirmed for FokI nuclease domains in ZFN proteins as well (Smith, 2000). It was demonstrated, that for efficient DSB generation by ZFNs, the substrate with two binding sites in close proximity (~6 bp) is required. Finally, taking into account the FokI nuclease domain dimerization, ZF number and spacer length between ZF binding sites requirements (Figure 3C), experiments with synthetic extrachromosomal DNA and designed ZFNs showed efficient cleavage and recombination of DNA *in vivo* in *Xenopus* oocytes (Bibikova et al., 2001). Soon, the first successful

experiments using ZFN pair designed for a natural genomic target took place. Both efficient targeted mutagenesis (Bibikova et al., 2002) and gene replacement (Bibikova et al., 2003) were demonstrated in *Drosophila* cells.

After showing, that ZFNs can be used for targeted genome editing, improving the efficiency and specificity became the next main goal for subsequent work. Even though it was well known, that ZF recognition of 3 nucleotides mostly lies in three main contacts made with residues at -1, 3 and 6 helical positions (Pavletich and Pabo, 1991), initial studies aiming to redesign ZF ignored additional contact to adjacent nucleotide made from amino acid at helical position 2 (Fairall et al., 1993) that was shown that it could significantly contribute to specificity (Isalan et al., 1997, 1998). By redesigning previously used phage library screening methods, scientists were able to take into account this interaction and create an archive of ZF proteins that selectively recognize a vast number of DNA sequences (Isalan et al., 2001). Another step in improving ZFN technology was increasing the length of recognizable DNA sequences, thus increasing the rarity of targets in genomic DNA and reducing possible off-targets. Simple addition of ZF modules to existing three ZF proteins had minor effect, because the periodicity of ZF modules linked with conventional linkers does not quite match the periodicity of DNA (Klug, 2010), leading only to the small increase in binding affinity. By using engineered longer and more flexible linkers between pre-existing two (Moore et al., 2001) or three finger peptides (Kim and Pabo, 1998), scientists were able to get an array of up to six ZF units with drastically increased binding affinity and specificity. During the next few years the method of choice became engineering ZF proteins by linking together pre-existing two finger peptides (Klug, 2010), because this design proved to be more sensitive to the DNA sequence mismatches (Moore et al., 2001). Possibility to add ZF units in ZFN proteins drastically increased specificity and reduced off-target cleavage, but in some cases this is not sufficient (Shimizu et al., 2011). ZFNs toxicity that is related to promiscuous binding and cleavage in off-target sites could be linked with inadequate specificity of one ZF protein in ZFNs pair, resulting in homodimerization and

undesirable cleavage at off-target sites. Requirement of FokI nuclease domains in ZFNs pair to form a dimer, suggested a perfect possibility to overcome this issue and further refine the system. Redesigning FokI dimer interface to prevent homodimer formation, but allowing heterodimerization solved the problem, but lowered on-target cleavage efficiency (Miller et al., 2007; Szczepek et al., 2007). Few years later, by improving obligate heterodimer formation cleavage efficiency was restored (Doyon et al., 2011).

More than twenty years of ongoing ZFNs research and development made ZFNs a very attractive tool for genome engineering. Until now, more than 25 species have been genetically modified using ZFNs (Carroll, 2014). The most notable achievement confirming the maturity of technology is ongoing Phase I/II clinical trials on ZFN driven *CCR5* gene modification in T-cells of patients that are HIV-infected initiated by Sangamo Biosciences (Porteus, 2016).

It might look like there is a smooth path from DNA target selection to genetically modified cell using ZFNs. In reality, considerable proportion of designed ZFN pairs fail, due to complex protein-protein and protein-DNA interaction problems (Ramirez et al., 2008). Fortunately, for those who have only a few DNA targets and an adequate budget, designed ZFNs can be obtained from Sangamo Biosciences archive marketed by Sigma-Aldrich that have been extensively tested for activity and specificity. For the rest looking for an affordable tool, the hope was restored, when alternative technology based on transcription activator-like effector (TALE) nucleases (TALENs) emerged.

### 1.1.3. TALENs

Naturally, transcription activator-like effectors (TALEs) are produced by plant pathogens from *Xanthomonas* spp. and was first described as early as 1989 (Bonas et al., 1989). During infection, these proteins enter to the plant cells using Type III secretion pathway. Inside the cells, TALEs imported to the

nucleus bind to specific DNA targets and activate gene expression resulting in increased plant susceptibility to colonization by pathogen (reviewed in Boch and Bonas, 2010).

The most interesting feature of TALE proteins are its modular DNA binding motifs, which consist of as many as 30 repeats of 33-35 amino acids (Figure 4A and B). Each repeat is highly conservative except for hypervariable adjacent amino acids at position 12 and 13. These repeat variable di-residues (RVDs) are responsible for specific single DNA base recognition (Boch et al., 2009; Moscou and Bogdanove, 2009). Later, crystal structures of TALEs bound to DNA revealed, that only 13$^{th}$ residue makes base specific contact from major DNA groove, while 12$^{th}$ residue stabilizes the RVD containing loop, between two helices (Deng et al., 2012; Mak et al., 2012) (Figure 4A and B). Thus, in contrast to ZF proteins, where each finger dominantly recognizes three DNA bases, TALE repeats interact with DNA in one to one fashion (one TALE repeat recognizes one base pair).

After identification and characterization of TALE DNA recognition code, the analogy to ZF DNA recognition became obvious, and it was a matter of time when TALE DNA binding arrays would be fused with FokI nuclease. Almost twenty years of pioneering work developing ZFNs helped tremendously and it took only more than a year before TALE fusion proteins with FokI were produced showing engineered TALE nucleases (TALENs) pair ability to induce DSB *in vivo* (Christian et al., 2010; Li et al., 2011; Miller et al., 2011). However, in contrast to ZF proteins that can bind DNA without any additional motifs, assembled TALE modules require additional protein sequences on both sides of the TALE array (Christian et al., 2010; Miller et al., 2011) (Figure 4C). Upstream of the TALE binding domain there is a requirement for protein motifs that form pseudo repeats involved in making contacts with T base (Mak et al., 2012). C-terminal end of the TALE array contains half repeat followed by additional residues from natural TALE protein. These requirements determine DNA sequences that can be efficiently targeted with TALENs. Therefore, to allow dimerization of FokI nuclease, two

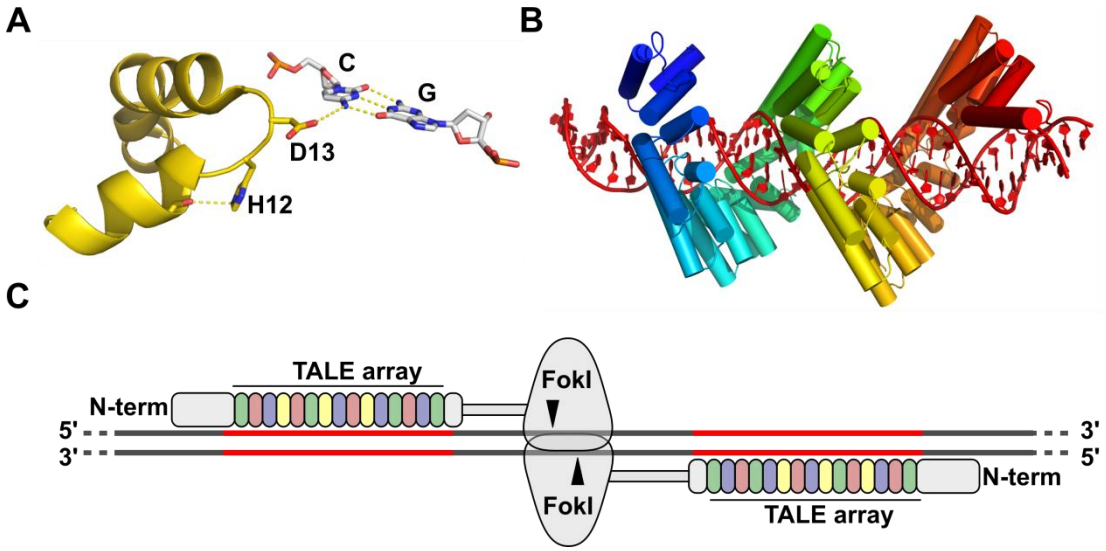


**Figure 4. Transcription activator-like effector (TALE) nucleases (TALENs).** (**A**) Single TALE motif highlighting repeat variable di-residues (RVDs) responsible for specific single DNA base recognition. 13$^{th}$ residue makes base specific contact from major DNA groove, while 12$^{th}$ residue stabilizes the RVD containing loop. (**B**) The crystal structure of TAL effector PthXo1 bound to its DNA target (PDB id: 3UGM). (**C**) Schematic representation of TALENs bound to DNA target. TALE array with additional protein sequences on both sides forms DNA binding module that is linked with FokI nuclease. Each TALEN binds separate half-site sequence (indicated in red), and a dimeric FokI nuclease cleaves target DNA (cleavage position is indicated by black triangles).

opposite orientation DNA binding sites separated by 12-20 bp are required. Longer 12-20 bp spacer for TALENs pair *versus* 5-7 bp for ZFNs is the result of additional protein sequences between TALE array at C-terminal end and FokI nuclease domain. Furthermore, each DNA target should start with T nucleotide, followed by 15-21 bp sequences to promote efficient binding by TALE DNA recognition array (Christian et al., 2010; Miller et al., 2011).

Although, single base discrimination without crosstalk between adjacent protein repeats offered greater ease and flexibility to design TALE arrays for novel DNA targets than with ZF proteins, but highly repetitive nature of these proteins raised new technical DNA assembly challenges. To overcome this issue, different strategies were adapted that can be grouped into four categories: standard restriction and ligation based cloning, solid phase assembly, "Golden Gate" cloning and commercial DNA synthesis (reviewed in Joung and Sander, 2013). For those, not interested in designing TALENs, companies like Cellectis Bioresearch, Thermo Fisher Scientific and

Transposagen Biopharmaceuticals offer custom TALENs construction. However, the complexity of protein engineering leads to the fact that on average only one third of engineered TALEN pairs shows reasonable activity at their intended target sites (Porteus, 2016).

Experience in developing ZFNs for genome editing greatly facilitated TALENs evolution and soon this technology became the method of choice. In all organisms tested, TALENs proved to be similar or more efficient than ZFNs (Carroll, 2014; Segal and Meckler, 2013). Ongoing preparations to enter Phase I clinical trials for immunotherapy using UCART19 T-cells (developed by Cellectis Bioresearch using TALENs) to treat $CD19^{+}$ B-cell leukemia shows the potential of this technology.

One major problem of all the aforementioned genome editing technologies is their DNA recognition that depends on protein sequence. Because of this, event after thorough protein design and synthesis it is necessary to extensively validate the activity *in vitro* and *in vivo* conditions, because it may vary in unpredictable ways. This leads to the construction of multiple sets of proteins, for the same locus to be targeted, resulting in genome editing experiments that are very complex, costly and time consuming. In 2012, no one even imagined that these problems soon will become history, after a new player called Cas9 from CRISPR-Cas bacterial adaptive immune system emerged.

### 1.1.4. CRISPR-Cas RNA-guided nucleases

The story of adaptable immunity system in prokaryotes composed of clustered regularly interspaced short palindromic repeats (CRISPR) and CRISPR associated (*cas*) genes (Figure 5), began in 1987, when first observation of CRISPR locus in *E. coli* genome were described (Ishino et al., 1987) and later in other organisms as well (Jansen et al., 2002; Mojica et al., 1993, 1995, 2000), but the function of these systems remained unknown.

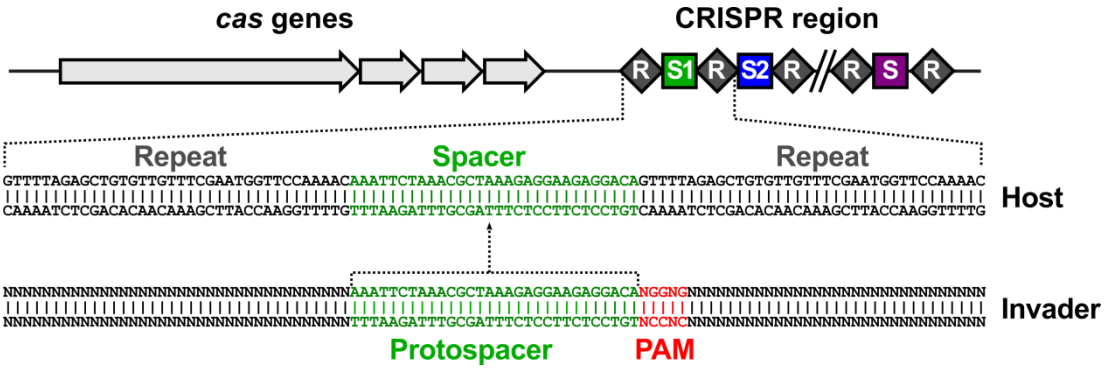


**Figure 5. Schematic representation of CRISPR-Cas system.** Usually, cluster of *cas* genes (represented by arrows) is located in the vicinity of CRISPR region, consisting of repeat sequences (represented by diamonds) interspaced by spacers (represented by boxes). Spacer sequences are originated from invading nucleic acids, where they are termed protospacers. Protospacer adjacent motif (PAM) is located in the vicinity of protospacer.

In 2005, after observation that non-repetitive sequences between repeats, termed spacers, match DNA sequences of phages and conjugative plasmids (Figure 5), the first hypotheses that these systems could be involved in RNA-guided adaptive immunity emerged (Bolotin et al., 2005; Makarova et al., 2006; Mojica et al., 2005; Pourcel et al., 2005). It took two years to experimentally demonstrate, that phage-sensitive *Streptococcus thermophilus* strains acquire immunity against phages when new phage derived spacers in pre-existing CRISPR locus were inserted (Barrangou et al., 2007). The central role of CRISPR RNA (crRNA) in this RNA-guided defense system was revealed shortly thereafter (Brouns et al., 2008).

Basically, the mechanism of action of CRISPR-Cas system can be divided into three stages: adaptation (spacer acquisition), crRNA biogenesis and interference (Figure 6). During adaptation, new spacers are selected from foreign nucleic acid and integrated in the CRISPR locus. Precursor CRISPR RNA (pre-crRNA) is then transcribed and further processed into smaller crRNAs, each containing a single or a part of the spacer adjacent to a repeat derived sequence or flanked by them. Matured crRNAs form ribonucleoprotein complexes with single or multiple Cas proteins, that recognize foreign nucleic acid by base pairing with crRNAs and triggers its degradation by Cas

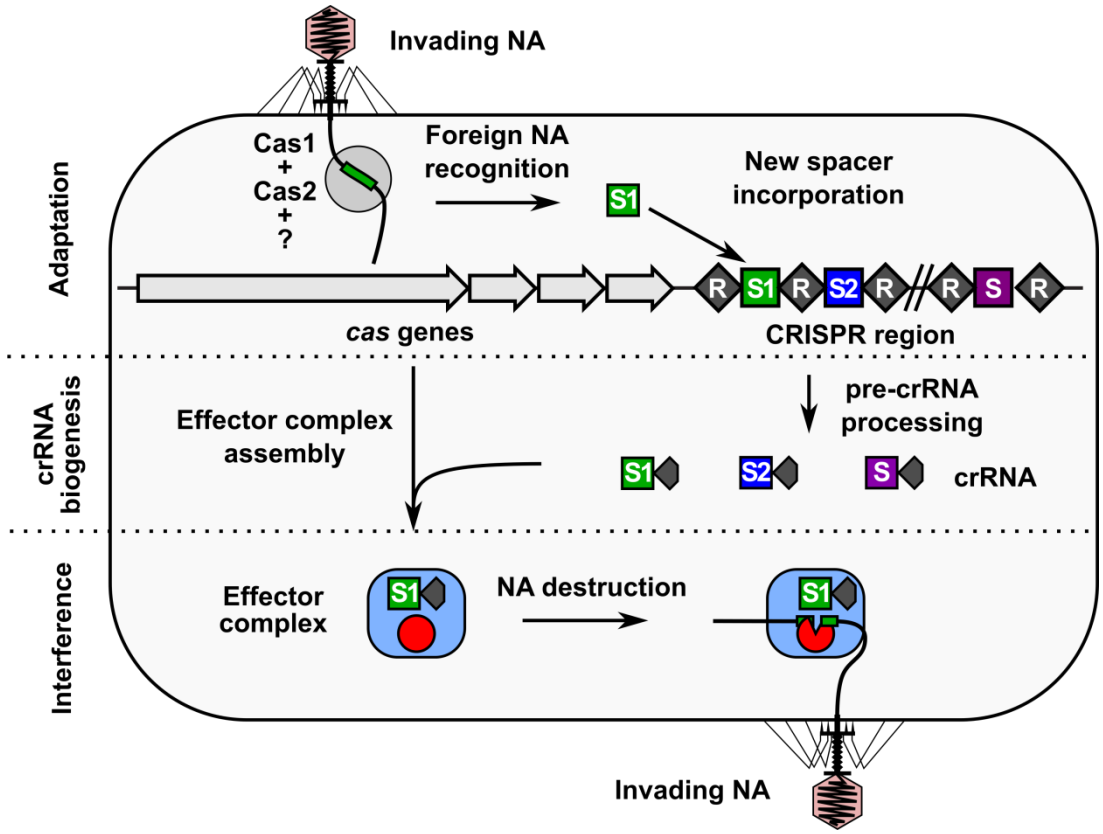


**Figure 6. Schematic representation of CRISPR-Cas mechanism.** The CRISPR-Cas system mechanism can be divided into three stages: adaptation, crRNA biogenesis and interference. During adaptation, Cas proteins recognize invasive nucleic acid (NA) and integrate short pieces of foreign DNA into the CRISPR region as new spacers. In the crRNA biogenesis stage, the CRISPR repeat-spacer array is transcribed into a long primary RNA transcript (pre-crRNA) that is further processed into a set of small crRNAs, containing a conserved repeat fragment and a variable spacer sequence (guide) complementary to the invading nucleic acid. crRNAs further combine with Cas proteins into an effector complex. In the interference stage, the effector complex recognizes the target sequence in the invasive nucleic acid by base pairing and induces sequence specific cleavage, thereby preventing proliferation and propagation of foreign genetic elements. Adapted from Gasiunas et al., 2013.

nucleases, resulting in aborted proliferation of foreign genetic elements (van der Oost et al., 2014; Wright et al., 2016). Extensive identification and characterization of CRISPR-Cas systems using bioinformatic and experimental approaches revealed the remarkable diversity of these systems. The most recent classification defines two classes, six types and 19 subtypes (Makarova et al., 2015; Shmakov et al., 2015). Despite this diversity, Cas proteins might be grouped into two main functional modules: adaptation module, which is responsible for spacer acquisition, and effector (interference) module, which, guided by crRNA recognizes and degrades foreign nucleic acids (Makarova et

al., 2015; Shmakov et al., 2015). The adaptation module consists of relatively conserved Cas1 and Cas2 proteins and share many common mechanistic features across different CRISPR-Cas systems. The effector module, on the other hand, shows extreme diversity. Thus, the recently introduced grouping of CRISPR-Cas systems into two classes are based on the effector modules. Class 1 group comprises CRISPR-Cas types with effector modules consisting of large multi-protein complexes: Cascade (CRISPR-associated complex for antiviral defense) accompanied with Cas3 protein to degrade DNA in Type I, Csm or Cmr in Type III, and putative interference complex from Type IV systems (Brouns et al., 2008; Hale et al., 2008, 2009; Kazlauskiene et al., 2016; Makarova et al., 2015; Marraffini and Sontheimer, 2008; Mulepati and Bailey, 2013; Sinkunas et al., 2011, 2013; Tamulaitis et al., 2014; Westra et al., 2012). In contrast, effector modules from Class 2 group consist of single Cas proteins: Cas9 in Type II, recently discovered Cpf1, C2c1 or C2c3 in Type V, and C2c2 in Type VI (Abudayyeh et al., 2016; Gasiunas et al., 2012; Jinek et al., 2012; Sapranauskas et al., 2011; Shmakov et al., 2015; Zetsche et al., 2015a). So, Class 2 CRISPR-Cas systems contain the simplest effector modules that have the potential to be easily designed for targeted DNA destruction and the development of bacterial Type II and V CRISPR-Cas RNA-guided nucleases to genome editing tools perfectly illustrates that.

After the initial experiments, demonstrating CRISPR-Cas immunity in action (Barrangou et al., 2007), the focus was shifted to understand how these systems work. Soon, by studying Type II CRISPR-Cas mediated plasmid interference in *S. thermophilus* scientists were able to capture linear DNA plasmid cleavage products. After sequencing and identification of blunt end cut 3 nt upstream of the DNA target 3'-end it was confirmed that Type II CRISPR-Cas system guided by RNA creates DSBs in target DNA (Garneau et al., 2010). Another unexpected key discovery came from a study focused on profiling *S. pyogenes* CRISPR-Cas regions expression using differential RNA sequencing. In addition to ~42 nt length crRNAs, consisting of 20 nt spacer derived 5' guide sequence and 22 nt repeat derived 3' sequence, additional

species of RNA molecules adjacent to the CRISPR-Cas locus were identified. These RNAs, named tracrRNAs (trans-activating crRNAs), contained sequence stretch with almost perfect complementarity to CRISPR repeats, suggesting base pairing between two RNA species. Experiments that followed confirmed formation of duplexes between tracrRNAs and precursor crRNAs (pre-crRNAs) that were further processed by endogenous RNase III, and after an additional crRNA's 5'-end trimming, resulted in matured forms of ~42 nt length crRNAs and ~75 nt length tracrRNAs (Figure 7). In agreement, genetic deletion experiments confirmed not only Cas9 and crRNA, but additionally tracrRNA and RNase III involvement in Type II CRISPR-Cas system DNA interference function *in vivo* (Deltcheva et al., 2011). Few months later another study came out regarding Type II systems. Scientists transferred entire *S. thermophilus* CRISPR-Cas locus with flanking sequences to distant *E. coli* bacteria and this proved to be sufficient to ensure DNA interference in the remote host (Sapranauskas et al., 2011). Additionally, it was shown that Cas9 is the only protein from entire CRISPR-Cas system required for the interference and mutations in Cas9 HNH or RuvC active sites abolishes this effect. Also, using plasmid interference assay, researchers experimentally confirmed the requirement of the short sequences (2-5 nt) adjacent to the DNA targets 3'-ends, termed PAMs (protospacer adjacent motifs) for efficient DNA interference (Sapranauskas et al., 2011). These sequences were identified earlier by analyzing DNA targets matching spacer sequences in CRISPR arrays (Bolotin et al., 2005; Deveau et al., 2008; Horvath et al., 2008; Mojica et al., 2009) and provide discrimination between self and non-self DNA.

Eventually all necessary puzzle components were characterized and identified: Cas9 nuclease, crRNA and tracrRNA. It became a matter of time to demonstrate the ability of this system to cleave DNA *in vitro* and *in vivo*. Finally, in 2012 two independent studies showed that Cas9 can be used to cleave DNA *in vitro* (Gasiunas et al., 2012; Jinek et al., 2012). *S. thermophilus* Cas9 complex purified from *E. coli* cells (Gasiunas et al., 2012) and

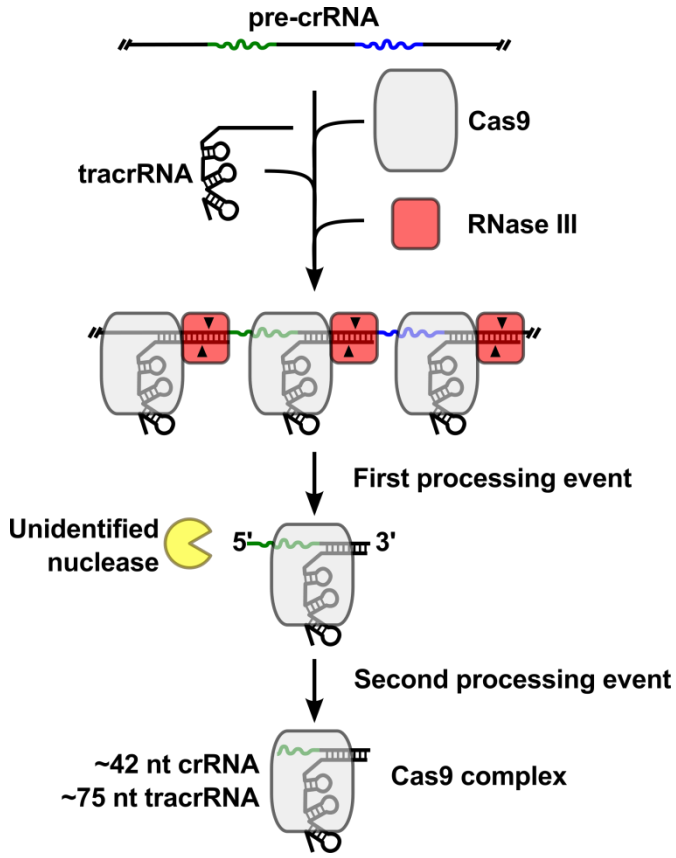


**Figure 7. Schematic representation of Type II CRISPR-Cas system crRNA maturation and Cas9 effector complex assembly.** In Type II CRISPR-Cas systems additional RNA molecule, termed tracrRNA, is required for pre-crRNA maturation. tracrRNA binds with almost perfect complementarity to each repeat sequence in pre-crRNA transcript. Resulting repeat-tracrRNA duplexes are recognized and cleaved by host RNase III in the presence of Cas9 (first processing event). crRNA undergo further trimming at the 5'-end by unidentified nuclease to produce mature Cas9 effector complex (second processing event).

*S. pyogenes* Cas9 complex assembled from separate components *in vitro* (Jinek et al., 2012) efficiently generated DSBs 3 nt upstream of the PAM sequence, confirming the observation from the data obtained in *in vivo* studies (Garneau et al., 2010). For the cleavage to take place, both crRNA and tracrRNA were required – 20 nt length 5'-end of the crRNA for base pairing with target DNA while 22 nt of 3'-end for duplex formation with tracrRNA to promote efficient complex formation. By fusing those into single-guide RNA (sgRNA) scientists were able to further simplify the system (Jinek et al., 2012). Cas9 complex recognizes DNA targets through Watson-Crick base pairing between guide RNA and DNA, thus by changing guide RNA sequences both studies were able to successfully reprogram Cas9s to cleave novel DNA targets (Gasiunas et

al., 2012; Jinek et al., 2012). In addition, mutating HNH and RuvC active sites revealed that each domain is responsible for cleaving opposite DNA strands – HNH cleaves complementary to crRNA DNA strand, while RuvC cuts displaced DNA strand. As a result, mutations in either one of the Cas9 active sites converted Cas9 endonuclease into a nickase (Cas9n) (Gasiunas et al., 2012; Jinek et al., 2012). At this point researches realized that Cas9 could potentially be the powerful new tools in genome editing field, but the fact that bacterial Cas9 protein could work in large and complex eukaryotic genome raised some concerns (Barrangou, 2012). Nevertheless, after few months all skepticism was dispersed when the first studies demonstrated that *S. pyogenes* Cas9 complex could be effectively used as a genome editing tool in human (Cong et al., 2013; Jinek et al., 2013; Mali et al., 2013a), mouse (Cong et al., 2013; Wang et al., 2013), zebrafish (Chang et al., 2013; Hwang et al., 2013), yeast (DiCarlo et al., 2013), bacteria (Jiang et al., 2013) and plant (Nekrasov et al., 2013) cells. Additionally, catalytically inactive Cas9 (dCas9) with sgRNA has been used as a programmable DNA binding scaffold. Using dCas9 with sgRNA investigators were able to sterically hinder RNA polymerase binding sites leading to the downregulated gene expression in bacteria (Bikard and Marraffini, 2013; Qi et al., 2013). Also, by fusing dCas9 with effector domains it was demonstrated the ability to address these proteins using sgRNA to specific DNA loci and induce or repress gene expression (Gilbert et al., 2013; Maeder et al., 2013; Perez-Pinera et al., 2013), modify histones or DNA bases (Hilton et al., 2015; Kearns et al., 2015; Komor et al., 2016; Vojta et al., 2016) and track DNA location *in vivo* (Chen et al., 2013). Furthermore, with PAM providing DNA oligonucleotides Cas9 can be used to bind and degrade RNA *in vitro* (O'Connell et al., 2014) or track RNA *in vivo* (Nelles et al., 2016). Cas9 was also adapted for systematic genetic analysis. Gene knockout screens using Cas9 and lentiviral sgRNA libraries in cells followed by high-throughput sgRNA sequencing after negative or positive selection allowed to identify genes essential for cell survival in the provided conditions (Koike-Yusa et al., 2013; Shalem et al., 2014; Wang et al., 2014; Zhou et al., 2014).

Recently, other Type V CRISPR-Cas RNA-guided nucleases from Class 2 have been used for genome editing in human cells. As Cas9, Type V-A signature protein Cpf1 uses 42-44 crRNA, consisting of 19 nt repeat derived 5' sequence and 23 nt spacer derived 3' guide sequence to recognize target DNA (Zetsche et al., 2015a). By utilizing RuvC and newly identified nuclease domain, Cpf1 makes staggered DSB in PAM dependent fashion that could be advantageous in some situations over Cas9 that produces blunt ends in target DNA (Dong et al., 2016; Yamano et al., 2016; Zetsche et al., 2015a). Furthermore, it was shown that in addition to DNAse activity, Cpf1 is also involved in pre-crRNA maturation. Cpf1 cleaves pre-crRNA transcripts (instead of RNAse III in Type II CRISPR-Cas systems) leading to intermediate crRNAs that are processed further into mature crRNAs (Fonfara et al., 2016). It is still too early to predict if Cpf1 will become as popular as Cas9, but the resemblance with Cas9 makes it an attractive alternative.

## 1.2. Comparison of technologies

In order to compare different technologies it is crucial to determine what features of nucleases are important for genome editing. Ideal nuclease has to have long target sites (>16 bp to be not found once by chance in human genome), ability to be easily reengineered, high on-target activity and high specificity (low or none off-targets). All above described platforms meet the target site length requirements (homing endonucleases can recognize >16 bp, ZFN pairs 24-36 bp, TALEN pairs 24-38 bp and Cas9 from *S. pyogenes* 22 bp). All technologies have more or less comparable on-target activity, reaching up to 80% editing efficiencies using Cas9 (Zuris et al., 2014), although off-target activity greatly varies, favoring TALENs and Cas9 platforms (Porteus, 2016). Another story is the ease of programmability. Despite the fact that genome editing tools evolution from homing endonucleases to TALENs dramatically simplified the reengineering of nucleases for novel target sites, it is hard to find the simpler alternative for Cas9 complex offered Watson-Crick

based crRNA and DNA base pairing for target recognition where no protein engineering is required. As a result, Cas9 based genome engineering technology became method of choice for practically all laboratories interested in genome editing.

## 1.3. Cas9 DNA cleavage mechanism

Rising Cas9 popularity as a genome editing field demanded to understand the mechanism by which Cas9 complex functions. A number of *S. pyogenes* and few other Cas9 structures in different substrate bound state were solved revealing the basic principles of molecular mechanism of DNA targeting (Anders et al., 2014, 2016, Hirano et al., 2016a, 2016b, Jiang et al., 2015, 2016; Jinek et al., 2014; Nishimasu et al., 2014, 2015). Basically, Cas9 complex consists of nuclease (NUC) and recognition (REC) lobes with nucleic acids positioned in positively charged groove between them (Figure 8A and B). NUC lobe composed of HNH and RuvC domains is responsible for DNA cleavage, and PAM recognition with PAM interacting (PI) domain. REC lobe is essential for binding nucleic acids and contains positively charged bridge helix (BH), connecting two lobes (Anders et al., 2014; Hirano et al., 2016a; Jinek et al., 2014; Nishimasu et al., 2014, 2015). During target recognition, Cas9 complex scans double stranded DNA for PAM sequences making specific protein-DNA contacts before DNA strand separation and R-loop formation initiation. R-loop is initiated at the PAM proximal nucleotides and propagates to PAM distal end leading to the both DNA strands cleavage (Sternberg et al., 2014). A perfect or near perfect complementarity between guide RNA and DNA is required, with some mismatches tolerated better at PAM distal end than in the first 10-12 nt from the PAM, named seed region (Figure 8C) (Cong et al., 2013; Jiang et al., 2013; Jinek et al., 2012; Sternberg et al., 2014). Recent studies revealed, that Cas9 bends DNA helix by 30°

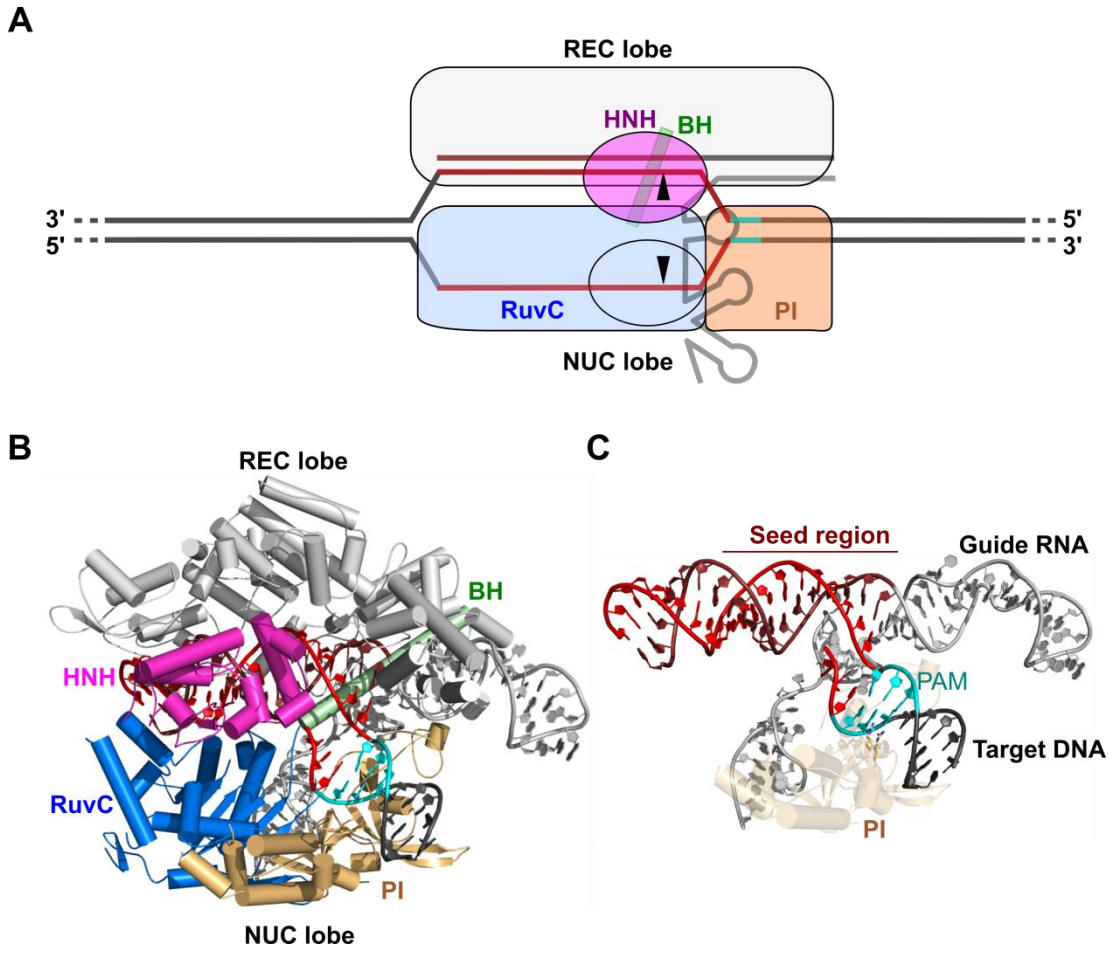


**Figure 8. CRISPR-Cas9 RNA-guided nuclease.** (**A**) Schematic representation of Cas9 complex bound to target DNA. Cas9 complex consists of nuclease (NUC) and recognition (REC) lobes with (BH) connecting two lobes. NUC lobe composed of HNH and RuvC domains is responsible for DNA cleavage, and PAM recognition with PAM interacting (PI) domain. DNA cleavage position is indicated by black triangles. (**B**) Structure of *S. pyogenes* Cas9 bound to DNA (PDB id: 4UN3). (**C**) DNA target can be divided into two parts: PAM sequence (indicated in cyan) and protospacer (indicated in red). PAM is recognized by Cas9 protein (residues located in PI domain make base specific contacts) while protospacer by base pairing with spacer component of the guide RNA.

providing structural distortion required for R-loop formation initiation (Jiang et al., 2016). In addition, seed region of guide RNA is positioned in A-form helix leading to the energetic compensation of dsDNA unwinding and hybridization to guide RNA (Jiang et al., 2015). Mechanism by which Cas9 is activated to cleave target DNA might be linked with HNH domain structural flexibility, since HNH domain position in crystal structures varies and are distant from the complementary strand cleavage site (Anders et al., 2014; Hirano et al., 2016a; Jiang et al., 2016; Nishimasu et al., 2014, 2015). In agreement, FRET experiments confirmed HNH domain movement towards scissile phosphate in

complementary DNA strand after full length heteroduplex formation leading to the RuvC domain activation and concerted DNA cleavage forming DSB (Sternberg et al., 2015).

## 1.4. Off-target cleavage

Initial experiments demonstrating Cas9 ability to be used as efficient genome editing tool were immediately followed by concerns regarding off-target cleavage. Early attempts to identify possible off-targets were limited to the analysis of similar to target sequences that were cleaved by Cas9 complex in various experimental setups. Thus this type of off-target characterization was biased by ignoring all available sequences in the genome wide context. Nevertheless, these experiments revealed that *S. pyogenes* Cas9 can cleave up to 5 nt mismatches containing target sites (favoring PAM-distal end) with high efficiency. Tolerable mismatches were not limited to the spacer sequences but also for the PAMs, where targets containing non-canonical 5'-NAG-3' PAM sequences (compared to canonical 5'-NGG-3') were efficiently cleaved as well (Fu et al., 2013; Hsu et al., 2013; Mali et al., 2013b; Pattanayak et al., 2013). To overcome these limitations and get more comprehensive profile of possible Cas9 induced genome wide off-targets, a number of methods were adapted or developed (see Table 1). Whole genome sequencing (WGS) is potentially the least biased method to identify INDELs after NHEJ repair in Cas9 treated cells, but suffers from the low sensitivity. Available 30-50× read coverage of the genome after high-throughput sequencing (Tsai and Joung, 2016) allows to identify off-targets in single-cell clones or non-mosaic $F_1$ animals but only with the limited sensitivity (Smith et al., 2014; Veres et al., 2014). Similar problems are present in method named digested genome sequencing (Digenome-seq). In this case purified genomic DNA is cleaved with assembled Cas9 complex *in vitro* and sequenced. Those reads that are relatively enriched possessing the same start or end sequences compared to random DNA breaks are interpreted as Cas9 cleavage products (Kim et al., 2015, 2016). The

**Table 1. Comparison of methods used in detection of off-target cleavage events.**

| Method | Details | Reference |
|---|---|---|
| Bioinformatic analysis | Allows to identify off-target cleavage events at similar to target sequences. | Fu et al., 2013; Hsu et al., 2013; Mali et al., 2013b; Pattanayak et al., 2013 |
| WGS | Whole genome sequencing allows to identify INDELs after NHEJ repair in Cas9 treated cells in single-cell clones or non-mosaic $F_1$ animals. | Smith et al., 2014; Veres et al., 2014 |
| Digenome-seq | Purified genomic DNA is cleaved with assembled Cas9 complex *in vitro* and sequenced. | Kim et al., 2015, 2016 |
| HTGTS | On target sequence is used as "pray" to detect the translocation events with off-target cleavage sites, termed "baits". | Frock et al., 2014 |
| IDLV | Based on integration of linear integrase-defective lentiviral vector at DSBs followed by amplification. | Gabriel et al., 2011; Osborn et al., 2016; Wang et al., 2015 |
| GUIDE-seq | Based on integration of double-stranded oligodeoxynucleotide at DSBs followed by amplification. | Tsai et al., 2015 |
| BLESS | Adapters are ligated to DSBs *in situ* in fixed cells followed by enrichment on streptavidin, amplification and sequencing. | Crosetto et al., 2013; Ran et al., 2015; Slaymaker et al., 2016 |

possibility to use high amounts of Cas9 complex relative to DNA makes this approach attractive, because it might allow to maximize rare off-target cleavage events, but the bottle neck of sensitivity becomes available read coverage of high-throughput sequencing (Tsai and Joung, 2016). Another strategy in off-target detection is utilized in the method named high-throughput genome-wide translocation sequencing (HTGTS). In this case on target sequence is used as "pray" to detect the translocation events with off-target cleavage sites, termed "baits". After sequencing and analysis "bait" sequences are mapped to the genome to identify off-target sites (Frock et al., 2014). The main disadvantages of this method are that translocations are quite rare events in the cells and the resulting products might have negative effects on the cells leading to the underestimated score of off-targets. More sensitive methods, named IDLV (integrase-defective lentiviral vector) capture (Gabriel et al., 2011; Osborn et al., 2016; Wang et al., 2015) and genome wide unbiased

identification of DSBs enabled by sequencing (GUIDE-seq) (Tsai et al., 2014a) were also adapted for off-target identification. Both approaches are based on integration of DNA sequences (linear vector or double-stranded oligodeoxynucleotide respectively) at the formed DSB sites after treatment with Cas9 followed by DNA amplification. Because of the relatively high level of background integration these methods suffer from high false-positive error rates. BLESS (breaks labeling, enrichment on streptavidin and next-generation sequencing) allows to capture DSBs at the specific moment in time, because it uses direct adapter ligation to DSBs *in situ* in fixed cells followed by amplification and sequencing (Crosetto et al., 2013; Ran et al., 2015; Slaymaker et al., 2016). However those DSBs that have already been resolved before cells permeabilization and fixation remain undetected.

All above described genome wide Cas9 induced off-target detection studies revealed a relatively comprehensive picture of these events. The most important observation was that Cas9 complex is able to cleave similar to target sequences, tolerating mismatches not only in the spacers, but also in the PAMs. Thus, these experiments resulted in a number of studies focused on improving CIRSPR-Cas9 system specificity.

## 1.5. Strategies to minimize off-target effects

Basically, two general strategies could be adapted to minimize off-target effects of nucleases: improving the specificity and limiting the time when nuclease is active in the cells. As a result, both approaches were successfully utilized in the attempts to reduce off-target cleavage by Cas9.

### 1.5.1. Improving Cas9 specificity

A number of strategies were adapted in order to improve the Cas9 specificity (see Table 2). The least intuitive approaches were based on sgRNA modifications. sgRNA 5’-end truncation by 2-3 nt or addition of two G

nucleotides at the guide RNA 5'-end resulted in reduced cleavage rate in some of the validated off-target sites (Cho et al., 2014; Fu et al., 2014). The rationale behind this is not well understood, but could be related with decreasing the RNA-DNA interaction energy or disruption of protein interaction with guide RNA 5'-end, respectively, thus only producing sufficient amount of interaction energy with on target sites but reduced tolerance for off-targets.

Somewhat more familiar approaches adapted from ZFNs or TALENs technologies were based on paired Cas9 nickases (Cas9n) (Mali et al., 2013b; Ran et al., 2013) or catalytically inactivated Cas9 (dCas9) fused with dimerization dependent FokI DNA cleavage domain (Guilinger et al., 2014; Tsai et al., 2014b) to induce DSBs in the target DNA. Co-localization of paired nickases in PAM-out orientation where each Cas9n independently cleaves single DNA strand resulted in efficient DSB formation (Mali et al., 2013b; Ran et al., 2013). The remaining uncertainty in this method was that each Cas9n could cleave elsewhere in the genome resulting in off-target mutations. These possible monomeric cleavage behaviors were eliminated in the approach based on fusing dCas9 with FokI, where DNA cleavage was only dependent on the

**Table 2. Strategies used to improve Cas9 specificity.**

| Strategy | Details | Reference |
|---|---|---|
| Guide RNA modification | Guide RNA 5'-end truncation by 2-3 nt (reduced excessive RNA–DNA interaction energy). | Fu et al., 2014 |
| | Addition of two G nucleotides at the guide RNA 5'-end (destabilization of protein interaction with the 5′-end of the guide RNA). | Cho et al., 2014 |
| Paired Cas9 nickases (Cas9n) | Co-localization of paired nickases in PAM-out orientation where each Cas9n independently cleaves single DNA strand. | Mali et al., 2013b; Ran et al., 2013 |
| Fusing dCas9 with FokI nuclease | Fusing dCas9 with FokI, where DNA cleavage is dependent on the dimerization of FokI nuclease domains. | Guilinger et al., 2014; Tsai et al., 2014 |
| Engineered Cas9 variants | SpCas9-HF1 (high-fidelity variant 1) contains alanine substitutions of residues responsible for nonspecific interaction with target DNA strand. | Kleinstiver et al., 2016 |
| | eSpCas9 1.1 (enhanced SpCas9 version 1.1) contains alanine substitutions of residues responsible for nonspecific interaction with non-target DNA strand. | Slaymaker et al., 2016 |

dimerization of FokI nuclease domains (Guilinger et al., 2014; Tsai et al., 2014b). Although, both strategies proved to be very effective in increasing specificity and yielding in some cases non-distinguishable off-target cleavage levels from background, but introduced requirement to target two adjacent DNA sequences restricted the DNA targeting range and limited the wider use of these methods.

Recently, there have been rationally engineered two Cas9 protein variants (Kleinstiver et al., 2016; Slaymaker et al., 2016) based on crystal structures of *S pyogenes* Cas9 complexes (Anders et al., 2014; Nishimasu et al., 2014). SpCas9-HF1 (high-fidelity variant 1) (Kleinstiver et al., 2016) and eSpCas9 1.1 (enhanced SpCas9 version 1.1) (Slaymaker et al., 2016) contained alanine substitutions of residues responsible for nonspecific interaction with complementary or non-complementary DNA strands, respectively. Disruption of these contacts allowed to retain enough binding energy for efficient on-target cleavage, but activity on the most off-target sites was reduced to the undetectable levels.

### 1.5.2. Modulating Cas9 activity in the cells

Another way to regulate Cas9 activity in the cells lies in the way how Cas9 is delivered. There are three methods based on introducing DNA expression vectors, RNAs or RNP complexes. In the latter two cases the half-life of Cas9 complex in the cells is reduced comparing to the Cas9 delivered by plasmid DNA, where protein can be expressed for several days (Kim et al., 2014), possessing the increased potential for off-target effects. In agreement, analyzes of previously identified off-target sites after RNP delivery to the cells revealed reduced off-target activity when compared to plasmid DNA transfection (Kim et al., 2014; Liang et al., 2015; Zuris et al., 2014). Also, different strategy to limit the time when Cas9 is active in the cells was utilized by reengineering Cas9 protein. By splitting Cas9 into domains fused with proteins capable to dimerize when stimulus is present (Nihongaki et al., 2015;

Zetsche et al., 2015b) or by catalytically repressing Cas9 by inserting a drug inducible intein (Davis et al., 2015) investigators were able to successfully improve the on-target and off-target cleavage ratio.

All these examples perfectly illustrates what a great progress has been made both in increasing specificity of Cas9 and off-target detection, but the room for improvements still remains. Comparison of off-target detection methods using the same guide RNAs in the same experimental conditions would provide the information allowing to evaluate and standardize assays to detect off-targets. Also, combining strategies from different studies could allow to further increase the specificity of Cas9, allowing to get closer to the "perfect" tool for genome editing.

## 1.6. Expanding the Cas9 toolbox for genome editing

DNA target of Cas9 RNP complex can be divided into two components. Variable component, termed protospacer is recognized by Watson-Crick base pairing with spacer sequence (~20 nt) of the guide RNA. As a result, by customizing guide RNA spacer sequence it is possible to program Cas9 to recognize, in theory, any protospacer. However, invariable component that is absolutely required to initiate guide RNA-mediated DNA binding is a short nucleotide sequence (2-5 nt) (Gasiunas et al., 2012; Jinek et al., 2012; Sternberg et al., 2014), termed PAM, which is usually located adjacent to a protospacer sequence and is uniquely associated with each Cas9 protein (Deveau et al., 2008; Fonfara et al., 2014; Mojica et al., 2009). Cas9 using hydrogen bonding and Van der Waals interactions between amino acid residues located in the PI (PAM interacting) domain specifically recognizes PAM sequence and licenses subsequent DNA unwinding and base pairing with guide RNA (Anders et al., 2014, 2016, Hirano et al., 2016a, 2016b). Thus, recognizable PAM sequence is the major factor restricting the choice of targets in the genome. However, PAM recognition mechanism suggested the possible solution for this issue. Experiments, using rational design or directed evolution

approaches revealed the possibility to alter Cas9 PAM recognition specificity. In engineered *S. pyogenes* Cas9 variants specificity from initial 5'-NGG-3' PAM were altered to 5'-NGA-3', NGCG-3', 5'-NGNG-3' or 5'-NAAG-3' PAMs (Anders et al., 2016; Kleinstiver et al., 2015a). Same strategy resulted in the variant of *S. aureus* Cas9 with relaxed PAM recognition from 5'-NNGRRT-3' to 5'-NNNRRT-3' (Kleinstiver et al., 2015b). *F. novicida* Cas9 engineered protein was able to recognize 5'-NYG-3' PAM, where wild-type Cas9 were limited to 5'-NGG-3' (Hirano et al., 2016a). These examples illustrate the feasibility of reengineering approaches to expand the toolbox of Cas9s for genome editing, but available repertoire is still too limited to target any desired DNA sequence with nucleotide precision.

Another strategy to overcome this issue was offered by nature itself. With >1000 Cas9 sequences available in sequence databases and the continued sequencing of microbial genomes, Cas9 orthologues are abundant (Fonfara et al., 2014; Makarova et al., 2015). The availability of new Cas9 proteins may also open the way for orthogonal genome engineering allowing different modifications (for example, DNA cleavage and transcriptional activation or silencing) to be performed simultaneously (Esvelt et al., 2013). However, methods to determine the PAM sequence requirement for new Cas9 proteins are limited and can be divided into three main categories. First, PAM sequences of new Cas9 proteins can be identified by bioinformatic analysis of sequences immediately flanking matching protospacers in bacteriophage genomes (Shah et al., 2013). With most of the spacers in available Type II CRISPR arrays exhibiting only a few if any matches to available phage sequences in the databases, this approach limits the exploration of Cas9 protein diversity for genomic applications. Another method is based on *in vitro* cleavage assays with DNA substrates containing various PAMs (Fonfara et al., 2014; Jinek et al., 2012). This approach, in theory, allows to identify consensus PAM sequence, but suffers of its low-throughput nature. Finally, the most recent methods are based on interrogation of PAM libraries and allow

examination of all available PAM sequences spanning up to 6-8 nucleotides from protospacer.

High-throughput *in vivo* screens are based on plasmid depletion experiments (Esvelt et al., 2013; Jiang et al., 2013; Kleinstiver et al., 2015a) or binding of dCas9 to *lacI* repressor promoter (Leenay et al., 2016). In plasmid depletion experiments, bacterial cells bearing plasmids with Cas9 are co-transformed with plasmid library containing randomized PAMs and gene providing resistance to antibiotic. Plasmids with functional PAMs are recognized by Cas9 and cleaved leading to cell death. As a result, sequencing and analysis of the plasmids purified from the survived cells allows identification of depleted functional PAMs (Esvelt et al., 2013; Jiang et al., 2013). Different strategy is used in the method, based on dCas9 binding to *lacI* repressor promoter. To associate Cas9 binding with positive signal, protospacer with randomized PAMs are cloned within the promoter upstream of *lacI*, which product blocks the promoter of *gfp*. Thus, only functional PAMs would result in Cas9 binding and lacI repression leading to reporter expression. After fluorescence activated cell sorting (FACS), plasmid purification and sequencing it is possible to identify functional PAMs (Leenay et al., 2016). Another group of high-throughput methods is based on PAM library cleavage *in vitro*, followed by adapter ligation and amplification (Pattanayak et al., 2013; Ran et al., 2015). These experiments require active RNP complexes, which can be assembled from purified components (Pattanayak et al., 2013) or directly in the cell extracts (Ran et al., 2015).

These strategies proved to be successful in identifying PAMs for novel Cas9 proteins, but suffered from low sensitivity. The point is perfectly reflected in the inability to precisely reproduce the PAM recognition of *S. thermophilus* Cas9 protein from CRISPR1 system (Esvelt et al., 2013; Leenay et al., 2016; Ran et al., 2015) originally reported by Horvath et al., 2008. Thus, in order to successfully expand Cas9 toolbox for genome editing there is still a need for more sensitive high-throughput PAM identification strategies.

# 2. MATERIALS AND METHODS

## 2.1. Materials

### 2.1.1. Chemicals

All chemicals used in this study were of the highest quality available.

### 2.1.2. Enzymes

T4 Polynucleotide Kinase, FastAP Thermosensitive Alkaline Phosphatase, Pfu and DreamTaq DNA polymerases, T4 DNA ligase, FastDigest restriction enzymes, Proteinase K, RNase A and RNase III used in this study were obtained from Thermo Fisher Scientific. Phusion DNA polymerases – from Thermo Fisher Scientific and NEB. All these products were used according to the manufacturer's protocols.

### 2.1.3. Kits for molecular biology

“Rapid DNA Ligation Kit”, “GeneJET Gel Extraction Kit”, “GeneJET PCR Purification Kit”, “GeneJET Plasmid Miniprep Kit” and “TranscriptAid T7 High Yield Transcription Kit”, “GeneJET RNA Cleanup and Concentration Micro Kit” were purchased from Thermo Fisher Scientific, “miRNeasy Mini Kit”, “QIAquick PCR Purification Kit” – from Qiagen and “SURVEYOR Mutation Detection Kit for Standard Gel Electrophoresis” – from Transgenomic. Kits were used according to the manufacturer’s protocols.

### 2.1.4. Bacterial strains

*E. coli* strain DH5α ($F^-$ endA1 glnV44 thi-1 recA1 relA1 gyrA96 deoR nupG Φ80dlacZΔM15 Δ(lacZYA-argF)U169, hsdR17($r_K^-$ $m_K^+$), λ–) was used for the cloning procedures.

*E. coli* strain RR1 (F- mcrB mrr hsdS20(rB- mB-) leuB6 ara-14 proA2 lacY1 galK2 xyl-5 mtl-1 rpsL20(SmR)glnV44 λ-) was used for plasmid interference assay.

*E. coli* strain BL21 (DE3) ($F^-$ ompT gal dcm lon $hsdS_B(r_B^- m_B^-)$ λ(DE3 [lacI lacUV5-T7 gene 1 ind1 sam7 nin5]) was used for plasmid interference assay and Cas9 proteins expression.

*E. coli* strain DH10B (F- endA1 recA1 galE15 galK16 nupG rpsL ΔlacX74 Φ80lacZΔM15 araD139 Δ(ara,leu)7697 mcrA Δ(mrr-hsdRMS-mcrBC) λ-) was used for Cas9 proteins expression.

### 2.1.5. Cell lines

CHO-K1: adherent cell line derived from the ovary of the Chinese hamster.

HEK293T: adherent cell line derived from human embryonic kidney cells grown in tissue culture and expressing SV40 Large T-antigen.

### 2.1.6. Plasmids

All plasmids used in this study are listed in Table 3.

### 2.1.7. RNAs

Origins of all RNA molecules used in this study are listed in Table 4.

### 2.1.8. Oligonucleotides

Sequences of all oligonucleotides and primers used in this study are listed in Table 5.

**Table 3. Plasmids used in this study.**

| Plasmid | Description | Reference |
|---|---|---|
| pBAD24-Sth3-Cas9-CHis | *cas9* of *Streptococcus thermophilus* CRISPR3-Cas system cloned into pBAD24-CHis expression vector | Karvelis et al., 2013 |
| pBAD24-Sth1-Cas9-CHis | *cas9* of *Streptococcus thermophilus* CRISPR1-Cas system cloned into pBAD24-CHis expression vector | Karvelis et al., 2015 |
| pBAD24-Spy-Cas9-CHis | *cas9* of *Streptococcus pyogenes* cloned into pBAD24-CHis expression vector | Karvelis et al., 2015 |
| pBAD24-Sth3-Cas9-NLS-CHis | *cas9* of *Streptococcus thermophilus* CRISPR3-Cas system fused with SV40T nuclear localization signal (NLS) cloned into pBAD24-CHis expression vector | Glemzaite et al., 2015 |
| pBAD24-Blat-Cas9-CHis | *cas9* of *Brevibacillus laterosporus* cloned into pBAD24-CHis expression vector | Karvelis et al., 2015 |
| pCRISPR3 | *Streptococcus thermophilus* complete CRISPR3-Cas system (~7.6 kb) cloned into pACYC184 plasmid | Sapranauskas et al., 2011 |
| pCRISPR3-Δt | pCRISPR3 plasmid lacking an entire tracrRNA encoding gene | Karvelis et al., 2013 |
| pCRISPR3-ΔtR | pCRISPR3 plasmid containing tracrRNA fragment from the transcription start site to the end of anti-repeat sequence | Karvelis et al., 2013 |
| ptracrRNA | pCDF-DUET plasmid with cloned tracrRNA encoding sequence under control of a T7 RNA polymerase promoter | Karvelis et al., 2013 |
| pMTC-RFP/EGFP dual reporter plasmid | Plasmid containing intron-split eGFP gene inserted into RFP gene, engineered with 400 bp direct homologous repeats on both sides of the eGFP insert (see Figure 15B) | Glemzaite et al., 2015 |
| pS1 | pUC18 plasmid with cloned 30 nt protospacer S1 (matching to the S1 spacer in the CRISPR3 system of *S. thermophilus*) and PAM – NGGNG (crRNA in the Cas9 complex matches only 20 nt in the 30 nt protospacer) | Sapranauskas et al., 2011 |
| pS1' | pUC18 plasmid with cloned 20 nt protospacer S1' matching to the 20 nt spacer of crRNA | Szczelkun et al., 2014 |
| pS1′Δ1 | pS1' plasmid with 1 nt 5′-end truncation of the S1′ protospacer (19 nt complementary to crRNA) | Szczelkun et al., 2014 |
| pS1′Δ3 | pS1' plasmid with 3 nt 5′-end truncation of the S1′ protospacer (17 nt complementary to crRNA) | Szczelkun et al., 2014 |

*See next page for Table 3 extension*

*Extension of the Table 3*

| | | |
|---|---|---|
| pS1′Δ5 | pS1' plasmid with 5 nt 5′-end truncation of the S1′ protospacer (15 nt complementary to crRNA) | Szczelkun et al., 2014 |
| pS1′Δ7 | pS1' plasmid with 7 nt 5′-end truncation of the S1′ protospacer (13 nt complementary to crRNA) | Szczelkun et al., 2014 |
| pS1′Δ9 | pS1' plasmid with 9 nt 5′-end truncation of the S1′ protospacer (11 nt complementary to crRNA) | Szczelkun et al., 2014 |
| pS1-G1C | pS1 plasmid with G1C mutation in PAM (30 nt S1 protospacer) | Szczelkun et al., 2014 |
| pS1-G2C | pS1 plasmid with G2C mutation in PAM (30 nt S1 protospacer) | Szczelkun et al., 2014 |
| pS1-G4C | pS1 plasmid with G4C mutation in PAM (30 nt S1 protospacer) | Szczelkun et al., 2014 |
| pS1-5N-PAM | pTZ57R/T pasmids library with cloned S1 protospacer and randomized 5 bp PAM | Karvelis et al., 2015 |
| pS1-7N-PAM | pTZ57R/T plasmids library with cloned S1 protospacer and randomized 7 bp PAM | Karvelis et al., 2015 |
| pS1-GTCCCGAA | pUC18 plasmid with cloned S1 protospacer and GTCCCGAA PAM sequence | Karvelis et al., 2015 |
| pS1-GTCaCGAA | pUC18 plasmid with cloned S1 ptotospacer and GTCACGAA PAM sequence | Karvelis et al., 2015 |
| pS1-GTCCtGAA | pUC18 plasmid with cloned S1 protospacer and GTCCTGAA PAM sequence | Karvelis et al., 2015 |
| pS1-GTCCCGcA | pUC18 plasmid with cloned S1 protospacer and GTCCCGCA PAM sequence | Karvelis et al., 2015 |
| pS1-GTCCCGAc | pUC18 plasmid with cloned S1 protospacer and GTCCCGAC PAM sequence | Karvelis et al., 2015 |
| pS1-GTCCCGcc | pUC18 plasmid with cloned S1 protospacer and GTCCCGCC PAM sequence | Karvelis et al., 2015 |

**Table 4. RNAs used in this study.**

| Name | Sequence (5' to 3') | Origin |
|---|---|---|
| Sth3 tracrRNA3 (105 nt) | gggguaauaauauugugguuugaaaccauucgaaacaacacagcgaguuaaaauaaggcuuaguccguacucaacuugaaaagguggcaccgauucgguguuuuu | *In vitro* transcription |
| Sth1 tracrRNA1 (105 nt) | gggguguaaggggcgccuuacacaguuacuuaaaucuugcagaagcuacaaagauaaggcuucaugccgaaaucaacacccugucauuuuauggcaggguguuuu | *In vitro* transcription |
| Sth3 pre-crRNA3 (150 nt) | ggguagaaaagauauccuacgaggguuuuagagcuguguuguuucgaaugguuccaaaacaaauucuaaacgcuaaagaggaagaggacaguuuuagagcuguguuguuucgaaugguuccaaaacuacugcuguauuagcuuggguuguug | *In vitro* transcription |
| Sth3 pre-crRNA3 (94 nt) | gggaaauucuaaacgcuaaagaggaagaggacaguuuuagagcuguguuguuucgaaugguuccaaaacuacugcuguauuagcuuggguuguug | *In vitro* transcription |
| Sth1 pre-crRNA1 (94 nt) | ggguguuugacagcaaaucaagauucgaauuguguuuuuguacucucaagauuuaaguaacuguacaacaaugacgaggagcuauuggcacaac | *In vitro* transcription |
| Sth3 tracrRNA (78 nt) | gggcgaaacaacacagcgaguuaaaauaaggcuuaguccguacucaacuugaaaagguggcaccgauucgguguuuuu | *In vitro* transcription |
| Sth3 tracrRNA (73 nt) | gggcgaaacaacacagcgaguuaaaauaaggcuuaguccguacucaacuugaaaagguggcaccgauucggug | *In vitro* transcription |
| Sth3 tracrRNA (68 nt) | gggcgaaacaacacagcgaguuaaaauaaggcuuaguccguacucaacuugaaaagguggcaccgauu | *In vitro* transcription |
| Sth3 tracrRNA (63 nt) | gggcgaaacaacacagcgaguuaaaauaaggcuuaguccguacucaacuugaaaagguggcac | *In vitro* transcription |
| Sth3 tracrRNA (58 nt) | gggcgaaacaacacagcgaguuaaaauaaggcuuaguccguacucaacuugaaaaggu | *In vitro* transcription |
| Sth3 tracrRNA (53 nt) | gggcgaaacaacacagcgaguuaaaauaaggcuuaguccguacucaacuugaa | *In vitro* transcription |
| Sth3 tracrRNA (48 nt) | gggcgaaacaacacagcgaguuaaaauaaggcuuaguccguacucaac | *In vitro* transcription |
| Sth3 tracrRNA (43 nt) | gggcgaaacaacacagcgaguuaaaauaaggcuuaguccguac | *In vitro* transcription |
| Sth3 tracrRNA (33 nt) | gggcgaaacaacacagcgaguuaaaauaaggcu | *In vitro* transcription |

*See next page for Table 4 extension*

*Extension of the Table 4*

| | | |
|---|---|---|
| Sth3 crRNA (S1) | cgcuaaagaggaagaggacaguuuuagagcuguguuguuucg | Synthetic oligonucleotide |
| Sth3 pre-crRNA (N1) | ggguagaaaagauauccuacgaggguuuuagagcuguguuguuucgaaugguuccaaaacugucaugauaauaaugguuuc uuagacgucguuuuagagcuguguuguuucgaaugguuccaaaacggauccucuacgccggacgcaucgug | *In vitro* transcription |
| Sth3 pre-crRNA (N2) | ggguagaaaagauauccuacgaggguuuuagagcuguguuguuucgaaugguuccaaaacacgagccggaagcauaaagug uaaagccuggguuuuagagcuguguuguuucgaaugguuccaaaacggauccucuacgccggacgcaucgug | *In vitro* transcription |
| Sth3 pre-crRNA (λ gDNA) | ggguagaaaagauauccuacgaggguuuuagagcuguguuguuucgaaugguuccaaaaccgggagggaagcugcaugaug cgauguuauguuuuagagcuguguuguuucgaaugguuccaaaacggauccucuacgccggacgcaucgug | *In vitro* transcription |
| Sth3 pre-crRNA (*E. coli* gDNA) | ggguagaaaagauauccuacgaggguuuuagagcuguguuguuucgaaugguuccaaaacucaagggagaauagaggcucu cguugcauuguuuuagagcuguguuguuucgaaugguuccaaaacggauccucuacgccggacgcaucgug | *In vitro* transcription |
| Sth3 crRNA (RASGEF1C) | gcucccggggcucgaugaagguuuuagagcuguguuguuucg | Synthetic oligonucleotide |
| Sth3 crRNA (ARL15) | ugaaucgugaaaucugcucaguuuuagagcuguguuguuucg | Synthetic oligonucleotide |
| Sth3 crRNA (NC) | cgcuaaagaggaagaggacaguuuuagagcuguguuguuucg | Synthetic oligonucleotide |
| Sth3 crRNA (EGFP-L1) | cuucagggucagcuugccguguuuuagagcuguguuguuucg | Synthetic oligonucleotide |
| Sth3 crRNA (EGFP-L2) | gcugaagggcaucgacuucaguuuuagagcuguguuguuucg | Synthetic oligonucleotide |
| Sth3 crRNA (DNMT3B) | gcugaauuacucacgccccaguuuuagagcuguguuguuucg | Synthetic oligonucleotide |
| Sth3 crRNA (PPIB) | guguauuuugaccuacgaauguuuuagagcuguguuguuucg | Synthetic oligonucleotide |
| Spy crRNA (S1) | cgcuaaagaggaagaggacaguuuuagagcuaugcuguuuug | Synthetic oligonucleotide |
| Spy tracrRNA | gggaaacagcauagcaaguuaaaauaaggcuaguccguuaucaacuugaaaaaguggcaccgagucggugcuuuuuuu | *In vitro* transcription |

*See next page for Table 4 extension*

*Extension of the Table 4*

| | | |
|---|---|---|
| Spy sgRNA (S1) | gggcgcuaaagaggaagaggacaguuuuagagcuagaaauagcaaguuaaaauaaggcuaguccguuaucaacuugaaaaaguggcaccgagucggugcuuuuuu | *In vitro* transcription |
| Sth3 crRNA (S1) | cgcuaaagaggaagaggacaguuuuagagcuguguuguuucg | Synthetic oligonucleotide |
| Sth3 sgRNA (S1) | gggcgcuaaagaggaagaggacaguuuuagagcuguguuguuucgguuaaaacaacacagcgaguuaaaauaaggcuuaguccguacucaacuugaaaagguggcaccgauucgguguuuuuu | *In vitro* transcription |
| Sth1 crRNA (S1) | cgcuaaagaggaagaggacaguuuuuguacucucaagauuua | Synthetic oligonucleotide |
| Sth1 tracrRNA | ggguaaaucuugcagaagcuacaaagauaaggcuucaugccgaaaucaacacccugucauuuuauggcaggguguuuucg | *In vitro* transcription |
| Sth1 sgRNA (S1) | gggcgctaaagaggaagaggacagtttttgtactctcaagattcaataatcttgcagaagctacaaagataaggcttcatgccgaaatcaacaccctgtcattttatggcagggtgttttcg | *In vitro* transcription |
| Blat sgRNA (S1) direct | gggcgcuaaagaggaagaggacagcuauaguuccuuacugaaagguaaguugcuauaguaagggcaacagacccgaggcguuggggaucgccuagcccguuuuuacgggcucucccccauauucaaaauaaugacagacgagcaccuuggagcauuuauuuccgaggugcuuuuuuuu | *In vitro* transcription |
| Blat sgRNA (S1) reverse | gggcgcuaaagaggaagaggacaaucauaucauaucgaggaaacuugauaugauaugauacuuucauuuuauauccauauaucaucgaagucaaucucauuuaucugucuauuuuaug | *In vitro* transcription |
| Blat sgRNA (S1-3) | gggaaacgctaaagaggaagagggcuauaguuccuuacugaaagguaaguugcuauaguaagggcaacagacccgaggcguuggggaucgccuagcccguuuuuacgggcucucccccauauucaaaauaaugacagacgagcaccuuggagcauuuauuuccgaggugcuuuuuuuu | *In vitro* transcription |

**Table 5. Oligonucleotides and primers used in this study.**

| Oligonucleotide | Sequence (5' to 3') | Details |
|---|---|---|
| GG-322 | gtacggactaagccttattttaactcgctgtgttgt | anti-tracrRNA probe for Northern blot analysis |
| GG-321 | tcgaaacaacacagctctaaaactgtcctcttcctctttagc | anti-crRNA probe for Northern blot analysis |
| dir-pS1 | gcgtaagtctcgagaactagttccgtaagatgcttttctgtgact | Primers used to amplify 2.1 kbp PCR fragment from pS1 plasmid and its derivatives containing the protospacer S1 variants |
| rev-pS1 | gcgtaagtgcggccgcttcgttccactgagcgtcaga | |
| dir-pUC19 | gcgtaagtctcgagaactagtagaatagtgtatgcggcgacc | Amplification 1.0 kbp DNA fragment from pUC19 plasmid labeled with biotin or digoxigenin |
| rev-pUC19 | gcgtaagtgcggccgctgaccatgattacgccaagc | |
| TK-36 | cagcaattataagagatgtatcagaagaagatgc | Generation of 397 bp hybridization probe from *E. coli* BL21(DE3) genomic DNA |
| TK-34 | gcacctttattcccaactgttctttgatttaga | |
| pr-hs1f | tgctgctcgatgcacaggt | Primers used for qPCR to amplify RASGEF1C (HS1) locus |
| pr-hs1r | catcttcaccttcctgctgag | |
| pr-hs2f | ccaaattataagacagatgcctag | Primers used for qPCR to amplify ARL15 (HS2) locus |
| pr-hs2r | gccaacttctgtgaaactacact | |
| EGFP_L1/L2-f | agggcgaggagctgttcacc | Primers used to amplify eGFP DNA fragment (592 bp) for Surveyor nuclease assay |
| EGFP_L1/L2-r | tagtggttgtcgggcagcag | |
| DNMT3B-f | tgagaaggagccacttgctt | Primers used to amplify DNTMT3B DNA fragment (544 bp) for Surveyor nuclease assay |
| DNMT3B-r | gaccaagaacgggaaagtca | |
| PPIB-f | gaacttaggctccgctcctt | Primers used to amplify PPIB DNA fragment (505 bp) for Surveyor nuclease assay |
| PPIB-r | ctctgcaggtcagtttgctg | |
| GG-821N | tgaccatgattacgaattcnnnnntgtcctcttcctctttagcgagc | Oligonucleotides used in construction of a 5 bp randomized PAM library (pS1-5N-PAM) |
| GG-820 | aaggatccccgggtaccgagctgctcgctaaagaggaagaggac | |
| GG-940-G | gtgcacgccggcgacgttgggtcaactnngnnnntgtcctcttcctctttagcgtttag | Oligonucleotides used in construction of a 7 bp randomized PAM library (pS1-7N-PAM) |
| GG-940-C | gtgcacgccggcgacgttgggtcaactnncnnnntgtcctcttcctctttagcgtttag | |
| GG-940-A | gtgcacgccggcgacgttgggtcaactnnannnntgtcctcttcctctttagcgtttag | |
| GG-940-T | gtgcacgccggcgacgttgggtcaactnntnnnntgtcctcttcctctttagcgtttag | |
| GG-939 | gactagacctgcaggggatcccgtcgacaaattctaaacgctaaagaggaagaggac | |

*See next page for Table 5 extension*

*Extension of the Table 5*

| | | |
|---|---|---|
| TK-119 | gagctcgctaaagaggaagagg | Primers used to PCR amplify 5N plasmid library |
| pUC-dir | gccagggttttcccagtcacga | |
| TK-113 | gaaattctaaacgctaaagaggaagagg | Primer used with pUC-dir to PCR amplify 7N plasmid library |
| JKYS800.1 | ctacactctttccctacacgacgctcttccgatctaagtgagctcgctaaagaggaaga | Primers used to introduce barcode for Illumina sequencing to 5N amplified library |
| JKYS803 | caagcagaagacggcatacgagctcttccgatctgaattcgagctcggtacct | |
| JKYS921.1 | ctacactctttccctacacgacgctcttccgatctggaataaacgctaaagaggaagagg | Primers used to introduce barcode for Illumina sequencing to 7N amplified library |
| JKYS812 | caagcagaagacggcatacgagctcttccgatctcggcgacgttgggtc | |
| JKYS557 | aatgatacggcgaccaccgagatctacactctttccctacacg | Universal primers for secondary PCR amplification for Illumina sequencing |
| JKYS558 | caagcagaagacggcata | |
| TK-117 | cggcattcctgctgaaccgctcttccgatct | Oligonucleotides used for adapter generation |
| TK-111 | gatcggaagagcggttcagcaggaatgccg | |
| JKYX1.1 | ctacactctttccctacacgacgctcttccgatctaaggggcgctggccctcctagtc | Primers used to introduce barcode for Illumina sequencing and amplify *Ms45* exon 1 target |
| JKYS178Rd | caagcagaagacggcatacgagctcttccgatctgccggctggcattgtctctg | |
| JKYS1083.1 | ctacactctttccctacacgacgctcttccgatctggaaggcaggttcgcgaacacct | Primers used to introduce barcode for Illumina sequencing and amplify *Ms45* exon 4 target |
| JKYS1084 | caagcagaagacggcatacgagctcttccgatctctccgagacaacaaactgcaggt | |
| JKYX2.1 | ctacactctttccctacacgacgctcttccgatctaaggggccggacgcggtgtt | Primers used to introduce barcode for Illumina sequencing and amplify *liguleless-1* target |
| JKYX3 | caagcagaagacggcatacgagctcttccgatcttacatgcgcaggtgcaaagtctac | |

### 2.1.9. Buffers

Loading buffer: 20 mM $KH_2PO_4$ (pH 7.0), 0.5 M NaCl, 10 mM imidazole, 5 % glycerol.

Cas9 storage buffer: 10 mM Bis-Tris-HCl, (pH 7.0 at 25°C), 300 mM KCl, 1mM EDTA, 1 mM DTT and 50 % (v/v) glycerol.

Cas9 complex assembly buffer: 10 mM Tris-HCl (pH 7.5 at 37°C), 100 mM NaCl and 1 mM DTT, 0.05 mg/ml BSA.

Cas9 complex reaction buffer: 10 mM Tris-HCl (pH 7.5 at 37°C), 100 mM NaCl, 1 mM DTT and 10 mM $MgCl_2$,0.05 mg/ml BSA.

Phenol/chloroform solution: Phenol-Chloroform-Isoamyl Alcohol [25:24:1 (v/v/v)] saturated with 10 mM Tris (pH 8.0), 1 mM EDTA.

3× Loading dye solution: 0.01% (w/v) bromophenol blue and 75 mM EDTA in 50% (v/v) glycerol.

20× SSC: 3M NaCl, 300 mM sodium citrate (pH 7.2).

Adapter ligation buffer: 40 mM Tris-HCl (pH 7.8 at 25°C), 10 mM $MgCl_2$, 10 mM DTT, 0.5 mM ATP, 5 % (w/v) PEG 4000.

TE buffer: 10 mM Tris-HCl (pH 8.0), 0.1 mM EDTA.

## 2.2. Methods

### 2.2.1. Plasmid interference assay

Plasmid interference assays were performed as described previously (Sapranauskas et al., 2011). The *E. coli* RR1 strain was transformed with 1 ng, BL21 (DE3) strain – with 25 ng and HS115 strain – with 100 ng of pUC18 or pS1 plasmid (Table 3). The transformants were plated on LB agar supplemented with appropriate antibiotics and 0.1 mM IPTG (for BL21 (DE3) strain). All transformation experiments were repeated at least three times. Bars in the graphs represent mean values from three or more independent experiments ±1 SD.

### 2.2.2. Northern blot analysis

Northern blot analysis was performed as described previously (Gasiunas et al., 2012). Cas9 bound RNA was isolated from Strep-Tactin–purified Cas9-RNA complex (Gasiunas et al., 2012) using the "miRNeasy Mini Kit" (Qiagen). The RNA was probed with a [$\gamma$-$^{32}$P]ATP (Hartmann Analytic) labeled 36 nt oligodeoxynucleotide GG-322 (Table 5) complementary to tracrRNA or 42 nt oligodeoxynucleotide GG-321 (Table 5) complementary to crRNA. The size of RNAs was estimated by comparison with $^{33}$P-radiolabeled Decade RNA marker (Ambion) and RNA transcripts of different lengths.

### 2.2.3. Expression and purification of Cas9 proteins

Sth1, Sth3, Sth3-NLS and Spy Cas9 proteins were expressed in *E. coli* DH10B while Blat Cas9 in *E. coli* BL21 (DE3) strains grown in LB supplemented with ampicillin (100 mg/ml). After growing bacteria at 37°C and reaching an $OD_{600}$ of 0.5, temperature was decreased to 16°C and expression induced with 0.2 % (w/v) arabinose for 20 h. Cells were pelleted and resuspended in loading buffer and disrupted by sonication. Cell debris was removed by centrifugation. The supernatant was loaded onto the $Ni^{2+}$-charged 5ml HiTrap chelating HP column (GE Healthcare) and eluted with a linear gradient of increasing imidazole concentration. The fractions containing Cas9 were pooled and subsequently loaded onto HiTrap heparin HP column (GE Healthcare) for elution using a linear gradient of increasing NaCl concentration (from 0.5 to 1 M NaCl). The fractions containing Cas9 were pooled and stored at -20°C in Cas9 storage buffer.

### 2.2.4. RNA production

RNAs were either purchased as synthetic oligoribonucleotides (Metabion) or synthesized by *in vitro* transcription using the "TranscriptAid T7 High Yield Transcription Kit" (Thermo Fisher Scientific). For *in vitro*

transcription, T7 promoter at the proximal end of the RNA coding sequence were PCR introduced using plasmids or DNA fragments assembled from oligonucleotides as a templates. After *in vitro* transcription resulting RNA fragments were purified using "GeneJET RNA Cleanup and Concentration Micro Kit" (Thermo Fisher Scientific). All RNAs sequences are listed in Table 4.

For RNA strand annealing activity assay (see section 2.2.7), tracrRNAs, dephosphorylated with FastAP phosphatase (Thermo Fisher Scientific) were radiolabeled with [γ-33P]ATP (Hartmann Analytic) using T4 polynucleotide kinase (Thermo Fisher Scientific).

### 2.2.5. Assembly of Cas9 RNP complexes

Cas9-guide RNA complexes (0.1-2.5 μM) were assembled by mixing Cas9 protein with pre-annealed crRNA and tracrRNA duplex or sgRNA at 1:1 molar ratio followed by incubation in a complex assembly buffer at 37°C for 1 h. To reconstitute active Cas9 complexes using pre-crRNAs generated from PCR products by *in vitro* transcription, Cas9 protein was mixed with pre-crRNA and tracrRNA transcripts at 1:1:2 molar ratio and pre-incubated in a complex assembly buffer at 37°C for 30 min followed by addition of equimolar amount (0.1-2.5 μM) of *E. coli* RNase III (Thermo Fisher Scientific) and $MgCl_2$ (10 mM) and further incubation for additional 30 min at 37°C.

### 2.2.6. Cas9 plasmid DNA cleavage assay

Cas9 DNA cleavage reactions were initiated by mixing supercoiled plasmid DNA with preassembled Cas9 complex and performed at various temperatures. Final reaction mixtures contained 3 nM plasmid DNA and 50 nM Cas9 complex in 100 μl of reaction buffer. Aliquots were removed at timed intervals and quenched with phenol/chloroform. The aqueous phase was mixed with 3× loading dye solution and reaction products analyzed by

electrophoresis through agarose gel and ethidium bromide staining. The amount of supercoiled (SC), open circle (OC), and linear (FLL) DNA forms was evaluated by densitometric analysis of ethidium bromide stained gels using the software ImageJ (National Institutes of Health). Values of reaction rate constants were obtained as described earlier (Zaremba et al., 2006).

### 2.2.7. RNA strand annealing activity assay

The annealing activity of Cas9 was assayed in 10 µl reaction volumes. Reactions were initiated by mixing 5 nM pre-crRNA (94 nt) and varying concentrations of Cas9 with 5 nM tracrRNA (105 nt) radiolabeled with [$\gamma$-$^{32}$P]ATP (Hartmann Analytic) and 1 µg nonspecific 2 kb RNA transcript in complex assembly buffer and conducted for 10 min at 37°C. In order to inactivate Cas9, disrupt protein-nucleic acid complexes and prevent spontaneous pre-crRNA annealing to labeled tracrRNA, reactions were terminated by mixing 5 µl reaction aliquots with 20 µl of solution containing 0.625 mg/ml Proteinase K (Thermo Fisher Scientific), 0.625% SDS, 12.5% glycerol and RNA trap (187.5 nM solution of unlabeled tracrRNA). Reaction mixtures were incubated for additional 5 min at 37°C and analyzed in non-denaturing 8% PAGE. Gels were dried and visualized by a phosphorimager (FLA-5100, Fujifilm).

### 2.2.8. Genomic DNA cleavage *in vitro*

#### 2.2.8.1. λ DNA cleavage

The reactions were initiated by mixing λ DNA (Thermo Fisher Scientific) with assembled Cas9 complex (1:1 v/v ratio) and incubating at 37°C. Final reaction mixture contained 2 µg λ DNA and 50 nM Cas9 complex in 100 µl reaction buffer. Aliquots were removed at timed intervals and quenched with phenol/chloroform. The aqueous phase was mixed with 3× loading dye

solution and reaction products analyzed by electrophoresis through agarose gel and ethidium bromide staining.

#### 2.2.8.2. *E. coli* genomic DNA cleavage

Genomic DNA from *E. coli* (BL21) strain was isolated using the "GeneJET Genomic DNA Purification Kit" (Thermo Fisher Scientific). For the cleavage assay, genomic DNA was combined with assembled Cas9 complex (1:1 v/v ratio) and incubated for 3 hours at 37°C. Final reaction mixture contained 30 µg genomic DNA and 1 µM Cas9 in a 300 µl reaction buffer. Following incubation, PstI (Thermo Fisher Scientific) was added and the reaction mix was incubated for additional 16 hours at 37°C. The reaction was terminated by heating the mixture for 30 min at 55°C with Proteinase K (0.5 mg/ml, Thermo Fisher Scientific) and SDS (0.5%, w/v) following 30 min incubation at room temperature with RNase A (0.25 mg/ml, Thermo Fisher Scientific). After phenol/chloroform treatment, DNA was precipitated by isopropanol and dissolved in TE buffer. DNA cleavage reaction products were analyzed using Southern blot hybridization (see section 2.2.9).

#### 2.2.8.3. Human genomic DNA cleavage

Human genomic DNA extracted from human brain was kindly provided by Dr. Arturas Petronis (Centre for Addiction and Mental Health, Toronto, Canada). Cas9-HS1 and Cas9-HS2 complexes were assembled to target RASGEF1C or ARL15 loci, respectively. DNA was combined with assembled Cas9 complexes and incubated for 30 min at 37°C. Final reaction mixture contained 1 µg of genomic DNA and 100 nM of each Cas9 complex in 100 µl reaction buffer. DNA cleavage reaction products were analyzed using qPCR assay (see section 2.2.10).

### 2.2.9. Southern blot hybridization

Southern blot analysis was performed as described in (Sambrook et al., 1989) with the following modifications. 10 μg of fractionated DNA was transferred from 1% agarose gel onto "SensiBlot Plus Nylon membrane" (Thermo Fisher Scientific) via semi-dry transfer. DNA was denatured and fixed on the membrane by placing it on a paper towel saturated with 0.4 M NaOH for 10 min following rinsing with 2× SSC and air drying. The membrane was prehybridized with 6× SSC buffer containing 0.5% SDS and 100 μg/ml denatured salmon sperm DNA (Amresco) for 1 h at 65°C. The hybridization probe was generated by PCR using genomic *E. coli* BL21 (DE3) DNA as template and primers TK-36/TK-34 (Table 5) yielding a 397 bp product. 5'-ends were radiolabeled by incubating with [γ-$^{32}$P]ATP (Hartmann Analytic) and T4 PNK (Thermo Fisher Scientific). The labeled probe was purified using "GeneJET PCR Purification Kit" (Thermo Fisher Scientific), denatured by heating to 95°C for 5 min, rapidly cooled on ice and added directly to the prehybridization solution. The membrane was probed for 16 hours at 65°C and washed twice with 2× SSC, 0.5% SDS and twice with 2× SSC, 0.1% SDS at room temperature, air dried and visualized by phosphorimaging (FLA-5100, Fujifilm).

### 2.2.10. DNA quantification using qPCR assay

DNA products after cleavage reactions were analyzed using qPCR assay using primers listed in Table 5. 30 ng of DNA was used as a template for qPCR reaction (25 μl) with SYBR Green reagent (Maxima SYBR Green Master Mix 2×, Thermo Fisher Scientific). Relative quantification of DNA was performed using $2^{-\Delta\Delta Ct}$ method (Livak and Schmittgen, 2001).

### 2.2.11. Cell culture and transfection

One day before the experiment, the CHO-K1 or HEK293T cells were seeded in a 24-well plate at the density of $7\times10^4$ cells per 1 ml of RPMI-1640 culture medium supplemented with 10% fetal bovine serum (FBS). For protein transfection, cells were transferred into serum-free medium just before the experiment. For DNA transfection, plasmid DNA was mixed with TurboFect *in vitro* DNA transfection agent (Thermo Fisher Scientific) following manufacturer's protocol. The transfection efficiencies were analyzed using "Guava EasyCyte 8HT" flow cytometer and "Guava CytoSoft 5.3" cell acquisition/analysis software (Millipore).

For DNA/protein co-transfection experiments the dual reporter plasmid DNA (0.5 µg) and Cas9 complex (1 µg) plus extra 300 ng of crRNA:tracrRNA duplex (to ensure more efficient complex formation with cationic polymer TurboFect) were diluted in separate tubes with serum-free medium to a final volume of 50 µl. Next, 0.5 µl of TurboFect transfection agent was added into each tube, samples were incubated for 15-20 min at room temperature and added to the cell cultures at the same time. The serum-free cell culture medium was changed into the complete growth medium 3 h later and cells were incubated for an additional 48 h.

For protein transfections, 1 µg of the recombinant Cas9 or Cas9 complex plus excess of crRNA:tracrRNA duplex (300 ng) were diluted with serum-free medium to the final volume of 100 µl. The samples were mixed with 1µl of TurboFect protein transfection agent, incubated for 15 min and added to the cell culture. Three hours after transfection the serum-free cell culture medium was changed into the full growth medium. Cells were analyzed 48 h later.

For multiplex gene editing of DNMT3B and PPIB genes, 1 µg of each corresponding Cas9 complexes plus 300 ng of each RNA duplex were diluted with PBS to a final volume of 100 µl, mixed with 2.5 µl of TurboFect and following 15 min of incubation added to the cells. Serum-free medium was

replaced with complete growth medium 16h after transfection and the cells were incubated for another 48h.

The stable cell line expressing a dual reporter gene cassette was generated using "cGPS® CHO-K1 Full Kit DD" (Cellectis Bioresearch). TurboFect *in vitro* transfection reagent was used for plasmid delivery into the CHO-K1 cells.

### 2.2.12. INDEL analysis

DNA was extracted from CHO-K1 or HEK293T cells using "GeneJET genomic DNA purification kit" (Thermo Fisher Scientific). Regions surrounding Cas9 target sites were PCR amplified using Phusion Hot Start II DNA polymerase (Thermo Fisher Scientific). Corresponding primer sequences are provided in Table 5. For Surveyor assay, PCR products were reannealed and analyzed using "SURVEYOR Mutation Detection Kit for Standard Gel Electrophoresis" (Transgenomic). Amplicon sizes are provided in Table 5.

For Cas9 complex digestion assay, 3 nM of PCR products were mixed with 50 nM of Cas9 complex and incubated for 20 min at 37°C in a reaction buffer. Reaction was quenched with phenol/chloroform and digestion products were examined in 2% agarose gel. Quantitative analysis was done using ImageJ program (National Institutes of Health) as described in (Cong et al., 2013).

### 2.2.13. Single-molecule experiments

#### 2.2.13.1. Construction of DNA substrates for magnetic tweezers experiments

All DNA constructs were based on pUC18 plasmids into which single protospacer/PAM elements were inserted via the EcoRI sites (see Tables 3 and 5 for the details). For preparing constructs for the tweezers experiments, a 2.1 kbp fragment containing a single protospacer/PAM combination was made by PCR from the recombinant plasmids (see Table 5 for primer sequences),

digested with NotI and SpeI and purified from the agarose gel. Biotin- or digoxigenin-modified attachment handles were made using 1.0 kbp DNA fragments that where labeled with biotin- or digoxigenin-dUTP by PCR (see Table 5 for primer sequences), and which were digested with either NotI or SpeI. The protospacer fragment was ligated with the biotin/digoxigenin-labeled handles using T4 DNA ligase.

#### 2.2.13.2. Single-molecule observation of R-loop formation

Single-molecule magnetic tweezers experiments with Cas9 were carried out as previously described (Revyakin et al., 2005; Seidel et al., 2005) using Picotwist instrument (Saint Romain de Popey, France, equipped with a Jai CV-A10 GE camera, image acquisition at 60 Hz). The fluidic cells for the Cas9 experiments were constructed from an uncoated 24x60 mm coverslip (Menzel-Gläser No. 1), double-sided adhesive tape (3M 467MP, 50 µm depth) and polyester film (Melinex 401, DuPont, 50 µm depth). Anti-digoxigenin (Roche) and BSA were adsorbed directly to the glass by incubation for >3 hours at room temperature. Each DNA construct was bound at its biotin-modified end to excess streptavidin-coated magnetic beads (1 µm diameter, MyOne, Invitrogen) and added into the fluidic cell to allow the DNA to bind the surface via its digoxigenin-modified end. Non-magnetic particles (3.2 µm tosyl activated polystyrene or 2.0 µm aldehyde/sulphate latex, Invitrogen) were adhered to the glass (in phosphate buffered saline or 50 mM MES, pH 5.5, respectively) to correct for instrument drift. The three-dimensional position of the magnetic bead and thus the orientation and length of the attached DNA molecule was determined from video images at the camera frame rate (see above) using real-time 3D particle tracking with sub-nm accuracy (Klaue and Seidel, 2009; Lionnet et al., 2012; Otto et al., 2010). Suitable topologically-constrained DNA were identified from rotations curves and the rotational zero reference set as determined from a rotation curve at 0.3 pN. Experiments with Cas9 were carried out in complex assembly buffer. Measurements were

performed using 1 nM Cas9 (at 25°C). When recording the shift in rotational zero due to R-loop formation magnets were rotated at 0.5 – 1.0 Hz. For measuring the on/off times as a function of torque, the magnets were turned at 10 Hz. In all time trajectories and rotation curves depicted, raw DNA length data taken at the camera acquisition rate is shown in light colours (light grey, blue or red), while data smoothed with either a 1 or 2 Hz moving average is shown as in dark colours (dark, blue, green, grey and red).

#### 2.2.13.3. Determination of rotational shifts upon R-loop binding

On the majority of Cas9 DNA substrates, R-loop formation and dissociation events occurred in the postbuckling/constant torque regime (see Figure 16D). The position of the rotational zero could therefore be estimated before and after R-loop formation by fitting the peak of the induction and probe rotation curves to a parabolic function.

### 2.2.14. Rapid characterization of Cas9 PAM sequence elements

#### 2.2.14.1. Construction of a 5 bp randomized PAM library

Construction of the 5 bp randomized PAM plasmid DNA library was initiated with the synthesis of a single oligonucleotide, GG-821N (Table 5), with hand-mixing used to create a random incorporation of nucleotides across the 5 random residues (represented as N in the sequence of GG-821N). To convert the single-stranded template of GG-821N into a double-stranded DNA template for cloning into the plasmid vector, a second oligonucleotide, GG-820 (Table 5), was synthesized with complementation to the 3’ end of GG-821N to form a partial oligonucleotide duplex. The partial duplex was then extended by PCR using DreamTaq polymerase (Thermo Fisher Scientific) to generate a full duplex containing the target sequence, 5 randomized base pairs downstream of the target sequence and cleavage site for the BamHI restriction enzyme. To generate the plasmid library, the oligoduplex, purified using “GeneJET PCR

Purification Kit" (Thermo Fisher Scientific), was digested with BamHI and ligated into pTZ57R/T vector (Thermo Fisher Scientific) pre-cleaved with BamHI. Linear pTZ57R/T vector contains protruding ddT nucleotide at the 3' ends, whereas PCR fragments generated with DreamTaq polymerase contains dA at the 3' ends. Therefore one end of the PCR fragment is ligated into the vector through BamHI sticky ends, while another through A/T ends. DH5α $Ca^{2+}$ competent cells were transformed with the ligated plasmid library and plated onto LB containing agar. The transformation efficiency was estimated from plated dilutions. Overall, ~12,000 colonies were recovered. The colonies were harvested from the plate by gently resuspending them in liquid LB media and plasmid DNA was purified using "GeneJET Plasmid Miniprep Kit" (Thermo Fisher Scientific).

### 2.2.14.2. Construction of a 7 bp randomized PAM library

The 7 bp randomized PAM plasmid DNA library was constructed as described for the 5 bp library with the following modifications. Randomization of the PAM sequence was generated through the synthesis of four oligonucleotides, GG-940-G, GG-940-C, GG-940-A and GG-940-T (Table 5), with hand-mixing used to create a random incorporation of nucleotides across the random residues (represented as N). The randomized single-stranded oligonucleotides were each separately converted into double-stranded DNA templates for cloning into the plasmid vector using a second oligonucleotide, GG-939 (Table5), with complementation to the 3' end of GG-940-G, GG-940-C, GG-940-A and GG-940-T and by PCR extension with DreamTaq polymerase (Thermo Fisher Scientific). To avoid cleavage of some species of the randomized positions, the resulting double-stranded templates were each digested with an 8 bp cutting restriction endonuclease, SdaI, so that overhangs were present at each end; a PstI compatible overhang and a Taq added single 3' A overhang. The resulting overhangs were used to directionally ligate the 4 double-stranded templates into pTZ57R/T (Thermo Fisher Scientific) pre-

cleaved with PstI. The ligations were transformed into DH5α $Ca^{2+}$ competent cells, plasmid DNA was recovered and combined from each of the 4 transformants derived from GG-940-G, GG-940-C, GG-940-A and GG-940-T to generate the randomized 7 bp PAM plasmid DNA library.

### 2.2.14.3. PAM libraries validation

To validate the randomness of the resulting PAM libraries, PCR fragments spanning the 5 bp and 7 bp randomized PAM regions were generated by Phusion High-Fidelity DNA Polymerase (Thermo Fisher Scientific) amplification (15 cycles of a 2-step amplification protocol) using the primer pair combinations TK-119/pUC-dir and TK-113/pUC-dir (Table 5) for the 5 bp and 7 bp libraries, respectively. The resulting 122 bp (for 5 bp PAM library) and 145 bp (for 7 bp PAM library) PCR products were purified using "GeneJET PCR Purification Kit" (Thermo Fisher Scientific) and the sequences necessary for amplicon-specific barcodes and Illumina sequencing were "tailed" on through two rounds of PCR each consisting of 10 cycles. The primer pair combinations in the first round of PCR were JKYS800.1/JKYS803 and JKYS921.1/JKYS812 (Table 5) for the 5 bp and 7 bp libraries, respectively. A set of primers, JKYS557/JKYS558 (Table 5), universal to all primary PCR reactions was utilized for the secondary PCR amplification. The resulting PCR amplifications were purified with a Qiagen PCR purification spin column, concentration measured with a Hoechst dye-based fluorometric assay, combined in an equimolar ratio and single read 60-100 nucleotide-length deep sequencing was performed on Illumina's MiSeq Personal Sequencer with a 5-10% (v/v) spike of PhiX control v3 (Illumina, FC-110-3001) to off-set sequence bias. After sequencing, reads were trimmed to a minimum Phred quality (Q score) of 13 and different treatments were deconvoluted by identifying a perfectly matching 4-6 nt barcode sequence present at the 5 prime end. The PAM sequence for only those reads containing a perfect 12 nt sequence match flanking either side of the randomized PAM

sequence were captured. The collection of resulting PAM sequences were then collapsed into like sequences, counted and frequency of each PAM calculated. A position frequency matrix (PFM) was then performed by first aligning the collapsed PAM sequences. Next, each nucleotide (G, C, A or T) at each position of the PAM was weighted based on the frequency of the PAM sequence with which it was associated. Finally, the total contribution of each nucleotide (G, C, A or T) at each PAM position was summed to generate the overall probability of identifying a given nucleotide at each PAM position within the dataset.

#### 2.2.14.4. Capture and identification of PAM preferences

To identify PAM preferences, 1 µg (5.6 nM) of plasmid DNA library with randomized PAM was cleaved with 0.5 nM and 50 nM of Cas9-guide RNA complex in a reaction buffer for 60 min at 37°C in a 100 µl reaction volume. To efficiently capture the blunt-ends of the plasmid library generated by Cas9 complex cleavage, a 3' dA was added by incubating the completed digestion reactions with 2.5 U of DreamTaq DNA Polymerase (Thermo Fisher Scientific) and 0.5 µl of 10 mM dATP (or dNTP) for an additional 30 min at 72°C. Reaction products were purified using "GeneJET PCR Purification Kit" (Thermo Fisher Scientific). Next adapters with a 3' dT overhang were generated by annealing TK-117 and phosphorylated TK-111 oligonucleotides (Table 5). 100 ng of the resulting adapter was ligated to an equal concentration of the purified 3' dA overhanging cleavage products for 1 hour at 22°C in a 25 µl reaction volume in adapter ligation buffer with 0.5 U T4 Ligase (Thermo Fisher Scientific). Next, to selectively enrich for cleaved products containing the PAM sequence, PCR amplification was performed with a forward primer, pUC-dir specific to the PAM-side of the cleaved pTZ57R/T plasmid vector and with a reverse primer, TK-117 specific to the ligated TK-117/TK-111 adapter sequence. PCR fragments were generated by Phusion High-Fidelity DNA Polymerase (Thermo Fisher Scientific) amplification (15 cycles of a 2-step

amplification protocol) with 10 µl of ligation reaction mixtures as a template (in 100 µl total volume). The resulting PCR products amplified from the Cas9-guide RNA complex cleaved plasmid libraries were purified with "GeneJET PCR Purification Kit" (Thermo Fisher Scientific) and prepared for Illumina deep sequencing as described in the PAM library validation section except the barcode containing forward primers used in the primary reaction were specific to the TK-117/TK-111 adapter sequence (Table 5). Illumina deep sequencing, post-processing and position frequency matrices (PFMs) were performed as described in the PAM library validation section (see 2.2.14.3). WebLogos were generated as described by (Crooks, 2004).

### 2.2.15. *In planta* mutation detection

The DNA region surrounding the expected site of cleavage for each Cas9-guide RNA was amplified by PCR using Phusion® High Fidelity PCR Master Mix (NEB) "tailing" on the sequences necessary for amplicon-specific barcodes and Illumina sequences through two rounds of PCR each consisting of 20 cycles. The primer pairs used in the primary PCR were JKYX1.1/JKYS178Rd, JKYS1083.1/JKYS1084 and JKYX2.1/JKYX3 (Table 5) each corresponding to *Ms45* exon 1, *Ms45* exon 4 and *liguleless-1* targets, respectively. A set of primers universal to the products from the primary reactions, JKYS557/JKYS558, were used in the secondary PCR reaction. The resulting PCR amplifications were purified using "QIAquick PCR Purification Kit" (Qiagen), concentration measured with a Hoechst dye-based fluorometric assay, combined in an equimolar ratio, and single read 100 nucleotide-length amplicon sequencing was performed on Illumina's MiSeq Personal Sequencer with a 5-10% (v/v) spike of PhiX control v3 (Illumina, FC-110-3001) to off-set sequence bias. Post-processing on the resulting sequences was performed as described in the PAM library validation section and only those reads with a ≥1 nucleotide INDEL arising within the 10 nt window centered over the expected site of cleavage and not found in the negative controls were classified as

mutations. Mutant reads with an identical mutation were counted and collapsed into a single read and the top 10 most prevalent mutations were visually confirmed as arising within the expected site of cleavage. The total numbers of visually confirmed mutations were then used to calculate the percentage of mutant reads based on the total number of reads of an appropriate length containing a perfect match to the barcode and forward primer.

# 3. RESULTS AND DISCUSSION

## 3.1. crRNA and tracrRNA guide Cas9-mediated DNA interference in *Streptococcus thermophilus*

*S. thermophilus* DGCC7710 strain contains four distinct CRISPR-Cas systems: CRISPR1, CRISPR2, CRISPR3 and CRISPR4 (Horvath and Barrangou, 2010). The CRISPR2 and CRISPR4 systems belong to the Type III and Type I, respectively, while CRISPR1 and CRISPR3 both belong to Type II CRISPR-Cas systems (Horvath and Barrangou, 2010; Makarova et al., 2015). Type II CRISPR-Cas systems dramatically differ from other types of CRISPR-Cas systems and are one of the simplest systems in terms of number of genes. All CRISPR-Cas systems in subtype II-A share a similar architecture, with *tracrRNA* and four *cas* genes located in the vicinity of the CRISPR spacer-repeat array (Figure 9A). Type II signature gene *cas9* encodes a sole multidomain protein, responsible for target DNA cleavage (Gasiunas et al., 2012; Jinek et al., 2012; Sapranauskas et al., 2011). In addition, it has recently been shown *cas9* involvement in adaptation step with *cas1*, *cas2* and *csn2* genes (Heler et al., 2015; Wei et al., 2015).

The effector complex of the Type II-A *S. thermophilus* CRISPR3-Cas system contains a ~42 nt crRNA combined with Cas9 (Gasiunas et al., 2012). While tracrRNA involvement in the *S. pyogenes* effector complex has been documented (Jinek et al., 2012), the role of tracrRNA in DNA silencing provided by the *S. thermophilus* CRISPR3-Cas effector complex remains to be established.

### 3.1.1. Genetic location of the tracrRNA-encoding sequence in the *S. thermophilus* DGCC7710 CRISPR3-Cas system

Sequence analysis of the DNA region upstream of the *cas9* gene in the *S. thermophilus* DGCC7710 CRISPR3-Cas system revealed that, similarly to the

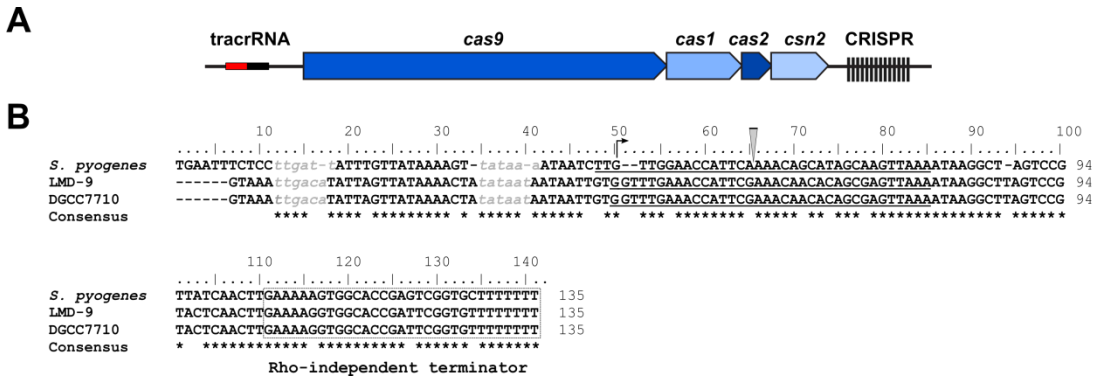


**Figure 9. Schematic representation of Type II-A CRISPR-Cas systems in *S. pyogenes*, *S. thermophilus* LMD-9 and DGCC7710 strains.** (**A**) Four *cas* genes are located in the vicinity of the CRISPR spacer-repeat array. tracrRNA is encoded upstream of the *cas* operon. (**B**) Alignment of tracrRNA encoding loci in *S. pyogenes*, *S. thermophilus* LMD-9 and DGCC7710 CRISPR-Cas systems. The anti-repeat regions complementary to the repeat sequences in crRNA are underlined, the putative promoters are shown in grey and the Rho independent terminators are boxed. The RNase III cleavage site and transcription site in *S. pyogenes* is indicated by a gray triangle and arrow, respectively.

*S. pyogenes* CRISPR-Cas system (Deltcheva et al., 2011), a putative tracrRNA is encoded upstream of the *cas* operon (Figure 9A). In *S. pyogenes*, deep sequencing data (Deltcheva et al., 2011) revealed 171 nt and 89 nt tracrRNAs that would result from transcription at two distinct promoters to a shared transcriptional terminator. Northern blot analysis of the *S. thermophilus* LMD-9 strain, which contains a CRISPR3-Cas system nearly identical to the *S. thermophilus* DGCC7710 CRISPR3-Cas system (Gasiunas et al., 2012), revealed a ~100 nt tracrRNA transcript (Deltcheva et al., 2011).

To identify a putative transcriptional start position for the tracrRNA transcript from *S. thermophilus* DGCC7710 CRISPR3-Cas system, we compared DNA fragments containing anti-repeat region of the tracrRNA (Figure 9B). The tracrRNA encoding sequences are very similar (87% identical nucleotides) and contain Rho-independent transcription terminators. Therefore, we assumed that the transcription of *S. thermophilus* DGCC7710 tracrRNA could start at the same position as in *S. pyogenes* and the estimated size of *S. thermophilus* DGCC7710 tracrRNA is ~100 nt, which would be consistent with the size of *S. thermophilus* LMD-9 tracrRNA (Deltcheva et al., 2011).

### 3.1.2. tracrRNA is necessary for *in vivo* DNA interference by *S. thermophilus* CRISPR3-Cas

The pCRISPR3 plasmid, which carries a complete *S. thermophilus* DGCC7710 CRISPR3-Cas locus including a CRISPR3 array comprised of 12 spacer-repeat units and a tracrRNA encoding fragment located upstream of the *cas9* gene (Figure 10A), provides interference against transformation of a donor pS1 plasmid which contains a protospacer identical to the S1 spacer in the CRISPR3 array and the accompanying 5'-NGGNG-3' PAM sequence (Gasiunas et al., 2012; Sapranauskas et al., 2011). To establish whether tracrRNA is required for *S. thermophilus* CRISPR3-Cas-mediated plasmid interference in *E. coli*, we generated pCRISPR3 plasmid variants with a compromised tracrRNA-encoding sequence (Figure 10A). In the pCRISPR3-Δt plasmid, the entire tracrRNA coding sequence is deleted, while in the pCRISPR3-ΔtR variant, the tracrRNA-encoding sequence is truncated at the 3' end, to leave only the region from the transcription start site to the end of anti-repeat sequence (Figure 10A). Next, we analyzed transformation efficiency of two recipient *E. coli* strains carrying tracrRNA-deficient pCRISPR3-Δt or pCRISPR3-ΔtR variants by the pS1 donor plasmid. The pUC18 plasmid, which lacks a protospacer but contains multiple PAMs was used as a control in the plasmid transformation assay. We found that, in contrast to pCRISPR3 carrying recipient strain, which was resistant to pS1 transformation, the tracrRNA-deficient strains became permissive for transformation by pS1 plasmid (Figure 10B). Thus, the plasmid immunity provided by the heterologous *S. thermophilus* CRISPR3-Cas system is compromised when either part of or the full tracrRNA encoding sequence is eliminated, indicating that tracrRNA is necessary for Cas9-mediated interference.

To confirm that the *S. thermophilus* CRISPR3-Cas mediated plasmid interference is lost due to the tracrRNA gene deletion/truncation, we introduced a full length tracrRNA *in trans* into the tracrRNA-deficient permissive recipient strain and evaluated pS1 plasmid transformation

efficiency. More specifically, the fragment encoding tracrRNA was cloned into the pCDF-DUET plasmid under T7 promoter control, the construct expressed in the pS1 permissive *E. coli* BL21 (DE3) strain carrying the pCRISPR3-Δt plasmid (Figure 10C), and transformation efficiency was evaluated by counting colonies on Ap, Str, Cm and IPTG-supplemented agar plates. Under these conditions, we found that transformation by the pUC18 plasmid yielded ~1,000 colonies, while no colonies were obtained in the case of pS1 plasmid (Figure 10C). The same results were obtained in a recipient host, which carried the

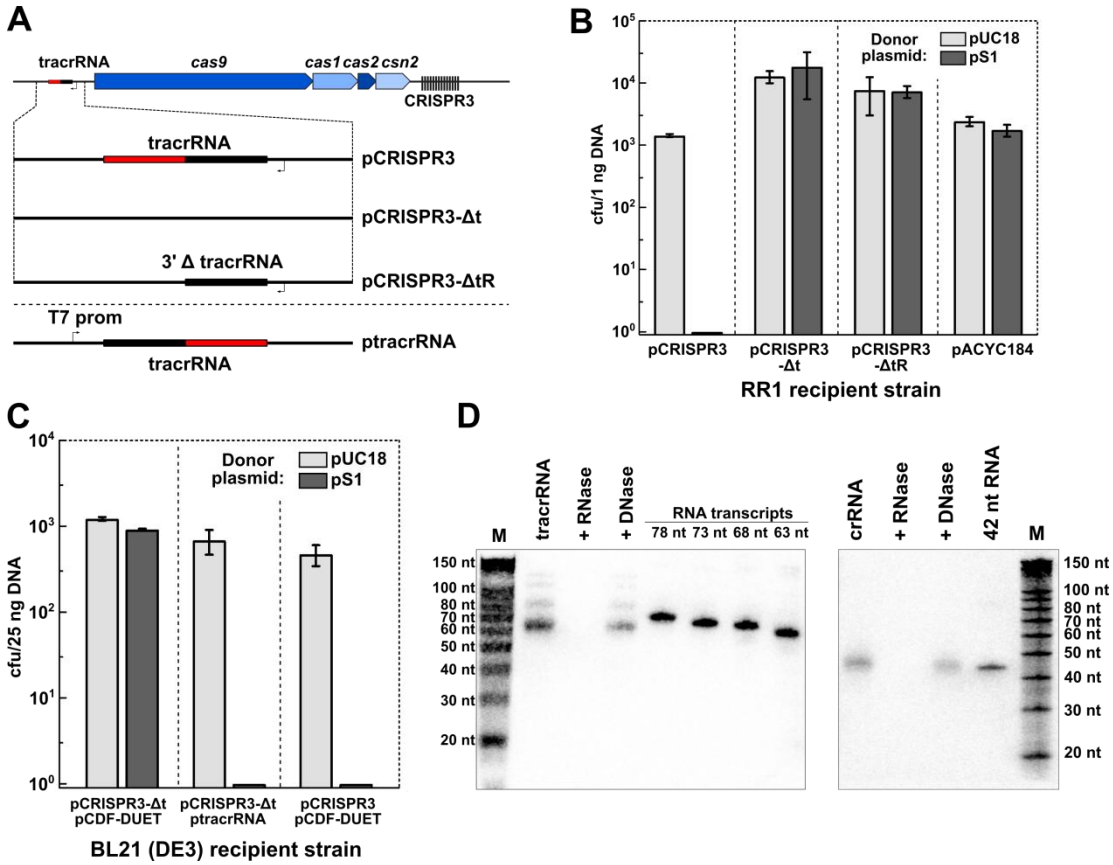


**Figure 10. The tracrRNA is required for interference. (A**) Schematic representation of plasmids used for plasmid transformation interference assays. The pCRISPR3-Δt plasmid encodes a CRISPR3-Cas system without tracrRNA. In pCRISPR3-ΔtR, tracrRNA contains only the anti-repeat region and lacks 3' end. ptracrRNA plasmid was obtained by inserting a full length tracrRNA encoding sequence under control of T7 RNA polymerase promoter in the pCDF-DUET plasmid. (**B**) The deletion or shortening of tracrRNA inactivates CRISPR3-Cas interference. (**C**) tracrRNA can be provided *in trans* on a separate plasmid. (**D**) Cas9 co-purifies with ~65 nt tracrRNA and ~42 nt crRNA. Northern blot analysis of nucleic acids extracted from purified active Sth3 Cas9 complex using anti-tracrRNA (left panel) and anti-crRNA (right panel) oligonucleotide probes. The estimated size of the tracrRNA is ~65 nt, minor amounts of longer tracrRNA intermediates are present. M – RNA size markers.

pCRISPR3 plasmid containing tracrRNA *in cis* (Figure 10C). These results show that trans-complementation of tracrRNA converts a pS1-permissive *E. coli* strain into a transformation resistant (non-permissive) strain. Taken together, plasmid transformation assays demonstrate that tracrRNA is necessary for DNA interference provided by the *S. thermophilus* CRISPR3-Cas effector complex, *in vivo*.

### 3.1.3. tracrRNA co-purifies with Cas9 protein

*S. thermophilus* CRISPRR3 (Sth3) Cas9 is the sole Cas protein required for the *S. thermophilus* CRISPR3-Cas mediated immunity (Gasiunas et al., 2012; Sapranauskas et al., 2011). Sth3 Cas9 protein co-purifies with a ~42 nt crRNA (Figure 10D) (Gasiunas et al., 2012). To probe whether tracrRNA also co-purifies with Cas9 and crRNA, we performed Northern blot analysis using an anti-tracrRNA 36 nt oligodeoxynucleotide probe. Nucleic acids extracted from the Strep-Tactin–purified Cas9 preparation (Gasiunas et al., 2012) hybridized with the anti-tracrRNA probe and were sensitive to RNase, but not to DNase treatment (Figure 10D). The estimated size of the tracrRNA which co-purified with Cas9 is ~65 nt. Hence, the *S. thermophilus* DGCC7710 CRISPR3-Cas effector complex that provides interference against DNA consists of a ternary Cas9-crRNA-tracrRNA complex similarly to the *S. pyogenes* (Spy) effector complex (Jinek et al., 2012).

### 3.1.4. *In vitro* reconstitution of the Sth3 Cas9-crRNA-tracrRNA effector complex

Analysis of the protein and nucleic acid content of the Sth3 effector complex isolated from the heterologous *E. coli* strain revealed the presence of Cas9, a ~42 nt crRNA and a ~65 nt tracrRNA (Figure 10D). Next, we aimed to reconstitute an effector complex *in vitro*, by combining these three individual components. To assemble Cas9-crRNA-tracrRNA complex, Cas9 was pre-incubated with equimolar amounts of a synthetic 42 nt crRNA and 78 nt

tracrRNA, corresponding to a mature form, obtained by *in vitro* transcription. Extra nucleotides (78 nt vs. 65 nt) were introduced for *in vitro* transcription by a T7 RNA polymerase. The DNA cleavage activity of the reconstituted complex was monitored *in vitro* using the pS1 plasmid (see section 2.2.6). In the presence of a reconstituted Sth3 Cas9 complex, the pS1 plasmid which contains a protospacer 1 sequence flanked by the 5'-NGGNG-3' PAM is converted into a linear form (Figure 11A), indicating that both DNA strands are cleaved within the protospacer region. If one of the complex components (Cas9, crRNA or tracrRNA) is missing, no cleavage of pS1 is observed, indicating that all three components are necessary to form a functional Sth3 Cas9 complex. On the other hand, the pUC18 plasmid which lacks the protospacer is not cleaved by the Sth3 Cas9. Hence, a catalytically competent Cas9-crRNA-tracrRNA complex can be reconstituted *in vitro* by mixing individual components, similarly to the Cas9 complex of *S. pyogenes* (Spy) (Jinek et al., 2012).

Experiments with Spy Cas9 complex revealed that tracrRNA can be 3'-truncated in *in vitro* applications (Jinek et al., 2012). To define the minimal length of 3'-end of tracrRNA required for crRNA-guided DNA cleavage by the Sth3 Cas9, we designed and generated a set of tracrRNA molecules truncated by stretches of 5 nt from the 3' terminus. The overall size of tracrRNA variants varied between 33 nt and 78 nt (Figure 11B). The truncated tracrRNA variants were used for *in vitro* reconstitution of the Sth3 Cas9 complex, followed by analysis of the pS1 plasmid cleavage. We found that 38 nt and 33 nt tracrRNA variants do not support DNA cleavage by Sth3 Cas9 (Figure 11C), whereas 43, 48 and 53 nt variants showed decreased cleavage activity. On the other hand, tracrRNAs longer than 58 nt support efficient Sth3 Cas9 cleavage of the pS1 plasmid. Interestingly, secondary structure analysis of tracrRNA suggests formation of three putative hairpin structures in the non-complementary

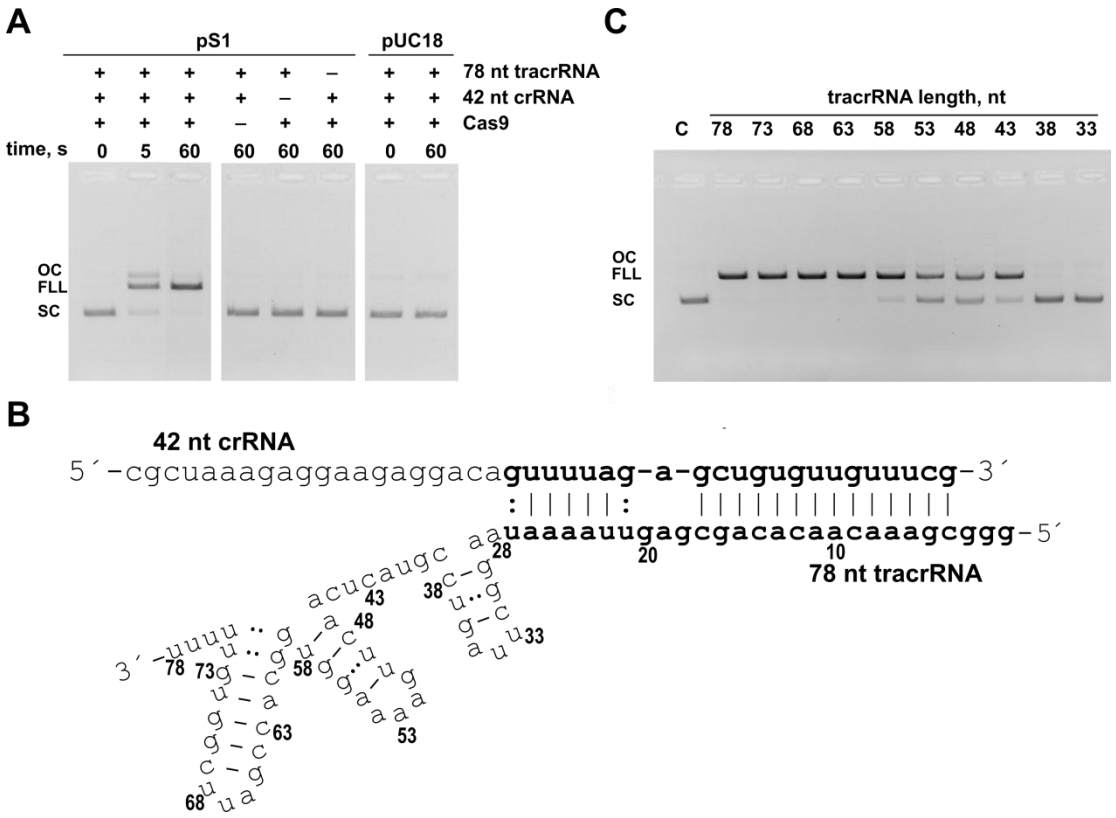


**Figure 11. Plasmid DNA cleavage by the *in vitro* assembled effector Sth3 Cas9 complex.** (**A**) Cas9, 42 nt crRNA and 78 nt tracrRNA were used for complex assembly. The Cas9 RNP complex was incubated with pS1 or pUC18 plasmids. pS1 plasmid contained a protospacer 1 sequence flanked by the 5'-NGGNG-3'PAM sequence. Protospacer 1 sequence is absent in pUC18. Reaction products were analyzed by electrophoresis in agarose gel. SC – supercoiled plasmid, OC – nicked reaction intermediate cut at either DNA strand, FLL – final reaction product cleaved at both DNA strands. Under these reaction conditions the pS1 plasmid is converted into a linear form while the pUC18 plasmid lacking a protospacer 1 sequence is resistant to cleavage. In the reaction mixes lacking one of the components (Cas9, crRNA or tracrRNA) pS1 plasmid is not cleaved. (**B**) Schematic representation of 42 nt crRNA:78 nt tracrRNA duplex structure. The putative secondary structures of the non-complementary part of the tracrRNA identified by RNAfold (Hofacker et al., 1994) are shown. The positions of 5 nt 3'-truncations in the non-complementary region are indicated. (**C**) pS1 plasmid cleavage by the effector complex assembled *in vitro* using tracrRNAs of different lengths.

3'-terminus of tracrRNA (Figure 11B). Our data indicate that the 3'-end proximal hairpin is not necessary for effector complex formation, while two others are required for Cas9 binding and/or cleavage.

### 3.1.5. Cas9 role in formation of the pre-crRNA:tracrRNA duplex

In the Sth3 and Spy Cas9 effector complexes, the target site recognition is achieved by the crRNA while the Cas9 protein provides two active sites for the

cleavage of opposite DNA strands in the protospacer (Gasiunas et al., 2012; Jinek et al., 2012). Indirect evidence obtained *in vivo* in *S. pyogenes* suggests that Cas9 is also an essential protein for crRNA maturation (Deltcheva et al., 2011) and may facilitate formation and stabilization of the pre-crRNA:tracrRNA duplex (Deltcheva et al., 2011), however experimental evidence and molecular details remain to be established. To probe experimentally whether the Sth3 Cas9 promotes pre-crRNA and tracrRNA annealing, we produced tracrRNA3 (105 nt) and pre-crRNA3 (94 nt containing a single 36 nt repeat sequence) from CRISPR3 by *in vitro* transcription and monitored tracrRNA3:pre-crRNA3 duplex formation in the presence of Sth3 Cas9 (Figure 12). In a control set of experiments, tracrRNA1 (105 nt) and pre-crRNA1 (94 nt) from the homologous *S. thermophilus* DGCC7710 Type II CRISPR1 system were used. To monitor duplex formation, radioactively labeled tracrRNAs were incubated for 10 min with pre-crRNAs in the presence of a non-specific 2 kb RNA transcript and different amounts of the Sth3 Cas9, the reaction was quenched by adding Proteinase K and an excess of unlabeled tracrRNA and samples were analyzed on the non-denaturing PAGE (Figure 12). Under these experimental conditions, in the absence of Cas9, no

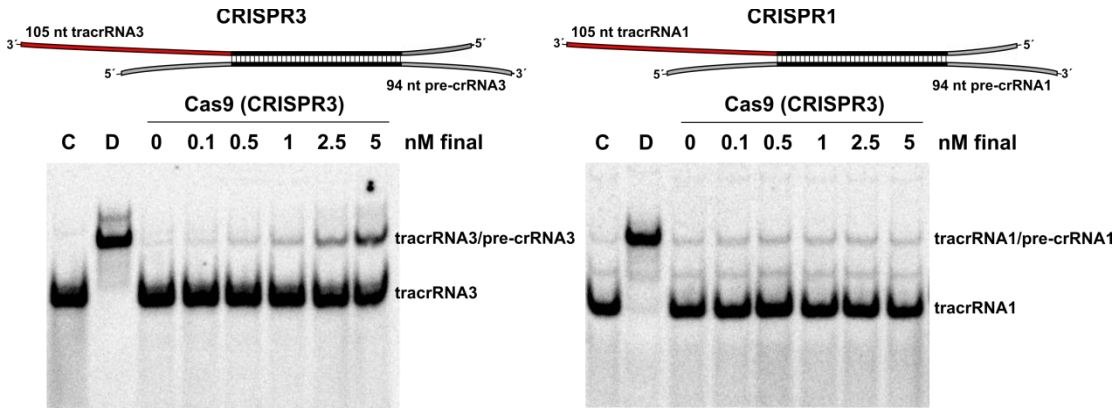


**Figure 12. tracrRNA and pre-crRNA annealing in the presence of Cas9.** 105 nt tracrRNA and 94 nt pre-crRNA used in the annealing assay are represented above the gels. Complementary sequences are indicated in black. Non-denaturing PAGE analysis of duplex assembly between tracrRNA3 and pre-crRNA3 of CRISPR3 system (left panel) and tracrRNA1 and pre-crRNA1 from CRISPR1 (right panel) at varying concentrations of Cas9 from CRISPR3. C – control lanes, containing only labeled tracrRNA, D – tracrRNA:pre-crRNA duplex, formed by heating tracrRNA:pre-crRNA mixture (1:100 molar ratio) to 95°C and slowly cooling to room temperature.

tracrRNA:pre-crRNA duplex is formed. Increasing amounts of Sth3 Cas9 promote tracrRNA3:pre-crRNA3 formation but have no effect with tracrRNA1 and pre-crRNA1 (Figure 12). Taken together these results indicate that Sth3 Cas9 specifically promotes pre-crRNA3 and tracrRNA3 annealing and facilitates formation of the pre-crRNA3:tracrRNA3 duplex.

## 3.2. Programmable DNA cleavage *in vitro* by Cas9

The simple modular organization of Cas9 complex, where specificity for the DNA target is encoded by a small crRNA and the cleavage reaction is executed by the Cas9 endonuclease (Gasiunas et al., 2012; Jinek et al., 2012), provides a versatile platform for the engineering of universal RNA-directed DNA endonucleases. By altering the crRNA sequence within the Cas9 complex, we tried to adapt programmable endonuclease for *in vitro* applications.

### 3.2.1. Sth3 Cas9 complex as a tool for DNA manipulation

Restriction endonucleases recognize short nucleotide sequences usually 4-8 bp in length and in the presence of $Mg^{2+}$-ions cut both DNA strands within or close to the target site. Due to their unique specificity, restriction enzymes became indispensable tools for DNA cloning (Roberts, 2005). The flexibility when inserting gene fragments into the plasmid vector is limited by the availability of recognition sequences for restriction enzymes. To demonstrate that Sth3 Cas9 could be used as a universal RNA-guided DNA endonuclease programmed by the crRNA to target any DNA site we adapted it for cloning procedure. We selected two target sites (N1 and N2) located near the AatII and SapI restriction endonucleases sites in the pUC18 plasmid (Figure 13A). Using *in vitro* transcription of the PCR products bearing N1 and N2 spacers encoding sequences flanked by two repeat units we generated pre-crRNAs matching protospacers N1 and N2.

The theoretical length of the fragment located between Cas9-N1 and Cas9-N2 sites is 612 bp (Figure 13A). Double digestion of pUC18 plasmid with Cas9-N1 and Cas9-N2 (Figure 13B) yielded two DNA fragments (~2000 bp and ~600 bp), indicating that plasmid was cut at sites targeted by the crRNAs in the Cas9-N1 and Cas9-N2 complexes. Sequencing data confirmed that plasmid DNA cleavage occurred at both sites 3 nt upstream of the PAM sequence generating blunt-ended plasmid vector (Figure 13C). To demonstrate that the plasmid vector pre-cleaved with Cas9 is suitable for further manipulations, we ligated a PCR fragment amplified from the region containing a promoter and a tetracycline resistance gene in the pACYC184 plasmid (Figure 13D) and used a resulting ligation mix to transform *E.coli* cells. The resulting clones were selected on media enriched by tetracycline and ampicillin. Sequencing data confirmed that the PCR fragment was inserted at the desired position and no mutations or INDELs were introduced upon Cas9 cleavage. Taken together these experiments demonstrate that i) Cas9 can be used as a programmable DNA endonuclease suitable for cloning experiments, ii) multiple cleavage can be performed simultaneously by different Cas9 complexes.

### 3.2.2. *In vitro* cleavage of genomic DNA by the Cas9 complex

The predicted length of the DNA target recognized by the Sth3 Cas9 is determined by the 20 nt sequence complimentary to the crRNA plus 3 nt of PAM sequence. Plasmid interference experiments *in vivo* indicate that 6 nt at the PAM distal end of the protospacer are not important for interference suggesting that only 14 nt matching to the crRNA and PAM sequence are absolutely required for DNA interference by the Sth3 Cas9 (Sapranauskas et al., 2011). In theory, the 17 nt recognition sequence allows to design Sth3 Cas9

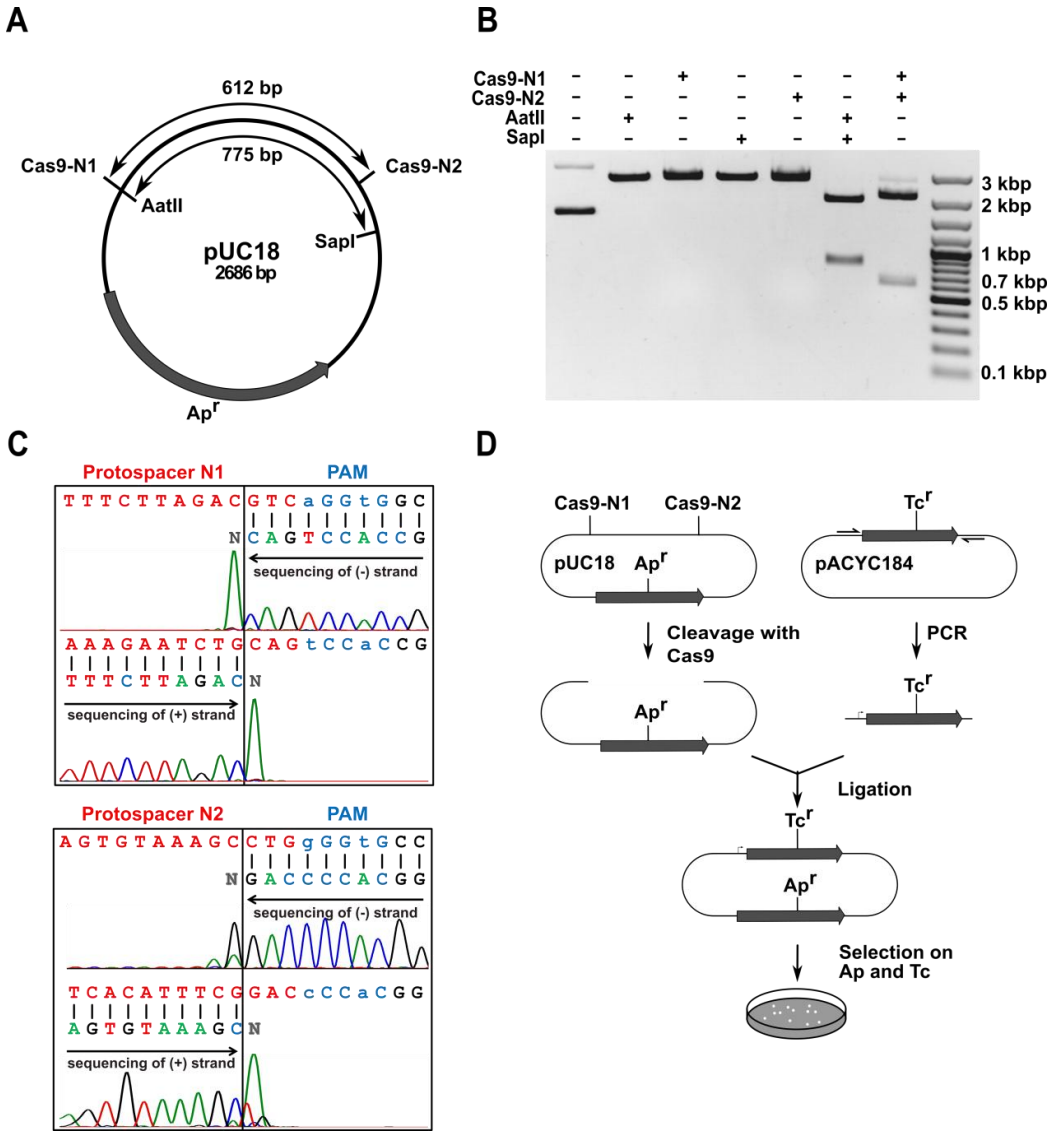


**Figure 13. Sth3 Cas9 complex as a tool for DNA manipulation.** (**A**) Schematic representation of the target sites in the pUC18 plasmid. The distance between SapI and AatII restriction enzymes sites is 775 bp, whereas the distance between two protospacer sites Cas9-N1 and Cas9-N2 is 612 bp. (**B**) pUC18 plasmid cleavage by reprogrammed Cas9 complexes. Cleavage products resulting from the double digestion using SapI and AatII restriction enzymes are also shown. Sizes are indicated in kbp. (**C**) Run-off sequencing data of pUC18 pre-cleaved with Cas9-N1 (upper panel) and Cas9-N2 (lower panel) complexes. Cleavage of plasmid DNA at both sites occurred 3 nt upstream of the PAM. (**D**) Schematic representation of the cloning experiment. Plasmid vector pre-cleaved with Cas9-N1 and Cas9-N2 complexes was ligated with a PCR fragment encoding a promoter and a tetracycline resistance gene. Clones were selected on media containing tetracycline (Tc) and ampicillin (Ap).

complexes that target unique sequences in genomic DNA. To find out whether Cas9 is able to locate and cleave target sites *in vitro* in the context of genomic DNA, we programmed the Cas9 complex for unique sites in bacteriophage λ

(49 kbp), *E. coli* (4.6 Mbp) and human (3.2 Gbp) DNA. The analysis of phage λ genomic DNA cleavage products in agarose gel confirmed that the 49 kbp DNA is cleaved only once yielding 42 kbp and 7 kbp products as predicted (Figure 14A). To identify cleavage products of the *E. coli* genomic DNA we used southern blot analysis. More specifically, we first treated genomic DNA with an excess of Cas9 complex, followed by PstI restriction enzyme digestion, and then analyzed reaction products by Southern blotting using a probe designed against the DNA fragment between the Cas9 cleavage site and a downstream PstI target (Figure 14B). The distance between two PstI targets is ~1500 bp, while the distance between a protospacer and the downstream PstI target is ~500 bp. After Cas9 cleavage, we detected only ~500 bp DNA fragment, which means that *E. coli* genomic DNA was cleaved by Cas9 at the desired position.

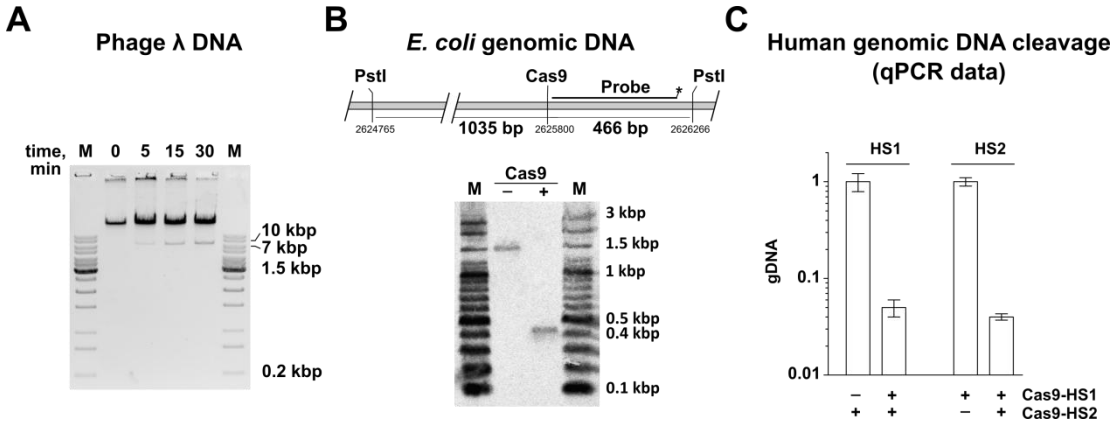


**Figure 14. Genomic DNA cleavage *in vitro* by the Cas9 complex.** (**A**) λ DNA cleavage. Phage λ DNA was incubated with the Cas9 complex for various times and reaction products were analyzed by agarose gel electrophoresis. The Cas9 is programmed for a target located 8 kb away from the *cos* site. (**B**) *E. coli* genomic DNA cleavage. Genomic DNA extracted from the *E. coli* BL21 strain was incubated for 30 min with Cas9 complex before fragmentation with PstI. The Cas9 is programmed for a target located between two PstI sites. If genomic DNA is cleaved by the Cas9, a 466 bp fragment should be detected by probe hybridization; in the absence of cleavage, the hybridization probe should detect the 1499 bp fragment. (**C**) Human genomic DNA cleavage by Cas9. Following Cas9-HS1 and Cas9-HS2 cleavage, the relative amounts of uncleaved DNA targets (HS1 or HS2) were estimated by qPCR. The ratio of cleaved to uncleaved DNA target quantified by qPCR is at least 25:1.

In the next set of experiments we analyzed human DNA cleavage by the Cas9-HS1 and Cas9-HS2 complexes targeted to RASGEF1C (HS1) and ARL15 (HS2) loci, respectively. Cleavage reaction was monitored by qPCR (quantitative real-time PCR) (Figure 14C). After treatment with Cas9-HS1 and Cas9-HS2, the amount of intact DNA targets in both loci decreased more than 25 times indicating that human genomic DNA was cleaved at the desired loci. Taken together these data demonstrate that Sth3 Cas9 effectively finds and cleaves its targets in the context of genomic DNA of different complexity including viral, bacterial and mammal DNA.

## 3.3. Targeted gene editing by transfection of *in vitro* reconstituted Sth3 Cas9 nuclease complex

Recently, Cas9 protein emerged as a promising tool for genome editing in human and other eukaryotic cells. Cas9 complex guided by RNA locates and binds to the target site, while the Cas9 protein cuts DNA generating a double strand break (DSB) within the target sequence (Gasiunas et al., 2012; Jinek et al., 2012). In eukaryotic cells, DSB is repaired by "error prone" non-homologous end joining (NHEJ) (Lieber, 2010) or by homology directed repair (HDR) (Moynahan and Jasin, 2010) mechanisms resulting in the genome modification or insertion of new genetic information (Carroll, 2011; Perez-Pinera et al., 2012; Urnov et al., 2010).

Spy Cas9 (Jinek et al., 2012) is currently used as a model system for genome editing applications. Typically, the DNA expression cassettes encoding nucleus targeted codon optimized Cas9 protein and single-guide RNAs (sgRNAs) are transfected into the cells (Cong et al., 2013; Jinek et al., 2013; Mali et al., 2013a). The efficiency of DNA cleavage by plasmid delivered Cas9 in eukaryotic cells depends on multiple factors, including expression vector design, transfection efficiency, cell type, recovery yield of functional Cas9 complex (Hsu et al., 2013), and usually requires optimization of a set of experimental conditions. Cas9 delivery by plasmid transfection is

still difficult to achieve for some hard to transform cell lines including human primary cells and pluripotent stem cells (Yamano et al., 2010; Kim and Eberwine, 2010; Li et al., 2014). Moreover, plasmid transfection occasionally results in undesirable integration of vector plasmid into the genome and is often inefficient and stressful to cells (Gabriel et al., 2011). To overcome these issues we tried to adapt an alternative way for the Cas9 mediated genome modification in eukaryotic cells (Figure 15A) by using chemical transfection of *in vitro* reconstituted functionally active Sth3 Cas9 complex.

Sth3 Cas9 complex bearing the nuclear localization signal (NLS) was reconstituted *in vitro* using crRNA:tracrRNA duplex or an artificial single guide RNA (sgRNA). To enable the delivery of reconstituted Sth3 Cas9 complex into CHO-K1 cells, transfection experiments were performed using a protein delivery agent TurboFect™. Alternatively, other transfection reagents like Lipofectamine® 2000 or 3000 can be used to transfect Sth3 Cas9 complexes into cells (data not shown).

To monitor the DNA cleavage activity of transfected Cas9 complexes in mammalian cells, we constructed a dual reporter cassette bearing Red Fluorescent Protein (RFP) and enhanced Green Fluorescent Protein (eGFP) genes (Figure 15B). eGFP gene contains two sites, L1 and L2, targeted by two different Sth3 Cas9 complexes. The I-CreI nuclease (Jurica et al., 1998) target site was also engineered into the cassette. In the absence of Cas9, eGFP fluorescence should be observed following intron processing *in vivo*. Cas9 facilitated DSB at L1 or L2 target site should trigger DNA repair either through NHEJ or HDR. In case of NHEJ, mutations within the eGFP gene would result in lost or diminished eGFP fluorescence. HDR, on the other hand, should result in the RFP expression due to reassembly of the RFP gene via the engineered homology arms, and enable quantification of HDR efficiency within the population of transfected cells. Integration of the dual reporter cassette into a plasmid vector generated a reporter plasmid while integration into the CHO-K1 cell genome produced a stable reporter cell line.

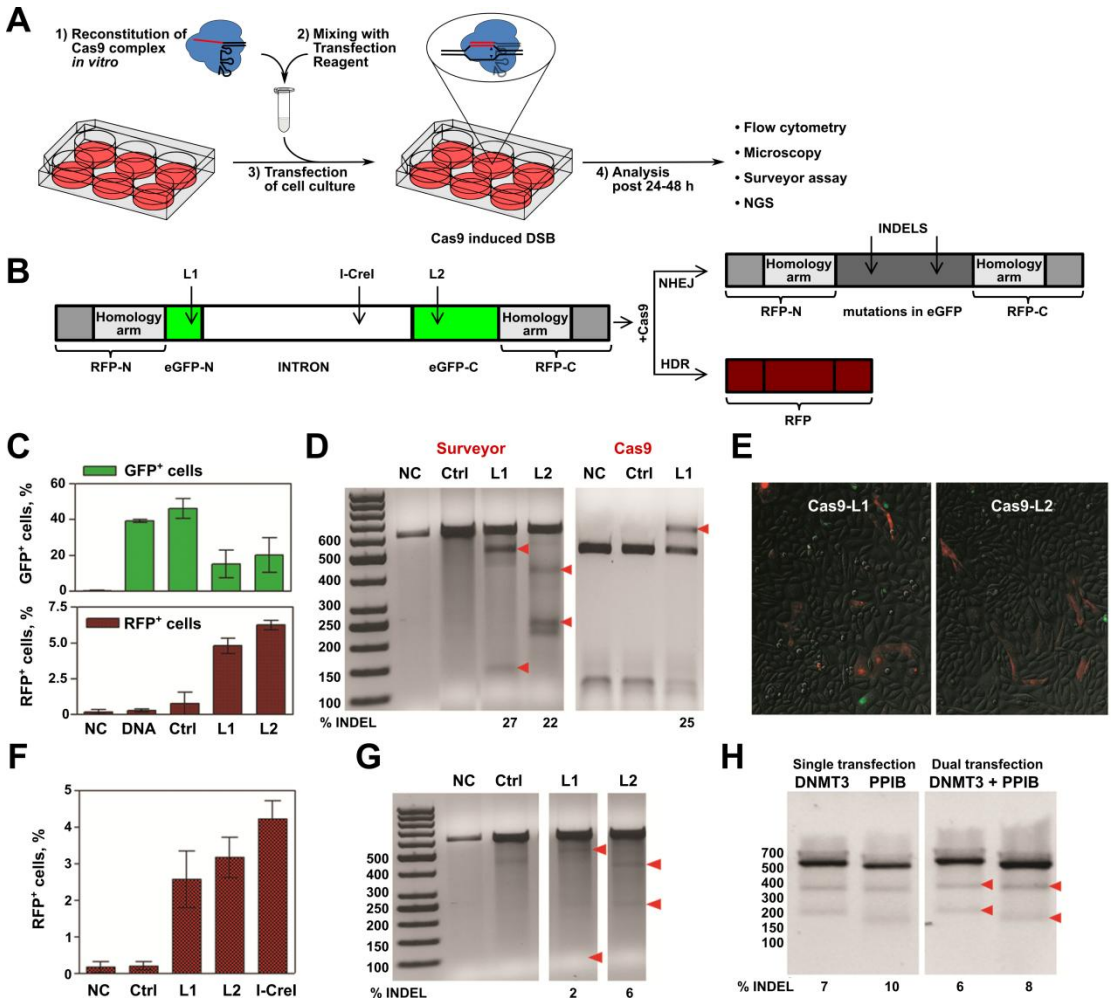


**Figure 15. Gene editing by *in vitro* reconstituted *S. thermophilus* Cas9 complex.** (**A**) Workflow of the cell transfection experiment. (**B**) CMV-promoter driven dual reporter gene cassette for the Cas9 cleavage analysis and expected NHEJ or HDR repair products. L1 and L2 designate 2 different Cas9 target sites. Insertion of the reporter gene cassette into a plasmid vector generates a reporter plasmid while insertion into the CHO-K1 cell genome generates a stable reporter cell line. (**C**) Percentage of $GFP^+$ and $RFP^+$ cells (in L1 and L2-targeted samples) estimated by a flow cytometry. (**D**) Cleavage analysis of the reporter plasmid. Regions surrounding Cas9 target sites in the reporter plasmid were PCR amplified and reannealed PCR amplicons (592 bp) were digested with Surveyor nuclease or L1 targeting Cas9 complex. Red arrows indicate 100+492 bp and 363+229 bp hydrolysis products for L1 and L2, respectively, and undigested 592 bp fragment. Numbers below indicate INDEL % calculated by densitometric analysis of corresponding bands. (**F**) FACS analysis of $RFP^+$ cells in the reporter CHO-K1 cell line transfected by Cas9 or I-CreI nucleases. (**E**) Detection of $RFP^+$ cells using fluorescent microscope imaging. (**G**) Cleavage analysis of the chromosomal Cas9 target sites in the CHO-K1 cell line. PCR amplicons (592 bp) were digested with Surveyor nuclease to produce 100+492 bp fragments (L1) and 363+229 bp fragments (L2), respectively. (**H**) HEK293T cells were transfected with Cas9 complexes specific for DNMT3B and/or PPIB gene loci. Surveyor digestion was performed on reannealed PCR amplicons - 544 bp (DNMT3B) or 505 bp (PPIB) yielding hydrolysis products of 335+209 bp or 330+174 bp, respectively (red arrowheads).

To demonstrate Sth3 Cas9 functional activity *in vivo*, the dual reporter plasmid was co-transfected together with preassembled Sth3 Cas9 complexes. Sth3 Cas9 and reporter DNA transfection mixtures were prepared in separate tubes and added to the cell culture at the same time; the percentage of $eGFP^+$ cells was estimated 48 h later by flow cytometry. When cells were transfected with the reporter plasmid alone, the percentage of $eGFP^+$ cells was ~40%, indicative of overall transfection efficiency (Figure 15C). Upon the Sth3 Cas9 complex transfection, the percentage of $eGFP^+$ cells was reduced to 5-15%. Cell transfection by Sth3 Cas9 complexes bearing non-targeting crRNA had no affect on GFP fluorescence. Surveyor nuclease assay (Guschin et al., 2010) revealed ~27% and ~22% of insertions/deletions (INDELs) at the L1 and L2 sites, respectively, indicating that Cas9 cleavage at these sites triggered DSB repair via NHEJ. These values were further verified by an independent *in vitro* cleavage assay. In this assay amplified eGFP gene fragments were digested *in vitro* with the Cas9 complex targeting L1 site. Assuming that NHEJ introduces INDELs, we expected that PCR fragments with INDELs will remain intact while unmodified fragments will be cut. It turned out that the percentage of the Cas9-resistant DNA (25%) is very similar to that established by the Surveyor assay. Thus, *in vitro* digestion of PCR amplicons with Cas9 provides an alternative to Surveyor assay for INDEL quantification in Cas9-mediated genome modification *in vivo* (Figure 15D). In this case the INDEL detection is faster compared to Surveyor assay since no additional steps of amplicon reannealing is required. Taken together these results demonstrate that *in vitro* reconstituted Cas9 complex delivered by transfection promotes generation of DSB at the target site, and subsequent repair by NHEJ produces INDELs and inactivates the eGFP gene. Importantly, INDEL percentage in the target gene after pre-assembled Sth3 Cas9 complex transfection (Figure 15D) was similar to that resulting from transfection of Sth3 Cas9 and sgRNA encoding plasmids (data not shown).

To find out whether HDR contributes to the DSB repair in the reporter plasmid, we looked at the appearance of $RFP^+$ cells: RFP was expressed in 5-

8% of cells transfected with L1 or L2 targeting Sth3 Cas9 complexes, whereas no red cells were detected in cultures transfected with non-targeting Cas9 complexes (Figure 15C). The presence of RFP positive cells was subsequently confirmed by fluorescence microscopy (Figure 15E).

For further analysis of Sth3 Cas9 cleavage activity on the chromosomal DNA we used the CHO-K1 cell line with a dual reporter cassette (Figure 15B) integrated into the genome. The weak eGFP fluorescence signal in the engineered cell line hindered a reliable quantification of the Cas9 cleavage efficiency by monitoring the decrease of the eGFP signal by flow cytometry. The Surveyor assay, however, revealed that INDEL values reached 2-6% for respective chromosomal target sites (Figure 15G). FACS analysis confirmed that about 3-4% of cells expressed RFP in reporter cell cultures (Figure 15F). Transfection of recombinant I-CreI nuclease yielded ~4.5% of $RFP^+$ cells, indicating that similar cleavage efficiencies can be achieved by different recombinant nucleases delivered using chemical transfection (Figure 15G). Cas9 cleavage specificity was further verified by deep sequencing using MiSeq system (Illumina). NHEJ-mediated INDELs were centered about the target site validating the cleavage specificity (data not shown).

To extend the study of recombinant Cas9 potential in multiplex genome modulation, a series of experiments were carried out using HEK293T cells and Cas9 complexes specific for endogenous genes DNMT3B or PPIB. The cells were transfected with each gene targeting complexes separately (single transfection) or together (dual transfection). Surveyor assay revealed that for Sth3 Cas9 complex INDEL values for genes DNMT3B or PPIB reached 7% and 10% in single transfections or 6% and 8% in dual transfections, respectively (Figure 15H). Overall, this data validated that *in vitro* preassembled Cas9 complexes could be successfully used for genome modulation at different genomic loci and in addition demonstrated the potential of targeting multiple loci at the same time.

## 3.4. DNA target recognition mechanism of Cas9 complex

In DNA-targeting CRISPR-Cas systems, the RNA component of the complex encodes target recognition by forming a site specific hybrid (R-loop) with its complement (protospacer) on the DNA while displacing the non-complementary strand. Subsequently, the R-loop structure triggers DNA degradation. Although we were able to reconstitute these reactions *in vitro* and *in vivo* (section 3.1-3.3), the exact mechanism of R-loop formation has not been fully resolved.

### 3.4.1. Direct observation of R-loop formation

To directly observe and quantify the dynamics of R-loop formation and dissociation for Sth3 Cas9 complex in single DNA molecule supercoiling experiments, we used magnetic tweezers (Brutzer et al., 2010; Mosconi et al., 2009) (Figure 16A). A 2.1 kbp DNA containing a single protospacer S1 and PAM was attached at one end to a magnetic bead and at the other end to the bottom of a fluidic cell. A pair of magnets above the cell were used to stretch the DNA and to supercoil it by rotating the magnets. Simultaneously the DNA length was measured (Klaue and Seidel, 2009). Upon supercoiling DNA at constant force, its length stays initially constant. Once a critical torque in the molecule is reached, its length starts to decrease due to formation of a plectonemic superhelix resulting in a characteristic rotation curve and an associated torque profile (Forth et al., 2008; Kauert et al., 2011; Mosconi et al., 2009; Oberstrass et al., 2012). Enzyme-dependent local DNA unwinding, e.g. due to R-loop formation, changes the DNA twist and can be seen as a shift of the whole rotation curve or as a DNA length change (Howan et al., 2012) (Figure 16A). To detect R-loop formation we carried out “R-loop cycles” with Cas9 on DNA with matching protospacer and canonical PAM sequence (Figure 16B). First we slightly untwisted the DNA (producing negative supercoiling) at low force to help R-loop formation (blue curves in Figures

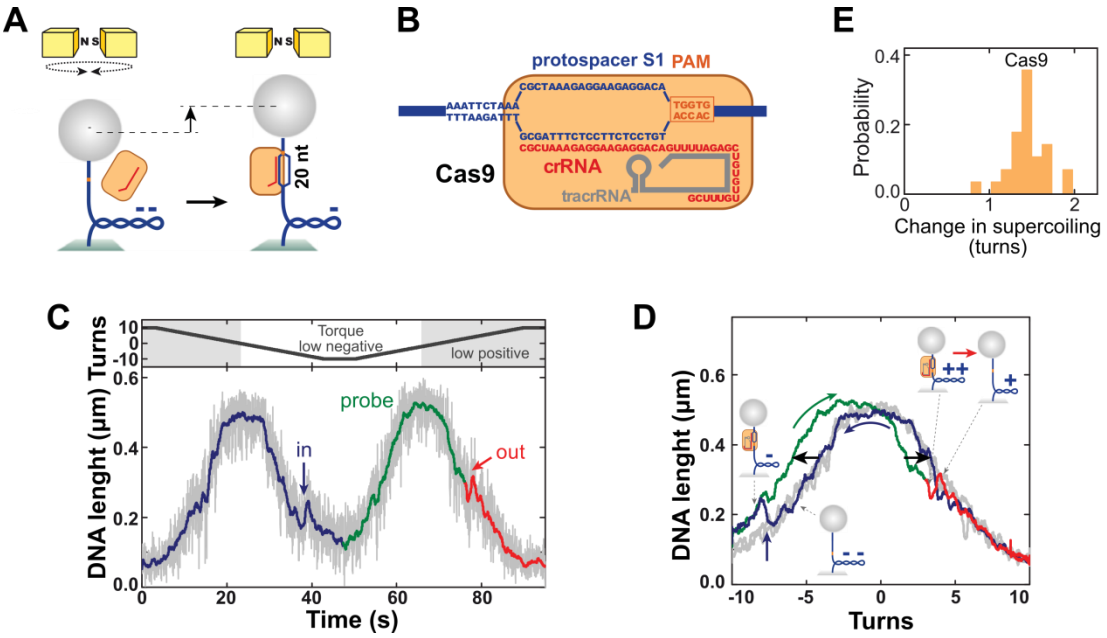


**Figure 16. R-loop formation and dissociation by Cas9 observed in single-molecule twisting experiments.** (**A**) Magnetic tweezers-based twisting assay. R-loop formation on supercoiled DNA molecules at fixed rotation causes local DNA untwisting. Compensatory overtwisting of the DNA changes the supercoiling, resulting in a DNA length change. (**B**) Schematics of the anticipated R-loops formed by Cas9 (20 bp). (**C** and **D**) R-loop cycle experiment in the presence of 1 nM Cas9. DNA with matching protospacer and PAM is negatively supercoiled at 0.31 pN to induce R-loop formation (blue area of trace), followed by positive supercoiling to probe the presence of the R-loop (green area of trace); and R-loop dissociation (red area of trace). R-loop dissociation occurs readily at low positive torque. Blue and red arrows indicate the positions of R-loop formation and dissociation, respectively. (**E**) Cas9-induced shift of the supercoiling curve (orange bars).

16C and D). Subsequently we probed R-loop dissociation by rewinding the DNA to produce positive supercoiling (green curves in Figures 16C and D). We observed efficient R-loop formation (100% of all cases, N = 50), which was seen as a shift of the left side of the probe curve towards negative turns compared to the curve in absence of the protein (Figure 16D). Cas9-induced R-loops dissociated readily at low positive force (Figures 16C and D). The observed shifts in the rotation curves were

dependent on the presence of a protein complex, a matching protospacer and a corresponding crRNA. For Cas9, 1.45±0.05 turns were obtained from the center shift (Figure 16E). This value is slightly smaller than anticipated (1.9 turns considering the predicted R-loop length (20 nt) and a DNA helical pitch of 10.5 bp), possibly due to compensatory writhe from DNA bending induced by Cas9 complex.

### 3.4.2. PAM mutations hinder R-loop formation but not its stability

To clarify whether the PAM regulates R-loop formation by kinetic inhibition or altered R-loop stability, we measured the dynamics of R-loop formation and dissociation on substrates with matching protospacers but mutated PAMs. Using magnetic tweezers experiments with Cas9 revealed, that R-loops could still be formed using a G4C mutant, albeit at a much lower rate than with WT (NGGNG) sequence (Figure 17). Once formed, R-loop stability was not compromised (Figure 17). R-loop formation using G1C and G2C was even slower while PAM deletion did not support R-loop formation at all, even under induced denaturation conditions (data not shown). In support of the tweezers data, DNA cleavage assays suggest that R-loop formation is hindered by PAM mutation (with the order G2C>G1C>G4C>>WT) but that subsequent R-loop cleavage is much less affected (Figure 18A, B and D). Thus, the PAM regulates R-loop formation and subsequent cleavage by kinetic inhibition.

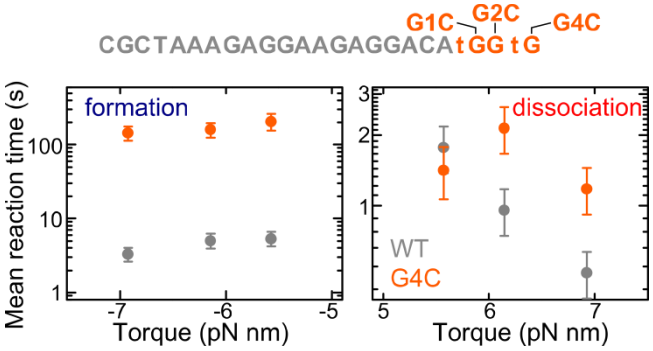


**Figure 17. PAM mutations affect primarily R-loop formation and not dissociation.** Mean R-loop formation and dissociation times as a function of torque for the G4C PAM compared with the canonical PAM.

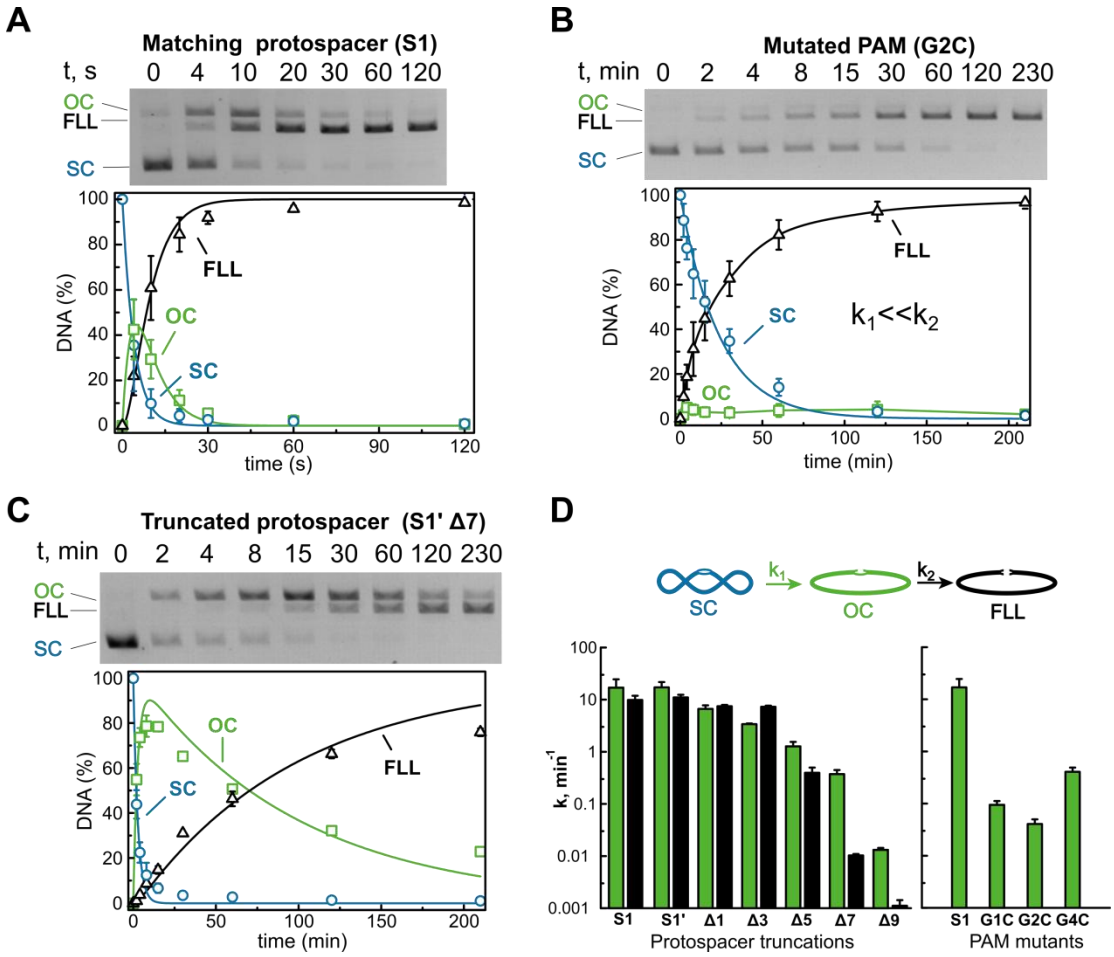


**Figure 18. Cleavage of plasmid DNA with end-truncated protospacers or mutated PAMs by the Cas9-crRNA-tracrRNA complex.** (**A**) Plasmid containing a consensus NGGNG PAM and a matching protospacer S1. Reaction products and intermediates were analyzed by agarose gel electrophoresis (gel image shown on top) and quantified (shown below). SC (blue open circles) – supercoiled plasmid, OC (green open squares) – nicked reaction intermediate cut at either DNA strand, FLL (black open triangles) – final reaction product cleaved at both DNA strands. Solid lines are fits to the data. (**B**) Cleavage reaction for a plasmid containing a G2C PAM mutation, and a matching protospacer S1. (**C**) Cleavage reaction for a plasmid containing a consensus NGGNG PAM and 7 bp protospacer truncation (S1′ Δ7 bp). (**D**) Cleavage rates for the first and second DNA strand in plasmids containing protospacers with end truncations of different lengths but a consensus PAM (left) or a mutated PAM but a matching protospacer S1 (right). $k_1$ is the rate constant for formation of the nicked intermediate (including the first strand cleavage step and any preceding rate-limiting R-loop formation steps) whilst $k_2$ is the rate constant for the appearance of the final FLL product (including the second strand cleavage step) . $k_1$ and $k_2$ values were obtained by fitting the cleavage data as previously described (Zaremba et al., 2006). All data points are mean values from ≥3 independent experiments. Error bars are given as S.D. For the PAM mutants, the accumulation of nicked intermediate was not detected (see gel in panel **B**). Therefore only the $k_1$ value is provided. We can only estimate that $k_1 << k_2$ under these conditions. This is in agreement with a greatly reduced R-loop formation rate on these DNA. The actual cleavage rates of first and second strand after R-loop formation are expected to be as fast as for the consensus PAM hence $k_2$ being much faster than $k_1$.

### 3.4.3. Protospacer end truncations destabilize R-loops for Cas9

For Cas9, R-loops were detected for truncations up to 7 bp with little change in the association rate (Figure 19) but were not detected for a 9 bp truncation. For 1 or 5 bp truncations, R-loop stability and DNA cleavage rates were slightly reduced (Figures 19 and 18). For the 7 bp truncation however, while an R-loop population with reduced stability was also detected (Figure 19), many R-loops exhibited slower, torque-independent dissociation kinetics. This suggests that mismatches between the crRNA and protospacer can cause rearrangement of the Cas9 RNP into an inhibited, off-pathway intermediate; the protospacer truncation correlated with changes in DNA cleavage rate but also with an accumulation of nicked intermediates (Figure 18C and D; nicking was even detectable for the 9 bp truncation).

### 3.4.4. Model for R-loop formation and dissociation by Cas9

Data presented above, shows that the PAM controls tightly the R-loop formation kinetics but leaves the R-loop stability practically unchanged. Thus, the PAM provides a kinetic rather than a thermodynamic control of R-loop

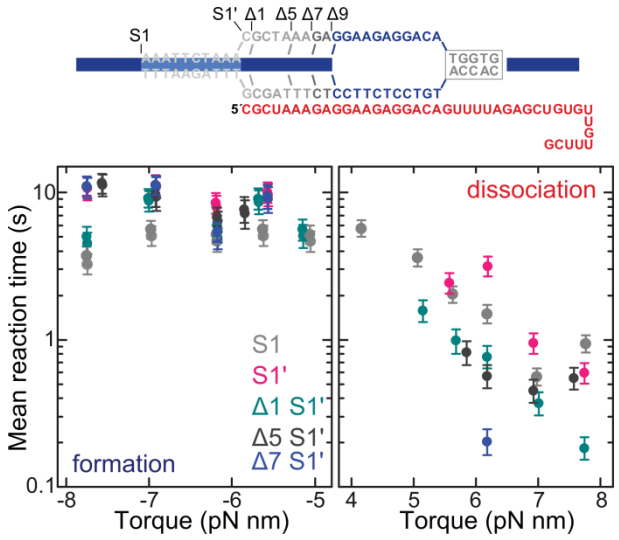


**Figure 19. Protospacer end truncations affect the R-loop stability.** Mean R-loop formation and dissociation times as a function of torque for different S1′ protospacer truncations.

formation. Distal protospacer mutations affected the R-loop stability but hardly altered the formation kinetics. This reveals a unidirectionality in the R-loop formation and dissociation cycle (Figure 20). Firstly the Cas9 complex uses DNA distortion, guided by PAM binding, to accomplish homology search. Since R-loop formation times are dependent on complex concentration (data not shown), 3D diffusion must be an integral part of the target search pathway. Matching hydrogen bonding between the crRNA and the protospacer then leads to propagation of R-loop formation over the adjacent base pairs. Under unfavorable energetic conditions (high positive torque or mismatches between the protospacer and crRNA), R-loop dissociation occurs in a PAM-independent manner.

Cas9 can efficiently cut a protospacer with a 7 bp truncations, albeit at a reduced rate (Figure 18). Nonetheless, Cas9 efficiently discriminates targets over 11 bp adjacent to the PAM similarly to Cas9 from *S. pyogenes* (Sternberg et al., 2014). Despite complementarity, R-loops of 11 bp or shorter were not formed, revealing that Cas9 can sense further into the protospacer to identify the correct target. Structurally this likely originates from the more extensive amino acid contacts with the heteroduplex in this particular region (Nishimasu et al., 2014).

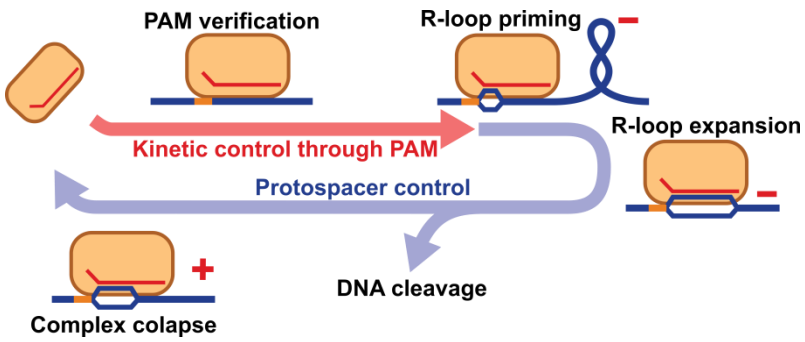


**Figure 20. Model for the differential control of R-loop formation and dissociation by PAM and protospacer sequences.** Red positive (+) and negative (–) symbols indicate where supercoiling of the respective sign can accelerate the step.

## 3.5. Rapid characterization of CRISPR-Cas9 protospacer adjacent motif (PAM) sequence elements

The DNA target site for Cas9 is composite and consists of a protospacer sequence and a short PAM sequence adjacent to the protospacer. As described previously (section 3.4) and in Sternberg et al., 2014, target recognition is achieved through a complex mechanism involving Cas9-mediated interaction with the PAM and crRNA-guided interactions with the complementary DNA of the protospacer. The process initiates with PAM recognition by Cas9 and then proceeds through crRNA-guided sequence-specific hybridization with a protospacer. In this respect, the PAM sequence plays a key role in target recognition by licensing crRNA-guided base pairing to the protospacer sequence (Sternberg et al., 2014). A strict PAM requirement constrains DNA target selection and poses a limit to Cas9 genome editing applications. Target site selection may be further confined if unique genomic sites are required especially in large complex plant genomes like maize (Xie et al., 2014). These constraints imposed by the PAM and the specificity of the Spy Cas9 could be overcome either by systematically redesigning the PAM specificity of a single Cas9 protein (Anders et al., 2016; Hirano et al., 2016a; Kleinstiver et al., 2015a, 2015b), by simply exploring the natural diversity of Cas9 proteins or by combining the two approaches. In addition to expanding the sequence space targeted by Cas9, orthologous Cas9 proteins with different biochemical activities may enhance genomic manipulation efforts. Cas9 systems with enhanced specificity or tunable activity may help mitigate off-target concerns while systems with incompatible guide RNAs or PAM sequences could be used to concertedly edit, activate or repress different targets. Thus, by combining these features unique control over genome expression and content may be afforded.

To tap into this unexplored diversity and expand the repertoire of Cas9s available for genome targeting applications, the development of a method that allows the direct-read out of Cas9 endonuclease PAM specificity is necessary.

### 3.5.1. Assaying Cas9 PAM preferences

PAM libraries containing randomized DNA sequences immediately downstream of a DNA sequence complementary to the spacer of a guide RNA were generated and used to empirically determine the protospacer adjacent motif (PAM) recognition of Type II Cas9 endonucleases (Figure 21). With the guide RNA spacer target sequence being fixed, the randomized bases serve as a substrate for the direct read-out of Cas9 endonuclease PAM specificity. Randomized sequences were introduced into a plasmid DNA vector in the PAM region of a protospacer target sequence demonstrating perfect homology to the guide RNA spacer S1. Two libraries increasing in size and complexity from 5 randomized base pairs (1,024 potential PAM combinations) to 7 randomized base pairs (16,384 potential PAM combinations) were generated as described in (sections 2.2.14.1 and 2.2.14.2). The randomized PAM libraries were subjected to *in vitro* digestion using different concentrations of recombinant Cas9 protein preloaded with guide RNA in order to assay Cas9 endonuclease PAM preferences in a dose-dependent manner. After digestion using Cas9 complexes, PAM sequence combinations from the randomized PAM library that supported cleavage were captured by ligating adapters to the free-ends of the plasmid DNA molecules cleaved by the Cas9 complex (Figures 21, steps A and B). To promote efficient ligation and capture of the cleaved ends, the blunt-ended double-stranded DNA cut generated by Cas9 endonucleases (Garneau et al., 2010; Gasiunas et al., 2012; Jinek et al., 2012) was modified to contain a 3’ dA overhang and adapters were modified to contain a complementary 3’ dT overhang. To generate sufficient quantities of DNA for sequencing, DNA fragments harboring the PAM sequence supporting cleavage were PCR amplified using a primer in the adapter and another directly adjacent to the PAM region (Figure 21, step C). The resulting PCR amplified Cas9 PAM libraries were converted into ampli-seq templates (Figure 21, step D) and single-read deep sequenced from the adapter-side of the amplicon. PAM sequences were identified from the resulting sequence data

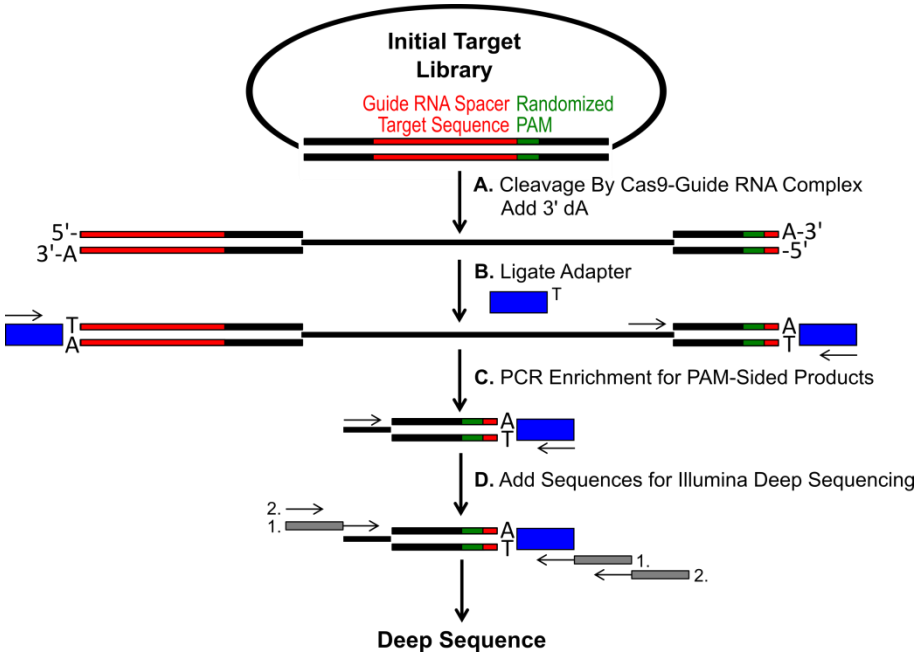


**Figure 21. Schematic for identification of PAM preferences by Cas9 cleavage *in vitro*. A**. Initial plasmid library with randomized PAM (green box) is cleaved with Cas9 complex and 3' dA overhangs are added. **B**. Adapters with 3' dT overhang (blue box) are ligated to both ends of the cleavage product. **C**. Primers are utilized to enrich for PAM-sided cleaved products by PCR. **D**. After PCR enrichment, DNA fragments are purified and Illumina compatible anchors and barcodes are "tailed-on" through two rounds of PCR (gray boxes) and Illumina deep sequenced.

by only selecting those reads containing a perfect 12 nt sequence match flanking either side of the 5 or 7 nt PAM sequence (depending on the randomized PAM library used); capturing only those PAM sequences resulting from perfect Cas9-guide RNA target site recognition and cleavage. To compensate for inherent bias in the initial randomized PAM libraries, the frequency of each PAM sequence was normalized to its frequency in the starting library. The results were visualized as a WebLogo (Crooks, 2004).

### 3.5.2. PAM preferences of *Streptococcus pyogenes* and S*treptococcus thermophilus* (CRISPR3 and CRISPR1 systems) Cas9 proteins

In order to validate the assay, the PAM preferences of Spy and Sth3 Cas9 proteins, whose PAM sequence requirement have been previously reported (Gasiunas et al., 2012; Horvath et al., 2008; Jinek et al., 2012; Sapranauskas et al., 2011), were examined. *In vitro* digests were carried-out with 1 μg (5.6 nM) of the 5 bp randomized PAM library at two concentrations, 0.5 and 50 nM, of pre-assembled Spy or Sth3 Cas9 protein, crRNA and tracrRNA RNP complexes (Gasiunas et al., 2012; Jinek et al., 2012). Based on their frequency in the 5 bp randomized PAM library, Spy and Sth3 Cas9 PAM sequences (NGG and NGGNG, respectively) were at final concentrations of 0.40 nM and 0.11 nM in the digestion, respectively. Members of the randomized PAM library that contained PAM sequences which supported cleavage were captured and identified as described in the previous section.

Examination of the PFM derived WebLogos (Figures 22A and B) reveal the presence of the canonical PAM preferences for the Spy and Sth3 Cas9 proteins, NGG (Jinek et al., 2012) and NGGNG (Gasiunas et al., 2012; Horvath et al., 2008; Sapranauskas et al., 2011), respectively. Although the PAM preferences reported for Spy and Sth3 Cas9 proteins are observed in both the 0.5 nM and 50 nM digests, there is a general broadening in specificity under the 50 nM digest conditions. This is most evident at position 2 for the Spy Cas9 protein where the frequency of a non-canonical A residue increases dramatically (Figure 22A). For Sth3, all PAM positions exhibit a marked decrease in specificity as a result of increasing the RNP complex concentration (Figure 22B).

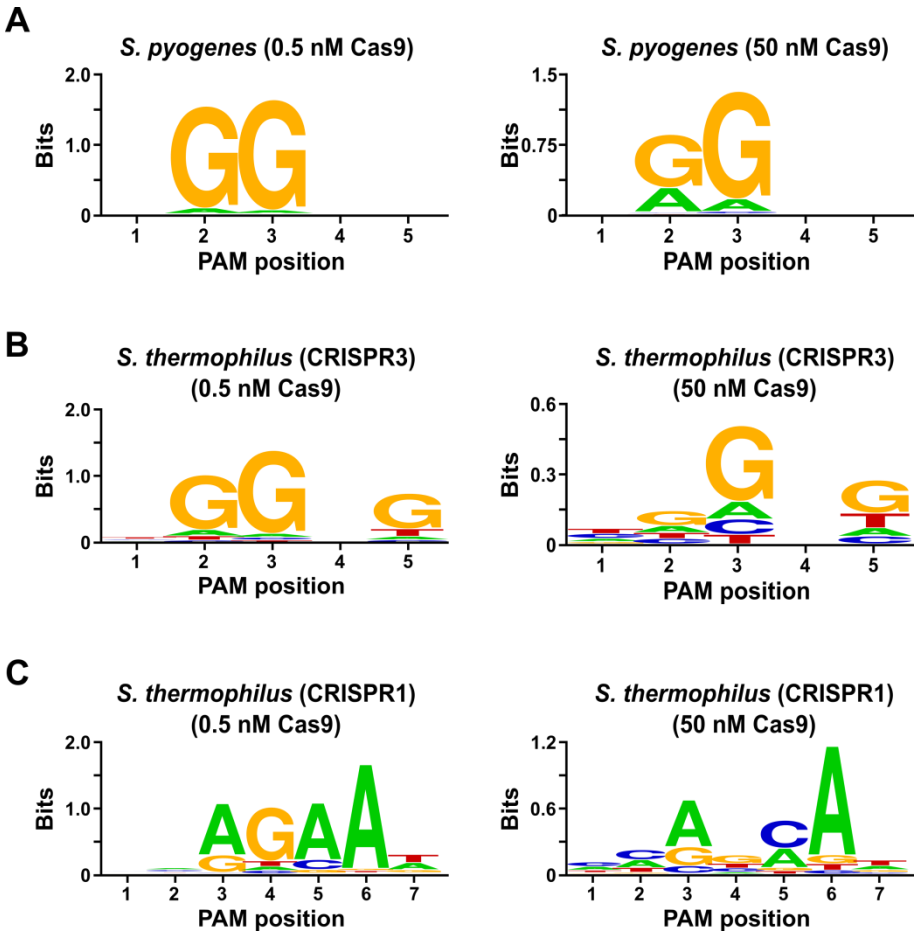


**Figure 22. PAM preferences for *S. pyogenes* (A), *S. thermophilus* CRISPR3 (B) and *S. thermophilus* CRISPR1 (C) Cas9 proteins.** Frequency of nucleotides at each PAM position was independently calculated using a position frequency matrix (PFM) (Stormo, 2013) and plotted as a WebLogo (Crooks, 2004).

Further validation of the assay was conducted by examining the PAM preferences for the *Streptococcus thermophilus* CRISPR1 (Sth1) Cas9 protein whose PAM specificity has been reported to extend out to 7 bp (Fonfara et al., 2014; Horvath et al., 2008). Using 1 µg (5.6 nM) of the 7 bp randomized PAM library as template, Sth1 Cas9-guide RNA digestions were carried-out at two concentrations, 0.5 nM and 50 nM, of RNP complex as described in section 2.2.14.4. Based on the frequency in the 7 bp randomized PAM library, the PAM sequences previoulsy reported for Sth1 (NNAGAAW), was at final concentrations of 0.01 nM. The PAM preferences for the Sth1 Cas9 protein closely matched that previously reported, NNAGAAW (Horvath et al., 2008), at the 0.5 nM Cas9-guide RNA complex concentration (Figure 22C). Similar to Spy and Sth3 Cas9 proteins, Sth1 Cas9 was capable of cleaving a more diverse

set of PAM sequences in the reactions containing a higher concentration of Cas9-guide RNA complex (50 nM), the most striking was the marked loss of the G residue requirement at position 4 and the near equal preference for a C and A bp at position 5 (Figure 22C). This resulted in a different PAM consensus than that obtained at lower concentrations. This finding corroborates previous studies which demonstrated that lowering Cas9 concentration and shortening cleavage time prevents off-target cleavage by Spy Cas9 *in vivo* (Lin et al., 2014; Pattanayak et al., 2013). Additionally, most other PAM determination methods have been performed in cells or cell extracts by expressing Cas9 at undefined concentrations (Esvelt et al., 2013; Jiang et al., 2013; Kleinstiver et al., 2015a; Leenay et al., 2016; Ran et al., 2015). Given this, the Cas9 PAM recognition results from these studies may be difficult to accurately interpret. A case in point is reflected in the inability of previous attempts (Esvelt et al., 2013; Fonfara et al., 2014; Ran et al., 2015) to precisely reproduce the PAM recognition of *S. thermophilus* CRISPR1 (Sth1) Cas9 protein originally reported by Horvath et al., 2008, while the methods described here accurately recapitulate the PAM recognition of Sth1 Cas9 albeit at lower Cas9-guide RNA ribonucleoprotein complex concentrations. Taken together, the method established here further refines PAM specificity assessments by the dose-dependent control of recombinant Cas9 protein *in vitro* enabling an accurate detailed examination of Cas9 PAM recognition as a function of Cas9 and guide RNA complex concentration.

### 3.5.3. Identification of sgRNA and PAM preferences for the *Brevibacillus laterosporus* Cas9 protein

To empirically examine the PAM preferences for a Cas9 protein whose PAM was undefined, an uncharacterized Type II-C CRISPR-Cas locus from *Brevibacillus laterosporus* strain SSP360D4 (Blat) was identified by searching internal DuPont Pioneer databases for Cas9 orthologues. The locus (approximately 4.5 kb) contained a *cas9* gene capable of encoding a 1,092

polypeptide, a CRISPR array comprised of 7 repeat-spacer units just downstream of the *cas9* gene and a tracrRNA encoding region located upstream of the *cas9* gene with partial homology to the CRISPR array repeats (Figure 23A). The repeat and spacer length (36 and 30 bp, accordingly) is similar to other Type II CRISPR-Cas systems with 5 of the 8 repeats containing 1 or 2 bp mutations (Figure 23B). Other genes typically found in a Type II CRISPR-Cas locus were either truncated (*cas1*) or missing (Figure 23A).

The guide RNA requirement for the Blat Cas9 protein was determined by generating two sgRNA variants. These variants were generated to account for both possible sense or anti-sense expression scenarios of the tracrRNA and CRISPR array (Figure 23C) and used to probe which expression scenario supported cleavage activity of Blat Cas9 in the randomized PAM library. Single guide RNAs were designed by first identifying the boundaries of the putative tracrRNA molecules by analyzing regions which were partially complementary to the 22 nt 5' terminus of the repeat (anti-repeat). Next, to determine the 3' end of the tracrRNA, possible secondary structures and terminators were used to predict the region of termination in the downstream fragment. This was accomplished by screening for the presence of Rho independent-like termination sequences in the DNA surrounding the anti-repeat, converting the surrounding DNA into RNA sequence and examining the resulting structures using UNAfold (Markham and Zuker, 2008). The resultant sgRNAs were designed to contain a T7 polymerase transcription initiation recognition signal at the 5' end followed by a 20 nt target recognition sequence, 16 nt of crRNA repeat, 4 nt self-folding hairpin loop and anti-repeat sequence complementary to the repeat region of the crRNA followed by the remaining 3' part of the putative tracrRNA. The sgRNA variant which contains a putative tracrRNA transcribed in the same direction as the *cas9* gene (Figure 23C) is termed "direct" sgRNA, while the sgRNA containing the tracrRNA transcribed in the opposite direction a "reverse" sgRNA. Fifty nM of Blat Cas9

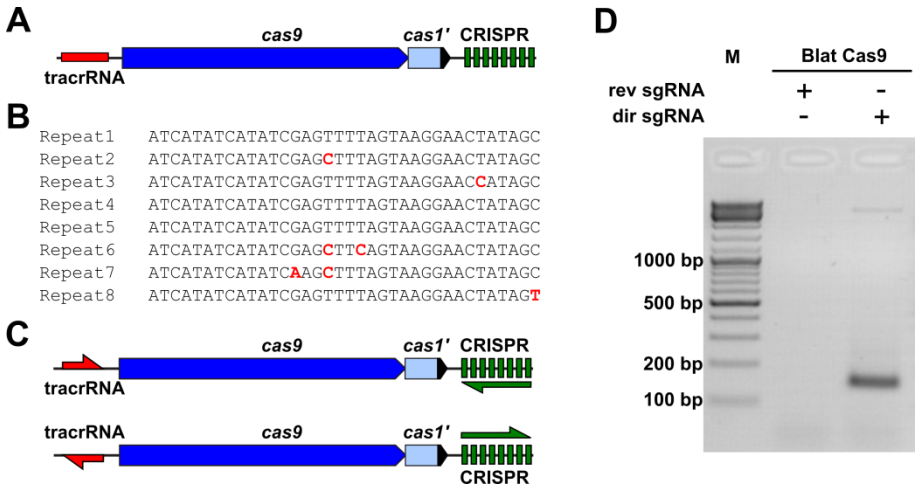


**Figure 23. Identification of elements in *Brevibacillus laterosporus* SSP360D4 CRISPR-Cas system.** (**A**) An illustration of the genomic DNA region from the Type II CRISPR-Cas system from *Brevibacillus laterosporus* SSP360D4. (**B**) Comparison of Type II CRISPR array repeat sequences identified in *Brevibacillus laterosporus* SSP360D4. (**C**) The "direct" and "reverse" tracrRNA and CRISPR array transcriptional scenarios for the Type II CRISPR-Cas system from *Brevibacillus laterosporus* SSP360D4. (**D**) An agarose gel with reaction products, indicating that only the "direct" sgRNA (dir sgRNA), but not the "reverse" sgRNA (rev sgRNA) support plasmid library cleavage in combination with the Cas9 endonuclease originating from *Brevibacillus laterosporus* SSP360D4 (Blat Cas9).

sgRNA RNP complex, pre-loaded with either the "direct" or "reverse" sgRNAs, respectively, were incubated with 1 µg (5.6 nM) of the 7bp randomized PAM library. After library digestion and addition of 3' dA overhangs, adapters were ligated and cleavage products were PCR amplified (Figure 20). Analysis of reaction products by agarose gel electrophoresis revealed that the "direct" sgRNA, but not the "reverse" sgRNA supported plasmid library cleavage (Figure 23D).

After determining the appropriate guide RNA for Blat Cas9, PAM identification was performed similarly to that described above for the Spy, Sth3 and Sth1 Cas9 proteins against the 7 bp randomized PAM library with two concentrations, 0.5 and 50 nM, of pre-assembled Blat Cas9 "direct" sgRNA RNP complex. As shown in Figure 24A, the PFM WebLogo PAM consensus for the Blat Cas9 protein under the 0.5 nM digest conditions was NNNNCND (N=G,C, A or T; D=A, G or T) with a strong preference for a C at position 5 of the PAM sequence. A moderate preference for an A was observed at position 7 and slight preferences for a C or T at position 4 and G, C or A

over T at position 6 were also present. Similarly to Spy, Sth3 and Sth1 Cas9 proteins, the PAM specificity broadens as the Cas9 RNP complex concentration increases. This is most evident at position 5 where a larger proportion of PAM sequences containing an A residue support cleavage at 50 nM compared with 0.5 nM digest conditions.

Since Blat Cas9 may accept any base in the first 3 positions of its PAM sequence (Figure 24A), the spacer S1 was shifted by 3 nucleotides in the 5' direction to allow PAM identification to be extended from 7 to 10 bp. The shifted S1 spacer, S1-3, was incorporated into the Blat "direct" sgRNA and PAM identification was performed as described previously for Spy, Sth3, Sth1 and Blat Cas9 proteins. PAM preference analysis revealed the PAM specificity for Blat Cas9 may be extended out to position 8 where there is a moderate preference for an additional A (Figure 24B).

PAM specificity for Blat Cas9 was confirmed by generating plasmids to contain mutations in the most conserved residues of the PAM (Figure 24C). Replacement of the C nucleotide at the 5th position abolished plasmid DNA cleavage confirming its key role in Blat Cas9 PAM recognition. Replacement of A nucleotides at the 7th and 8th positions significantly reduced (43× and 12×, respectively) the cleavage rate of supercoiled plasmid also indicating the importance of these nucleotides in Blat Cas9 PAM recognition.

To identify the DNA target cleavage positions for the Blat Cas9 protein, a plasmid containing a 20 base pair region matching the spacer S1 followed by a PAM sequence, GTCCCGAA, falling within the PAM consensus for Blat Cas9, NNNNCNDD, was generated and digested with Blat Cas9-guide RNA ribonucleoprotein complex. Direct DNA sequencing was used to determine the ends of the linear DNA molecule generated by the Blat Cas9 RNP complex. The sequencing results confirmed that plasmid DNA cleavage occurred in the protospacer 3 nt 5' of the PAM sequence (Figure 24D) similar to that observed for Spy, Sth3 and Sth1 Cas9 proteins (Garneau et al., 2010; Gasiunas et al., 2012; Jinek et al., 2012).

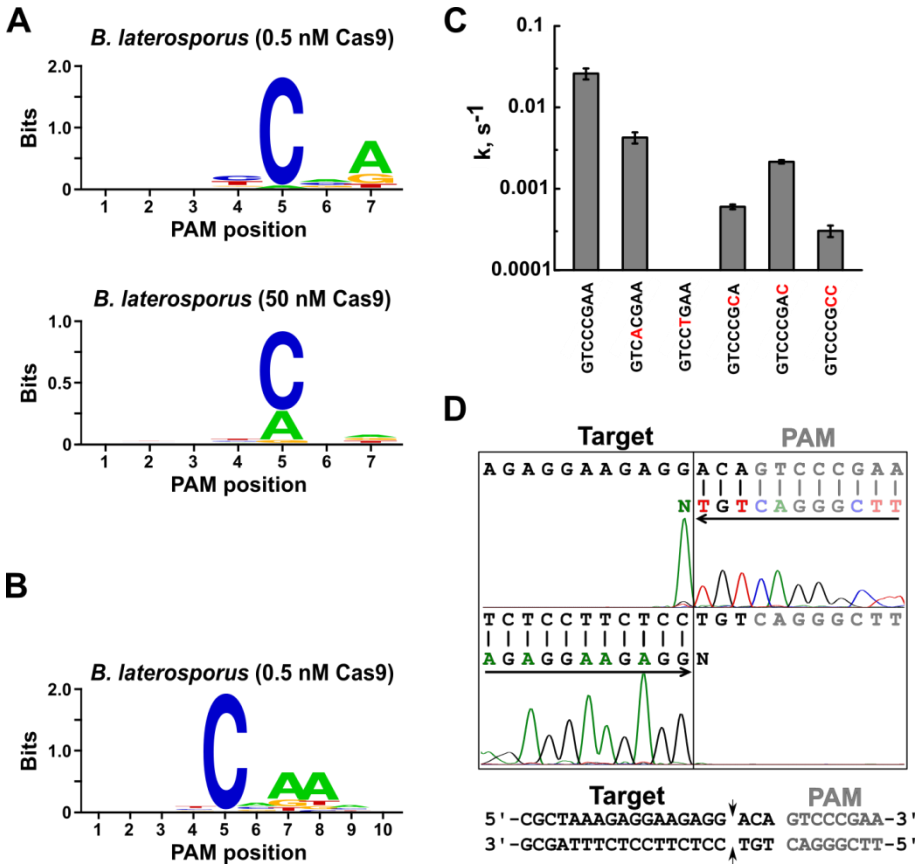


**Figure 24. PAM preferences and cleavage positions of Blat Cas9 enzyme.** Blat Cas9 PAM preferences when 1 μg of library DNA was cleaved with 0.5 nM or 50 nM Cas9-sgRNA complex (**A**), extended out to position 10 by shifting the protospacer target by 3 bp (**B**). Frequency of nucleotides at each PAM position was independently calculated using a position frequency matrix (PFM) (Stormo, 2013) and plotted as a WebLogo (Crooks, 2004). (**C**) Cleavage rates of supercoiled plasmid DNA substrates containing mutations (shown in red) in GTCCCGAA PAM sequence. All data points are mean values from ≥3 independent experiments. Error bars are given as S.D. (**D**) Run-off sequencing from both sense and anti-sense directions of plasmid DNA cleaved with Blat Cas9.

### 3.5.4. *In planta* genome editing using Blat Cas9 and sgRNA

Following elucidation of the sgRNA and PAM preferences for Blat Cas9, maize optimized Cas9 and sgRNA expression cassettes were generated for *in planta* testing as previously described for the Spy *cas9* gene and sgRNA (Svitashev et al., 2015). Briefly, the Blat *cas9* gene was maize codon optimized and intron 2 of the potato *STH3 LSI* gene was inserted to disrupt expression in *E. coli* and facilitate optimal splicing *in planta* (Libiakova et al., 2001). Nuclear localization of the Blat Cas9 protein in maize cells was facilitated by

the addition of both amino and carboxyl-terminal nuclear locations signals, SV40 (MAPKKKRKV) and *Agrobacterium tumefaciens* VirD2 (KRPRDRHDGELGGRKRAR), respectively. The Blat *cas9* gene was constitutively expressed in plant cells by linking the optimized *cas9* to a maize Ubiquitin promoter (Christensen et al., 1992) and pinII terminator (An et al., 1989) in a plasmid DNA vector. To confer efficient sgRNA expression in maize cells, a maize U6 polymerase III promoter and terminator (TTTTTTTT) were isolated and fused to the 5' and 3' ends of a modified Blat sgRNA encoding DNA sequence, respectively.

To accurately compare the mutational efficiency resulting from the imperfect non-homologous end-joining (NHEJ) repair of DNA double-strand breaks (DSBs) resulting from Spy and Blat Cas9 cleavage, protospacer identical genomic target sites were selected by identifying targets with Spy and Blat Cas9 compatible PAMs, NGGNCNDD. Identical spacer sequences were selected for Blat and Spy Cas9 by capturing the 18 to 21 nt sequence immediately upstream of the PAM. To ensure optimal U6 polymerase III expression and not introduce a mismatch within the sgRNA spacer, all target sequences were selected to naturally terminate in a G at their 5' end. Targets were identified and selected in exon 1 and 4 of the maize fertility gene *Ms45* and in a region upstream of the maize *liguleless-1* gene.

The mutational activity of Blat Cas9 in maize was examined by biolistically transforming 10 day old immature maize embryos (IMEs) with DNA vectors containing Cas9 and guide RNAs. Blat and the equivalent Spy Cas9 and sgRNA expression vectors were independently introduced into maize Hi-Type II (Armstrong and Green, 1985) IMEs by particle gun transformation similar to that described in (An et al., 1989; Ananiev et al., 2009). Since particle gun transformation can be highly variable, a visual marker DNA expression cassette, Ds-Red, was also co-delivered with the Cas9 and sgRNA expression vectors to aid in the selection of evenly transformed IMEs. In total, 3 transformation replicates were performed on 60-90 IMEs and 20-30 of the most evenly transformed IMEs from each replicate were harvested 3 days after

transformation. Total genomic DNA was extracted and the region surrounding the target site was amplified by PCR and amplicons sequenced to a read depth in excess of 300,000. The resulting reads were examined for the presence of mutations at the expected site of cleavage by comparison to control experiments where the sgRNA DNA expression cassette was omitted from the transformation. As shown in Figure 25A, mutations were observed at the expected site of cleavage for Blat Cas9 with the most prevalent types of mutations being single base pair insertions or deletions. The mutational activity for Blat Cas9 was robust at 2 of the 3 sites tested and exceeded that of the Spy Cas9 at the *Ms45* Exon 4 target site by ~30% (Figure 25B).

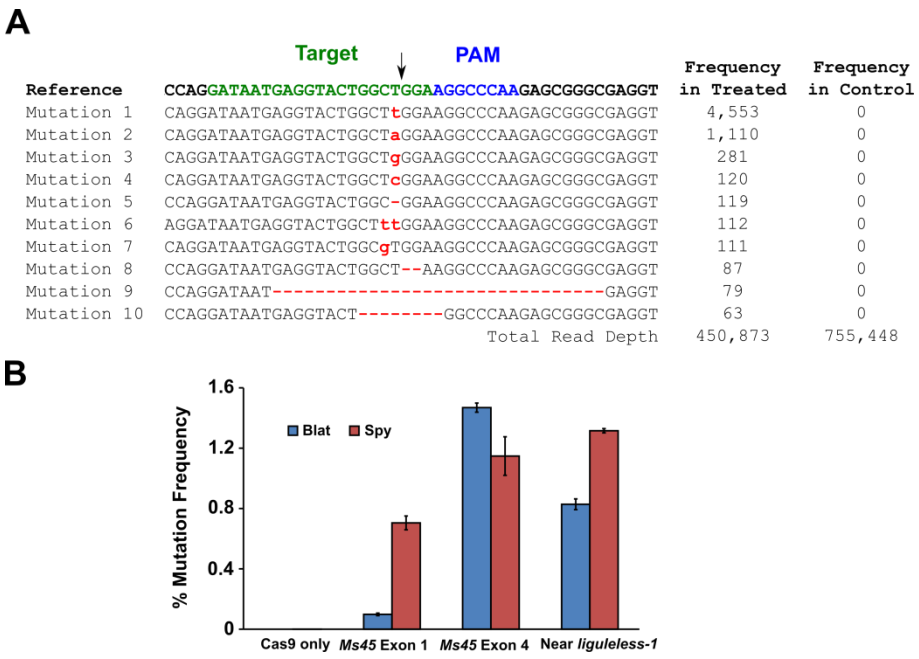


**Figure 25. *Brevibacillus laterosporus* Cas9 promotes NHEJ mutations in maize.** (**A**) Top 10 most prevalent types of NHEJ mutations detected with Blat Cas9 in exon 4 of the *Ms45* gene. A black arrow indicates the expected site of cleavage; mutations are highlighted in red; lower case font indicates an insertion; "-" indicates a deletion. (**B**) Comparison of Spy and Blat Cas9 NHEJ mutation frequencies at 3 proto-spacer identical target sites in maize. NHEJ mutations were detected by deep sequencing 3 days after transformation. Error bars represent SEM, n=3 particle gun transformations. Cas9 only is the negative control and represents the average (across all 3 target sites) background frequency of mutations resulting from PCR amplification and sequencing.

Taken together, the developed assays described here further refine Cas9 PAM discovery efforts by the use of recombinant Cas9 protein and reframe PAM specificity as being non-static and dependent on Cas9-guide RNA complex concentration. Proof of concept for the described methods is provided by identifying the PAM preferences of a novel Cas9 protein from *B. laterosporus* SSP360D4 and by demonstrating its functional activity in maize.

## 3.6. Final remarks

In this work we have characterized molecular components crucial for Type II CRISPR-Cas system activity and reconstituted active Cas9 complex from *Streptococcus thermophilus* CRISPR3-Cas system *in vitro*. These experiments allowed us not only to use Cas9 complex as a molecular tool for *in vitro* cloning and *in vivo* genome editing experiments, but also thoroughly analyze Cas9 complex DNA target recognition molecular mechanism. Identified PAM importance motivated us to develop PAM identification assay that allowed to characterize a novel Cas9 protein from *Brevibacillus laterosporous*.

The strategies and methods described here could be used not only for novel Cas9-based tools characterization that may have unique sequence recognition and enzymatic properties but also for evaluation of previously described Cas9 proteins (e.g. Cas9s with altered PAM recognition preferences). The offered approach for thorough identification of Cas9 complex components and inspection of target DNA requirements further contributes to the efforts of expanding Cas9-based genome editing toolbox.

# CONCLUSIONS

1. The functional Cas9 ribonucleopreotein complex of *S. thermophilus* CRISPR3-Cas system capable of cleaving target DNA can be assembled *in vitro* by mixing the Cas9 protein, tracrRNA and crRNA.

2. *In vitro* assembled Cas9 complex can be used as a molecular tool to cleave target DNA *in vitro*.

3. *In vitro* reconstituted Cas9 complex delivered by a chemical transfection agent can be used for genome editing in eukaryotic cells *in vivo*.

4. PAM sequence is used by Cas9 ribonucleoprotein complex as R-loop priming site for unidirectional crRNA hybridization to DNA target.

5. PAM characterization assay developed in this work allows identifying PAM preferences for Cas9 proteins.

6. Cas9 protein from *B. laterosporus* SSP360D4 recognizes NNNNCNDD PAM sequence and is functionally active in maize.

## LIST OF PUBLICATIONS

The thesis is based on the following original publications:

1. **Karvelis, T**., Gasiunas, G., Miksys, A., Barrangou, R., Horvath, P., and Siksnys, V. (2013). crRNA and tracrRNA guide Cas9-mediated DNA interference in *Streptococcus thermophilus*. RNA Biol. 10, 20–19.
2. **Karvelis, T**., Gasiunas, G., and Siksnys, V. (2013). Programmable DNA cleavage *in vitro* by Cas9. Biochem. Soc. Trans. 41, 1401–1406.
3. Szczelkun, M.D., Tikhomirova, M.S., Sinkunas, T., Gasiunas, G., **Karvelis, T**., Pschera, P., Siksnys, V., and Seidel, R. (2014). Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes. Proc. Natl. Acad. Sci. U. S. A. 111, 9798–9803.
4. Glemzaite, M., Balciunaite, E., **Karvelis, T**., Gasiunas, G., Grusyte, M.M., Alzbutas, G., Jurcyte, A., Anderson, E.M., Maksimova, E., Smith, A.J., et al. (2015). Targeted gene editing by transfection of *in vitro* reconstituted *Streptococcus thermophilus* Cas9 nuclease complex. RNA Biol. 12, 1–4.
5. **Karvelis, T**., Gasiunas, G., Young, J., Bigelyte, G., Silanskas, A., Cigan, M., and Siksnys, V. (2015). Rapid characterization of CRISPR-Cas9 protospacer adjacent motif sequence elements. Genome Biol. 16, 253.

## PATENT APPLICATIONS

1. Cigan A. M., Gasiunas G., **Karvelis T**., Siksnys V., Young J. K. BB2475-US-PSP2, 62/196535. 2015.07.24.
2. Cigan A. M., Gasiunas G., **Karvelis T**., Siksnys V., Young J. K. BB2539-US-PSP, 62/162377. 2015.05.15.

3. Cigan A. M., Gasiunas G., **Karvelis T**., Siksnys V., Young J. K. BB2475-US-PSP, 62/162353. 2015.05.15.
4. Siksnys V., Gasiunas G., **Karvelis T**., Lubys A., Zaliauskiene L., Glemzaite M., Smith A. WO/2013/142578, PCT/US2013/033106. 2013.03.20.
5. Šikšnys V., Gasiūnas G., **Karvelis T**. WO/2013/141680, PCT/LS2013/000006. 2013.03.15.

## CONFERENCE PRESENTATIONS

**Poster presentations:**

1. **Karvelis T**., Gasiunas G., Miksys A., Barrangou R., Horvath P. and Siksnys V. Cas9-mediated DNA interference in *Streptococcus thermophilus*: composition and assembly of the silencing complex. CRISPR: Evolution, Mechanisms and Infection; University of St Andrews, UK; 2013.06.17-19.
2. Gasiunas G., **Karvelis T**. and Siksnys V. Cas9 – a programmable DNA endonuclease for *in vitro* applications. CRISPR: Evolution, Mechanisms and Infection; University of St Andrews, UK; 2013.06.17-19.
3. Gasiunas G., **Karvelis T**., Barrangou R., Horvath P. and Siksnys V. Cas9 – a programmable RNA-guided DNA endonuclease from the bacterial adaptive immune system. 13th Young Scientists Forum (YSF); Saint Petersburg, Russia; 2013.07.03-06.
4. **Karvelis T**., Gasiunas G., Miksys A., Grusyte M. M., Siksnys V. Molecular determinants of DNA cleavage by Cas9 endonuclease. CRISPR 2014 - The Prokaryotic Immune System CRISPR/Cas; Berlin, Germany; 2014.05.13-16.
5. Szczelkun M. D., Gasiunas G., **Karvelis T**., Seidel R., Siksnys V. Direct observation of R-loop formation by RNA-guided Cas9 effector complex.

CRISPR 2014 - The Prokaryotic Immune System CRISPR/Cas; Berlin, Germany; 2014.05.13-16.

6. Gasiunas G, **Karvelis T**, Glemzaite M, Zaliauskiene L and Siksnys V. Cas9 is able to replace the restriction enzymes. 7th NEB Meeting on DNA Restriction and Modification. Gdansk, Poland; 2015.08.24‑29.
7. **Karvelis T**., Gasiunas G., Young J., Bigelyte G., Silanskas A., Cigan M. and Siksnys V. Rapid characterization of CRISPR-Cas9 protospacer adjacent motif sequence elements. CRISPR 2016; Rehovot, Israel; 2016.05.23-25.

**Oral presentation:**

1. **Karvelis T**., Gasiunas G., Young J., Bigelyte G., Silanskas A., Cigan M. and Siksnys V. Rapid characterization of CRISPR-Cas9 protospacer adjacent motif (PAM) sequence elements. XIVth Conference of the Lithuanian Biochemical Society; Druskininkai, Lithuania; 2016.06.28-30.

## FINANCIAL SUPPORT

The work was supported by the European Social Fund under the Global Grant measure (project R100) and Agency for Science, Innovation and Technology (MITA) grants (#31v-13 and #31v-36). Scholarships for academic achievements were received from the Research Council of Lithuania.

## ACKNOWLEDGEMENTS

I am grateful to my supervisor prof. dr. Virginijus Šikšnys for the opportunity to work in this interesting and dynamic field and also for invaluable discussions, suggestions and help with the project described in this doctoral thesis.

My special thanks go to dr. Giedrius Gasiūnas for the useful discussions and the help, that continue till now.

I would like to thank to prof. dr. M. D. Szczelkun (University of Bristol) for performing magnetic tweezers experiments, M. Glemžaitė and dr. L. Zaliauskienė (Thermo Fisher Scientific Baltics) for performing cell transfection experiments, J. Young and dr. M. Cigan (Dupont Pioneer) for the help with developing PAM identification assay, deep sequencing of the samples and genome editing experiments in maize cells, and for all co-authors of the publications, that were involved in the work described in this thesis.

I am grateful to all my colleagues in the Department of Protein - DNA Interactions for discussions, advice and help with experiments.

I also thank to dr. Rolandas Meškys and dr. Saulius Serva for the critical reading and comments on this doctoral thesis.

# REFERENCES


1. Abudayyeh, O.O., Gootenberg, J.S., Konermann, S., Joung, J., Slaymaker, I.M., Cox, D.B.T., Shmakov, S., Makarova, K.S., Semenova, E., Minakhin, L., et al. (2016). C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. Science *353*, aaf5573.
2. An, G., Mitra, A., Choi, H.K., Costa, M.A., An, K., Thornburg, R.W., and Ryan, C.A. (1989). Functional analysis of the 3' control region of the potato wound-inducible proteinase inhibitor II gene. Plant Cell *1*, 115–122.
3. Ananiev, E. V., Wu, C., Chamberlin, M.A., Svitashev, S., Schwartz, C., Gordon-Kamm, W., and Tingey, S. (2009). Artificial chromosome formation in maize (*Zea mays* L.). Chromosoma *118*, 157–177.
4. Anders, C., Niewoehner, O., Duerst, A., and Jinek, M. (2014). Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. Nature *513*, 569–573.
5. Anders, C., Bargsten, K., and Jinek, M. (2016). Structural Plasticity of PAM Recognition by Engineered Variants of the RNA-Guided Endonuclease Cas9. Mol. Cell *61*, 895–902.
6. Armstrong, C.L., and Green, C.E. (1985). Establishment and maintenance of friable, embryogenic maize callus and the involvement of L-proline. Planta *164*, 207–214.
7. Barrangou, R. (2012). RNA-mediated programmable DNA cleavage. Nat. Biotechnol. *30*, 836–838.
8. 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.
9. Baxter, S., Lambert, A.R., Kuhar, R., Jarjour, J., Kulshina, N., Parmeggiani, F., Danaher, P., Gano, J., Baker, D., Stoddard, B.L., et al. (2012). Engineering domain fusion chimeras from I-OnuI family LAGLIDADG homing endonucleases. Nucleic Acids Res. *40*, 7985–8000.
10. Bibikova, M., Carroll, D., Segal, D.J., Trautman, J.K., Smith, J., Kim, Y.G., and Chandrasegaran, S. (2001). Stimulation of homologous recombination through targeted cleavage by chimeric nucleases. Mol. Cell. Biol. *21*, 289–297.
11. Bibikova, M., Golic, M., Golic, K.G., and Carroll, D. (2002). Targeted chromosomal cleavage and mutagenesis in Drosophila using zinc-finger nucleases. Genetics *161*, 1169–1175.
12. Bibikova, M., Beumer, K., Trautman, J.K., and Carroll, D. (2003). Enhancing gene targeting with designed zinc finger nucleases. Science *300*, 764.
13. Bikard, D., and Marraffini, L.A. (2013). Control of gene expression by

CRISPR-Cas systems. F1000Prime Rep *5*, 47.
14. Bitinaite, J., Wah, D.A., Aggarwal, A.K., and Schildkraut, I. (1998). FokI dimerization is required for DNA cleavage. Proc. Natl. Acad. Sci. *95*, 10570–10575.
15. Boch, J., and Bonas, U. (2010). Xanthomonas AvrBs3 Family-Type III Effectors: Discovery and Function. Annu. Rev. Phytopathol. *48*, 419–436.
16. Boch, J., Scholze, H., Schornack, S., Landgraf, A., Hahn, S., Kay, S., Lahaye, T., Nickstadt, A., and Bonas, U. (2009). Breaking the code of DNA binding specificity of TAL-type III effectors. Science *326*, 1509–1512.
17. Bolotin, A., Quinquis, B., Sorokin, A., and Ehrlich, S.D. (2005). Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology *151*, 2551–2561.
18. Bonas, U., Stall, R.E., and Staskawicz, B. (1989). Genetic and structural characterization of the avirulence gene *avrBs3* from *Xanthomonas campestris* pv. *vesicatoria*. Mol. Gen. Genet. *218*, 127–136.
19. 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.
20. Brutzer, H., Luzzietti, N., Klaue, D., and Seidel, R. (2010). Energetics at the DNA Supercoiling Transition. Biophys. J. *98*, 1267–1276.
21. Capecchi, M. (1989). Altering the genome by homologous recombination. Science *244*, 1288–1292.
22. Carroll, D. (2011). Genome engineering with zinc-finger nucleases. Genetics *188*, 773–782.
23. Carroll, D. (2014). Genome engineering with targetable nucleases. Annu. Rev. Biochem. *83*, 409–439.
24. Chang, N., Sun, C., Gao, L., Zhu, D., Xu, X., Zhu, X., Xiong, J.-W., and Xi, J.J. (2013). Genome editing with RNA-guided Cas9 nuclease in Zebrafish embryos. Cell Res. *23*, 465–472.
25. Chen, B., Gilbert, L. a, Cimini, B. a, Schnitzbauer, J., Zhang, W., Li, G.-W., Park, J., Blackburn, E.H., Weissman, J.S., Qi, L.S., et al. (2013). Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. Cell *155*, 1479–1491.
26. Chevalier, B.S., and Stoddard, B.L. (2001). Homing endonucleases: structural and functional insight into the catalysts of intron/intein mobility. Nucleic Acids Res. *29*, 3757–3774.
27. Cho, S.W., Kim, S., Kim, Y., Kweon, J., Kim, H.S., Bae, S., and Kim, J.-S. (2014). Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. Genome Res. *24*, 132–141.
28. Choo, Y., and Klug, A. (1994). Toward a code for the interactions of zinc fingers with DNA: selection of randomized fingers displayed on phage. Proc. Natl. Acad. Sci. *91*, 11163–11167.
29. Choo, Y., Sánchez-García, I., and Klug, A. (1994). *In vivo* repression by

a site-specific DNA-binding protein designed against an oncogenic sequence. Nature *372*, 642–645.
30. Choulika, A., Perrin, A., Dujon, B., and Nicolas, J.F. (1995). Induction of homologous recombination in mammalian chromosomes by using the I-SceI system of *Saccharomyces cerevisiae*. Mol. Cell. Biol. *15*, 1968–1973.
31. Christensen, A.H., Sharrock, R.A., and Quail, P.H. (1992). Maize polyubiquitin genes: structure, thermal perturbation of expression and transcript splicing, and promoter activity following transfer to protoplasts by electroporation. Plant Mol. Biol. *18*, 675–689.
32. Christian, M., Cermak, T., Doyle, E.L., Schmidt, C., Zhang, F., Hummel, A., Bogdanove, A.J., and Voytas, D.F. (2010). Targeting DNA Double-Strand Breaks with TAL Effector Nucleases. Genetics *186*, 757–761.
33. Cohen, S.N., Chang, A.C., Boyer, H.W., and Helling, R.B. (1973). Construction of biologically functional bacterial plasmids *in vitro*. Proc. Natl. Acad. Sci. *70*, 3240–3244.
34. Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., et al. (2013). Multiplex Genome Engineering Using CRISPR/Cas Systems. Science *339*, 819–823.
35. Crooks, G.E. (2004). WebLogo: A Sequence Logo Generator. Genome Res. *14*, 1188–1190.
36. Crosetto, N., Mitra, A., Silva, M.J., Bienko, M., Dojer, N., Wang, Q., Karaca, E., Chiarle, R., Skrzypczak, M., Ginalski, K., et al. (2013). Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing. Nat. Methods *10*, 361–365.
37. Davis, K.M., Pattanayak, V., Thompson, D.B., Zuris, J.A., and Liu, D.R. (2015). Small molecule–triggered Cas9 protein with improved genome-editing specificity. Nat. Chem. Biol. *11*, 316–318.
38. 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.
39. Deng, D., Yan, C., Pan, X., Mahfouz, M., Wang, J., Zhu, J.-K., Shi, Y., and Yan, N. (2012). Structural Basis for Sequence-Specific Recognition of DNA by TAL Effectors. Science *335*, 720–723.
40. Desjarlais, J.R., and Berg, J.M. (1992). Toward rules relating zinc finger protein sequences and DNA binding site preferences. Proc. Natl. Acad. Sci. *89*, 7345–7349.
41. 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.
42. DiCarlo, J.E., Norville, J.E., Mali, P., Rios, X., Aach, J., and Church, G.M. (2013). Genome engineering in *Saccharomyces cerevisiae* using CRISPR-Cas systems. Nucleic Acids Res. *41*, 4336–4343.
43. Doyon, Y., Vo, T.D., Mendel, M.C., Greenberg, S.G., Wang, J., Xia,

D.F., Miller, J.C., Urnov, F.D., Gregory, P.D., and Holmes, M.C. (2011). Enhancing zinc-finger-nuclease activity with improved obligate heterodimeric architectures. Nat. Methods *8*, 74–79.

44. Dong, D., Ren, K., Qiu, X., Zheng, J., Guo, M., Guan, X., Liu, H., Li, N., Zhang, B., Yang, D., et al. (2016). The crystal structure of Cpf1 in complex with CRISPR RNA. Nature *532*, 522–526.
45. Esvelt, K.M., Mali, P., Braff, J.L., Moosburner, M., Yaung, S.J., and Church, G.M. (2013). Orthogonal Cas9 proteins for RNA-guided gene regulation and editing. Nat. Methods *10*, 1116–1121.
46. Fairall, L., Schwabe, J.W.R., Chapman, L., Finch, J.T., and Rhodes, D. (1993). The crystal structure of a two zinc-finger peptide reveals an extension to the rules for zinc-finger/DNA recognition. Nature *366*, 483–487.
47. Fonfara, I., Le Rhun, A., Chylinski, K., Makarova, K.S., Lécrivain, A.L., Bzdrenga, J., Koonin, E. V., and Charpentier, E. (2014). Phylogeny of Cas9 determines functional exchangeability of dual-RNA and Cas9 among orthologous type II CRISPR-Cas systems. Nucleic Acids Res. *42*, 2577–2590.
48. Fonfara, I., Richter, H., Bratovič, M., Le Rhun, A., and Charpentier, E. (2016). The CRISPR-associated DNA-cleaving enzyme Cpf1 also processes precursor CRISPR RNA. Nature *532*, 517–521.
49. Forth, S., Deufel, C., Sheinin, M.Y., Daniels, B., Sethna, J.P., and Wang, M.D. (2008). Abrupt Buckling Transition Observed during the Plectoneme Formation of Individual DNA Molecules. Phys. Rev. Lett. *100*, 148301.
50. Frock, R.L., Hu, J., Meyers, R.M., Ho, Y.-J., Kii, E., and Alt, F.W. (2014). Genome-wide detection of DNA double-stranded breaks induced by engineered nucleases. Nat. Biotechnol. *33*, 179–186.
51. Fu, Y., Foden, J.A., Khayter, C., Maeder, M.L., Reyon, D., Joung, J.K., and Sander, J.D. (2013). High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. Nat. Biotechnol. *31*, 822–826.
52. Fu, Y., Reyon, D., and Keith Joung, J. (2014). Targeted Genome Editing in Human Cells Using CRISPR/Cas Nucleases and Truncated Guide RNAs. In Methods in Enzymology, (Elsevier Inc.), pp. 21–45.
53. Gabriel, R., Lombardo, A., Arens, A., Miller, J.C., Genovese, P., Kaeppel, C., Nowrouzi, A., Bartholomae, C.C., Wang, J., Friedman, G., et al. (2011). An unbiased genome-wide analysis of zinc-finger nuclease specificity. Nat. Biotechnol. *29*, 816–823.
54. Gaj, T., Gersbach, C.A., and Barbas, C.F. (2013). ZFN, TALEN and CRISPR/Cas-based methods for genome engineering. Trends Biotechnol. *31*, 397–405.
55. Gao, H., Smith, J., Yang, M., Jones, S., Djukanovic, V., Nicholson, M.G., West, A., Bidney, D., Falco, S.C., Jantz, D., et al. (2010). Heritable targeted mutagenesis in maize using a designed endonuclease. Plant J. *61*, 176–187.
56. Garneau, J.E., Dupuis, M.-È., Villion, M., Romero, D. a, Barrangou, R.,

Boyaval, P., Fremaux, C., Horvath, P., Magadán, A.H., and Moineau, S. (2010). The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature *468*, 67–71.

57. 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. *109*, 2579–2586.
58. Gasiunas, G., Sinkunas, T., and Siksnys, V. (2014). Molecular mechanisms of CRISPR-mediated microbial immunity. Cell. Mol. Life Sci. *71*, 449–465.
59. Gilbert, L.A., Larson, M.H., Morsut, L., Liu, Z., Brar, G.A., Torres, S.E., Stern-Ginossar, N., Brandman, O., Whitehead, E.H., Doudna, J.A., et al. (2013). CRISPR-Mediated Modular RNA-Guided Regulation of Transcription in Eukaryotes. Cell *154*, 442–451.
60. Gimble, F.S., Moure, C.M., and Posey, K.L. (2003). Assessing the Plasticity of DNA Target Site Recognition of the PI-SceI Homing Endonuclease Using a Bacterial Two-hybrid Selection System. J. Mol. Biol. *334*, 993–1008.
61. Glemzaite, M., Balciunaite, E., Karvelis, T., Gasiunas, G., Grusyte, M.M., Alzbutas, G., Jurcyte, A., Anderson, E.M., Maksimova, E., Smith, A.J., et al. (2015). Targeted gene editing by transfection of *in vitro* reconstituted *Streptococcus thermophilus* Cas9 nuclease complex. RNA Biol. *12*, 1–4.
62. Grizot, S., Epinat, J.-C., Thomas, S., Duclert, A., Rolland, S., Paques, F., and Duchateau, P. (2010). Generation of redesigned homing endonucleases comprising DNA-binding domains derived from two different scaffolds. Nucleic Acids Res. *38*, 2006–2018.
63. Guilinger, J.P., Thompson, D.B., and Liu, D.R. (2014). Fusion of catalytically inactive Cas9 to FokI nuclease improves the specificity of genome modification. Nat. Biotechnol. *32*, 577–582.
64. Guschin, D.Y., Waite, A.J., Katibah, G.E., Miller, J.C., Holmes, M.C., and Rebar, E.J. (2010). A rapid and general assay for monitoring endogenous gene modification. Methods Mol. Biol. *649*, 247–256.
65. Hale, C., Kleppe, K., Terns, R.M., and Terns, M.P. (2008). Prokaryotic silencing (psi)RNAs in Pyrococcus furiosus. RNA *14*, 2572–2579.
66. Hale, C.R., Zhao, P., Olson, S., Duff, M.O., Graveley, B.R., Wells, L., Terns, R.M., and Terns, M.P. (2009). RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. Cell *139*, 945–956.
67. Heler, R., Samai, P., Modell, J.W., Weiner, C., Goldberg, G.W., Bikard, D., and Marraffini, L.A. (2015). Cas9 specifies functional viral targets during CRISPR-Cas adaptation. Nature *519*, 199–202.
68. Hilton, I.B., D'Ippolito, A.M., Vockley, C.M., Thakore, P.I., Crawford, G.E., Reddy, T.E., and Gersbach, C.A. (2015). Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. Nat. Biotechnol. *33*, 510–517.
69. Hirano, H., Gootenberg, J.S., Horii, T., Abudayyeh, O.O., Kimura, M., Hsu, P.D., Nakane, T., Ishitani, R., Hatada, I., Zhang, F., et al. (2016a).

Structure and Engineering of *Francisella novicida* Cas9. Cell *164*, 950–961.
70. Hirano, S., Nishimasu, H., Ishitani, R., and Nureki, O. (2016b). Structural Basis for the Altered PAM Specificities of Engineered CRISPR-Cas9. Mol. Cell *61*, 886–894.
71. Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, L.S., Tacker, M., and Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. Monatshefte Fur Chemie Chem. Mon. *125*, 167–188.
72. Horvath, P., and Barrangou, R. (2010). CRISPR/Cas, the immune system of bacteria and archaea. Science *327*, 167–170.
73. 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.
74. Howan, K., Smith, A.J., Westblade, L.F., Joly, N., Grange, W., Zorman, S., Darst, S.A., Savery, N.J., and Strick, T.R. (2012). Initiation of transcription-coupled repair characterized at single-molecule resolution. Nature *490*, 431–434.
75. Hsu, P.D., Scott, D.A., Weinstein, J.A., Ran, F.A., Konermann, S., Agarwala, V., Li, Y., Fine, E.J., Wu, X., Shalem, O., et al. (2013). DNA targeting specificity of RNA-guided Cas9 nucleases. Nat. Biotechnol. *31*, 827–832.
76. Hwang, W.Y., Fu, Y., Reyon, D., Maeder, M.L., Tsai, S.Q., Sander, J.D., Peterson, R.T., Yeh, J.-R.J., and Joung, J.K. (2013). Efficient genome editing in zebrafish using a CRISPR-Cas system. Nat. Biotechnol. *31*, 227–229.
77. Yamano, S., Dai, J., and Moursi, A.M. (2010). Comparison of transfection efficiency of nonviral gene transfer reagents. Mol. Biotechnol. *46*, 287–300.
78. Yamano, T., Nishimasu, H., Zetsche, B., Hirano, H., Slaymaker, I.M., Li, Y., Fedorova, I., Nakane, T., Makarova, K.S., Koonin, E.V., et al. (2016). Crystal Structure of Cpf1 in Complex with Guide RNA and Target DNA. Cell *165*, 949–962.
79. Isalan, M., Choo, Y., and Klug, A. (1997). Synergy between adjacent zinc fingers in sequence-specific DNA recognition. Proc. Natl. Acad. Sci. *94*, 5617–5621.
80. Isalan, M., Klug, A., and Choo, Y. (1998). Comprehensive DNA recognition through concerted interactions from adjacent zinc fingers. Biochemistry *37*, 12026–12033.
81. Isalan, M., Klug, A., and Choo, Y. (2001). No Title. Nat. Biotechnol. *19*, 656–660.
82. 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.
83. Jaenisch, R. (1976). Germ line integration and Mendelian transmission of

the exogenous Moloney leukemia virus. Proc. Natl. Acad. Sci. *73*, 1260–1264.
84. Jaenisch, R. (1977). Germ line integration of moloney leukemia virus: effect of homozygosity at the m-mulV locus. Cell *12*, 691–696.
85. Jaenisch, R., and Mintz, B. (1974). Simian virus 40 DNA sequences in DNA of healthy adult mice derived from preimplantation blastocysts injected with viral DNA. Proc. Natl. Acad. Sci. *71*, 1250–1254.
86. 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.
87. Jiang, F., Zhou, K., Ma, L., Gressel, S., and Doudna, J.A. (2015). A Cas9-guide RNA complex preorganized for target DNA recognition. Science *348*, 1477–1481.
88. Jiang, F., Taylor, D.W., Chen, J.S., Kornfeld, J.E., Zhou, K., Thompson, A.J., Nogales, E., and Doudna, J.A. (2016). Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. Science *351*, 867–871.
89. Jiang, W., Bikard, D., Cox, D., Zhang, F., and Marraffini, L.A. (2013). RNA-guided editing of bacterial genomes using CRISPR-Cas systems. Nat. Biotechnol. *31*, 233–239.
90. 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.
91. Jinek, M., East, A., Cheng, A., Lin, S., Ma, E., and Doudna, J. (2013). RNA-programmed genome editing in human cells. Elife *2*, 1–9.
92. Jinek, M., Jiang, F., Taylor, D.W., Sternberg, S.H., Kaya, E., Ma, E., Anders, C., Hauer, M., Zhou, K., Lin, S., et al. (2014). Structures of Cas9 Endonucleases Reveal RNA-Mediated Conformational Activation. Science *343*, 1247997–1247997.
93. Joung, J.K., and Sander, J.D. (2013). TALENs: a widely applicable technology for targeted genome editing. Nat. Rev. Mol. Cell Biol. *14*, 49–55.
94. Jurica, M.S., and Stoddard, B.L. (1999). Homing endonucleases: structure, function and evolution. Cell. Mol. Life Sci. *55*, 1304–1326.
95. Jurica, M.S., Monnat, R.J., and Stoddard, B.L. (1998). DNA recognition and cleavage by the LAGLIDADG homing endonuclease I-CreI. Mol. Cell *2*, 469–476.
96. Karvelis, T., Gasiunas, G., Miksys, A., Barrangou, R., Horvath, P., and Siksnys, V. (2013). crRNA and tracrRNA guide Cas9-mediated DNA interference in *Streptococcus thermophilus*. RNA Biol. *10*, 20–19.
97. Karvelis, T., Gasiunas, G., Young, J., Bigelyte, G., Silanskas, A., Cigan, M., and Siksnys, V. (2015). Rapid characterization of CRISPR-Cas9 protospacer adjacent motif sequence elements. Genome Biol. *16*, 253.
98. Kauert, D.J., Kurth, T., Liedl, T., and Seidel, R. (2011). Direct Mechanical Measurements Reveal the Material Properties of Three-Dimensional DNA Origami. Nano Lett. *11*, 5558–5563.
99. 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.

100. Kearns, N.A., Pham, H., Tabak, B., Genga, R.M., Silverstein, N.J., Garber, M., and Maehr, R. (2015). Functional annotation of native enhancers with a Cas9-histone demethylase fusion. Nat. Methods *12*, 401–403.
101. Kim, Y.G., and Chandrasegaran, S. (1994). Chimeric restriction endonuclease. Proc. Natl. Acad. Sci. *91*, 883–887.
102. Kim, J.S., and Pabo, C.O. (1998). Getting a handhold on DNA: design of poly-zinc finger proteins with femtomolar dissociation constants. Proc. Natl. Acad. Sci. *95*, 2812–2817.
103. Kim, T.K., and Eberwine, J.H. (2010). Mammalian cell transfection: the present and the future. Anal. Bioanal. Chem. *397*, 3173–3178.
104. Kim, D., Bae, S., Park, J., Kim, E., Kim, S., Yu, H.R., Hwang, J., Kim, J.-I., and Kim, J.-S. (2015). Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells. Nat. Methods *12*, 237–243.
105. Kim, D., Kim, S., Kim, S., Park, J., and Kim, J.-S. (2016). Genome-wide target specificities of CRISPR-Cas9 nucleases revealed by multiplex Digenome-seq. Genome Res. *26*, 406–415.
106. Kim, Y.G., Cha, J., and Chandrasegaran, S. (1996). Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain. Proc. Natl. Acad. Sci. *93*, 1156–1160.
107. Kim, S., Kim, D., Cho, S.W., Kim, J., and Kim, J.S. (2014). Highly efficient RNA-guided genome editing in human cells via delivery of purified Cas9 ribonucleoproteins. Genome Res. *24*, 1012–1019.
108. Klaue, D., and Seidel, R. (2009). Torsional Stiffness of Single Superparamagnetic Microspheres in an External Magnetic Field. Phys. Rev. Lett. *102*, 28302.
109. Kleinstiver, B.P., Prew, M.S., Tsai, S.Q., Topkar, V. V, Nguyen, N.T., Zheng, Z., Gonzales, A.P.W., Li, Z., Peterson, R.T., Yeh, J.J., et al. (2015a). Engineered CRISPR-Cas9 nucleases with altered PAM specificities. Nature *523*, 481–485.
110. Kleinstiver, B.P., Prew, M.S., Tsai, S.Q., Nguyen, N.T., Topkar, V. V, Zheng, Z., and Joung, J.K. (2015b). Broadening the targeting range of *Staphylococcus aureus* CRISPR-Cas9 by modifying PAM recognition. Nat. Biotechnol. *33*, 1293–1298.
111. Kleinstiver, B.P., Pattanayak, V., Prew, M.S., Tsai, S.Q., Nguyen, N.T., Zheng, Z., and Joung, J.K. (2016). High-fidelity CRISPR–Cas9 nucleases with no detectable genome-wide off-target effects. Nature *529*, 490–495.
112. Klug, A. (2010). The Discovery of Zinc Fingers and Their Applications in Gene Regulation and Genome Manipulation. Annu. Rev. Biochem. *79*, 213–231.
113. Koike-Yusa, H., Li, Y., Tan, E.-P., Velasco-Herrera, M.D.C., and Yusa, K. (2013). Genome-wide recessive genetic screening in mammalian cells

with a lentiviral CRISPR-guide RNA library. Nat. Biotechnol. *32*, 267–273.
114. Komor, A.C., Kim, Y.B., Packer, M.S., Zuris, J.A., and Liu, D.R. (2016). Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. Nature *533*, 420–424.
115. Lanio, T., Jeltsch, A., and Pingoud, A. (2000). On the possibilities and limitations of rational protein design to expand the specificity of restriction enzymes: a case study employing EcoRV as the target. Protein Eng. *13*, 275–281.
116. Leenay, R.T., Maksimchuk, K.R., Slotkowski, R.A., Agrawal, R.N., Gomaa, A.A., Briner, A.E., Barrangou, R., Beisel, C.L., Anders, C., Niewoehner, O., et al. (2016). Identifying and Visualizing Functional PAM Diversity across CRISPR-Cas Systems. Mol. Cell *62*, 137–147.
117. Li, K., Wang, G., Andersen, T., Zhou, P., and Pu, W.T. (2014). Optimization of Genome Engineering Approaches with the CRISPR/Cas9 System. PLoS One *9*, e105779.
118. Li, L., Wu, L.P., and Chandrasegaran, S. (1992). Functional domains in Fok I restriction endonuclease. Proc. Natl. Acad. Sci. *89*, 4275–4279.
119. Li, T., Huang, S., Jiang, W.Z., Wright, D., Spalding, M.H., Weeks, D.P., and Yang, B. (2011). TAL nucleases (TALNs): hybrid proteins composed of TAL effectors and FokI DNA-cleavage domain. Nucleic Acids Res. *39*, 359–372.
120. Liang, X., Potter, J., Kumar, S., Zou, Y., Quintanilla, R., Sridharan, M., Carte, J., Chen, W., Roark, N., Ranganathan, S., et al. (2015). Rapid and highly efficient mammalian cell engineering via Cas9 protein transfection. J. Biotechnol. *208*, 44–53.
121. Libiakova, G., Jørgensen, B., Palmgren, G., Ulvskov, P., and Johansen, E. (2001). Efficacy of an intron-containing kanamycin resistance gene as a selectable marker in plant transformation. Plant Cell Rep. *20*, 610–615.
122. Lieber, M.R. (2010). The Mechanism of Double-Strand DNA Break Repair by the Nonhomologous DNA End-Joining Pathway. Annu. Rev. Biochem. *79*, 181–211.
123. Lin, S., Staahl, B.T., Alla, R.K., and Doudna, J. a (2014). Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery. Elife *3*, 1–13.
124. Lionnet, T., Allemand, J.-F., Revyakin, A., Strick, T.R., Saleh, O.A., Bensimon, D., and Croquette, V. (2012). Magnetic trap construction. Cold Spring Harb. Protoc. *2012*, 133–138.
125. Livak, K.J., and Schmittgen, T.D. (2001). Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the $2^{-\Delta\Delta CT}$ Method. Methods *25*, 402–408.
126. Maeder, M.L., Linder, S.J., Cascio, V.M., Fu, Y., Ho, Q.H., and Joung, J.K. (2013). CRISPR RNA–guided activation of endogenous human genes. Nat. Methods *10*, 977–979.
127. Mak, A.N.-S.N.-S., Bradley, P., Cernadas, R.A., Bogdanove, A.J., and Stoddard, B.L. (2012). The Crystal Structure of TAL Effector PthXo1

Bound to Its DNA Target. Science *335*, 716–719.
128. Makarova, K.S., Grishin, N. V, Shabalina, S.A., Wolf, Y.I., and Koonin, E. V (2006). No Title. Biol. Direct *1*, 7.
129. 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.
130. Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E., and Church, G.M. (2013a). RNA-Guided Human Genome Engineering via Cas9. Science *339*, 823–826.
131. Mali, P., Aach, J., Stranges, P.B., Esvelt, K.M., Moosburner, M., Kosuri, S., Yang, L., and Church, G.M. (2013b). CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. Nat. Biotechnol. *31*, 833–838.
132. Markham, N.R., and Zuker, M. (2008). UNAFold: software for nucleic acid folding and hybridization. Methods Mol Biol *453*, 3–31.
133. Marraffini, L.A., and Sontheimer, E.J. (2008). CRISPR Interference Limits Horizontal Gene Transfer in *Staphylococci* by Targeting DNA. Science *322*, 1843–1845.
134. Miller, J., McLachlan, A.D., and Klug, A. (1985). Repetitive zinc-binding domains in the protein transcription factor IIIA from *Xenopus* oocytes. EMBO J. *4*, 1609–1614.
135. Miller, J.C., Holmes, M.C., Wang, J., Guschin, D.Y., Lee, Y.-L., Rupniewski, I., Beausejour, C.M., Waite, A.J., Wang, N.S., Kim, K.A., et al. (2007). An improved zinc-finger nuclease architecture for highly specific genome editing. Nat. Biotechnol. *25*, 778–785.
136. Miller, J.C., Tan, S., Qiao, G., Barlow, K.A., Wang, J., Xia, D.F., Meng, X., Paschon, D.E., Leung, E., Hinkley, S.J., et al. (2011). A TALE nuclease architecture for efficient genome editing. Nat. Biotechnol. *29*, 143–148.
137. Moynahan, M.E., and Jasin, M. (2010). Mitotic homologous recombination maintains genomic stability and suppresses tumorigenesis. Nat. Rev. Mol. Cell Biol. *11*, 196–207.
138. Mojica, F.J.M., Juez, G., and Rodriguez-Valera, F. (1993). Transcription at different salinities of *Haloferax mediterranei* sequences adjacent to partially modified PstI sites. Mol. Microbiol. *9*, 613–621.
139. Mojica, F.J.M., Ferrer, C., Juez, G., and Rodríguez-Valera, F. (1995). Long stretches of short tandem repeats are present in the largest replicons of the Archaea *Haloferax mediterranei* and *Haloferax volcanii* and could be involved in replicon partitioning. Mol. Microbiol. *17*, 85–93.
140. Mojica, F.J.M., Diez-Villasenor, 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.
141. 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.
142. 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.
143. Moore, M., Klug, A., and Choo, Y. (2001). Improved DNA binding specificity from polyzinc finger peptides by using strings of two-finger units. Proc. Natl. Acad. Sci. *98*, 1437–1441.
144. Morrow, J.F., Cohen, S.N., Chang, A.C., Boyer, H.W., Goodman, H.M., and Helling, R.B. (1974). Replication and transcription of eukaryotic DNA in *Escherichia coli*. Proc. Natl. Acad. Sci. *71*, 1743–1747.
145. Mosconi, F., Allemand, J.F., Bensimon, D., and Croquette, V. (2009). Measurement of the Torque on a Single Stretched and Twisted DNA Using Magnetic Tweezers. Phys. Rev. Lett. *102*, 78301.
146. Moscou, M.J., and Bogdanove, A.J. (2009). A Simple Cipher Governs DNA Recognition by TAL Effectors. Science *326*, 1501–1501.
147. 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.
148. Nekrasov, V., Staskawicz, B., Weigel, D., Jones, J.D.G., and Kamoun, S. (2013). Targeted mutagenesis in the model plant *Nicotiana benthamiana* using Cas9 RNA-guided endonuclease. Nat. Biotechnol. *31*, 691–693.
149. Nelles, D.A., Fang, M.Y., O'Connell, M.R., Xu, J.L., Markmiller, S.J., Doudna, J.A., and Yeo, G.W. (2016). Programmable RNA Tracking in Live Cells with CRISPR/Cas9. Cell *165*, 488–496.
150. Nihongaki, Y., Yamamoto, S., Kawano, F., Suzuki, H., and Sato, M. (2015). CRISPR-Cas9-based Photoactivatable Transcription System. Chem. Biol. *22*, 169–174.
151. Nishimasu, H., Ran, F.A., Hsu, P.D., Konermann, S., Shehata, S.I., Dohmae, N., Ishitani, R., Zhang, F., and Nureki, O. (2014). Crystal structure of Cas9 in complex with guide RNA and target DNA. Cell *156*, 935–949.
152. Nishimasu, H., Cong, L., Yan, W.X., Ishitani, R., Zhang, F., and Correspondence, O.N. (2015). Crystal Structure of *Staphylococcus aureus* Cas9. Cell *162*, 1113–1126.
153. O'Connell, M.R., L. Oakes, B., Sternberg, S.H., East-Seletsky, A., Kaplan, M., and Doudna, J.A. (2014). Programmable RNA recognition and cleavage by CRISPR/Cas9. Nature *516*, 263–266.
154. Oberstrass, F.C., Fernandes, L.E., and Bryant, Z. (2012). Torque measurements reveal sequence-specific cooperative transitions in supercoiled DNA. Proc. Natl. Acad. Sci. *109*, 6106–6111.
155. van der Oost, J., Westra, E.R., Jackson, R.N., and Wiedenheft, B. (2014). Unravelling the structural and mechanistic basis of CRISPR–Cas systems. Nat. Rev. Microbiol. *12*, 479–492.
156. Osborn, M.J., Webber, B.R., Knipping, F., Lonetree, C.-L., Tennis, N., DeFeo, A.P., McElroy, A.N., Starker, C.G., Lee, C., Merkel, S., et al.

(2016). Evaluation of TCR Gene Editing Achieved by TALENs, CRISPR/Cas9, and megaTAL Nucleases. Mol. Ther. *24*, 570–581.
157. Otto, O., Czerwinski, F., Gornall, J.L., Stober, G., Oddershede, L.B., Seidel, R., and Keyser, U.F. (2010). Real-time particle tracking at 10,000 fps using optical fiber illumination. Opt. Express *18*, 22722.
158. Pâques, F., and Duchateau, P. (2007). Meganucleases and DNA double-strand break-induced recombination: perspectives for gene therapy. Curr. Gene Ther. *7*, 49–66.
159. Pattanayak, V., Lin, S., Guilinger, J.P., Ma, E., Doudna, J.A., and Liu, D.R. (2013). High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. Nat. Biotechnol. *31*, 839–843.
160. Pavletich, N., and Pabo, C. (1991). Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. Science *252*, 809–817.
161. Pelham, H.R., and Brown, D.D. (1980). A specific transcription factor that can bind either the 5S RNA gene or 5S RNA. Proc. Natl. Acad. Sci. *77*, 4170–4174.
162. Perez-Pinera, P., Ousterout, D.G., and Gersbach, C.A. (2012). Advances in targeted genome editing. Curr. Opin. Chem. Biol. *16*, 268–277.
163. Perez-Pinera, P., Kocak, D.D., Vockley, C.M., Adler, A.F., Kabadi, A.M., Polstein, L.R., Thakore, P.I., Glass, K.A., Ousterout, D.G., Leong, K.W., et al. (2013). RNA-guided gene activation by CRISPR-Cas9–based transcription factors. Nat. Methods *10*, 973–976.
164. Picard, B., and Wegnez, M. (1979). Isolation of a 7S particle from *Xenopus laevis* oocytes: a 5S RNA-protein complex. Proc. Natl. Acad. Sci. *76*, 241–245.
165. Porteus, M. (2016). Genome Editing: A New Approach to Human Therapeutics. Annu. Rev. Pharmacol. Toxicol. *56*, 163–190.
166. 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 *151*, 653–663.
167. Qi, L.S., Larson, M.H., Gilbert, L.A., Doudna, J.A., Weissman, J.S., Arkin, A.P., and Lim, W.A. (2013). Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression. Cell *152*, 1173–1183.
168. Ramirez, C.L., Foley, J.E., Wright, D.A., Müller-Lerch, F., Rahman, S.H., Cornu, T.I., Winfrey, R.J., Sander, J.D., Fu, F., Townsend, J.A., et al. (2008). Unexpected failure rates for modular assembly of engineered zinc fingers. Nat. Methods *5*, 374–375.
169. Ran, F.A., Cong, L., Yan, W.X., Scott, D. a., Gootenberg, J.S., Kriz, A.J., Zetsche, B., Shalem, O., Wu, X., Makarova, K.S., et al. (2015). *In vivo* genome editing using *Staphylococcus aureus* Cas9. Nature *520*, 186–190.
170. Ran, F.A.A., Hsu, P.D.D., Lin, C.-Y.Y., Gootenberg, J.S.S., Konermann, S., Trevino, A.E., Scott, D.A. a., Inoue, A., Matoba, S., Zhang, Y., et al. (2013). Double Nicking by RNA-Guided CRISPR Cas9 for Enhanced

Genome Editing Specificity. Cell *154*, 1–10.
171. Revyakin, A., Ebright, R.H., and Strick, T.R. (2005). Single-molecule DNA nanomanipulation: Improved resolution through use of shorter DNA fragments. Nat. Methods *2*, 127–138.
172. Rimseliene, R., Maneliene, Z., Lubys, A., and Janulaitis, A. (2003). Engineering of restriction endonucleases: using methylation activity of the bifunctional endonuclease Eco57I to select the mutant with a novel sequence specificity. J. Mol. Biol. *327*, 383–391.
173. Roberts, R.J. (2005). How restriction enzymes became the workhorses of molecular biology. Proc. Natl. Acad. Sci. *102*, 5905–5908.
174. Rouet, P., Smih, F., and Jasin, M. (1994). Introduction of double-strand breaks into the genome of mouse cells by expression of a rare-cutting endonuclease. Mol. Cell. Biol. *14*, 8096–8106.
175. Rudin, N., Sugarman, E., and Haber, J.E. (1989). Genetic and physical analysis of double-strand break repair and recombination in *Saccharomyces cerevisiae*. Genetics *122*, 519–534.
176. Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989). Molecular Cloning: a Laboratory Manual, 2nd edn., Cold Spring Harbor Laboratory Press, Cold Spring Harbor.
177. Samuelson, J.C., and Xu, S. (2002). Directed Evolution of Restriction Endonuclease BstYI to Achieve Increased Substrate Specificity. J. Mol. Biol. *319*, 673–683.
178. Sapranauskas, R., Gasiunas, G., Fremaux, C., Barrangou, R., Horvath, P., and Siksnys, V. (2011). The *Streptococcus thermophilus* CRISPR/Cas system provides immunity in Escherichia coli. Nucleic Acids Res. *39*, 9275–9282.
179. Segal, D.J., and Meckler, J.F. (2013). Genome Engineering at the Dawn of the Golden Age. Annu. Rev. Genomics Hum. Genet. *14*, 135–158.
180. Seidel, R., Bloom, J.G., van Noort, J., Dutta, C.F., Dekker, N.H., Firman, K., Szczelkun, M.D., and Dekker, C. (2005). Dynamics of initiation, termination and reinitiation of DNA translocation by the motor protein Eco R124I. EMBO J. *24*, 4188–4197.
181. Shah, S.A., Erdmann, S., Mojica, F.J.M., and Garrett, R.A. (2013). Protospacer recognition motifs. RNA Biol. *10*, 891–899.
182. Shalem, O., Sanjana, N.E., Hartenian, E., Shi, X., Scott, D.A., Mikkelsen, T.S., Heckl, D., Ebert, B.L., Root, D.E., Doench, J.G., et al. (2014). Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells. Science *343*, 84–87.
183. Shimizu, Y., Şöllü, C., Meckler, J.F., Adriaenssens, A., Zykovich, A., Cathomen, T., and Segal, D.J. (2011). Adding fingers to an engineered zinc finger nuclease can reduce activity. Biochemistry *50*, 5033–5041.
184. Shmakov, S., Abudayyeh, O.O., Makarova, K.S., Wolf, Y.I., Gootenberg, J.S., Semenova, E., Minakhin, L., Joung, J., Konermann, S., Severinov, K., et al. (2015). Discovery and Functional Characterization of Diverse Class 2 CRISPR-Cas Systems. Mol. Cell *60*, 385–397.
185. 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.
186. Sinkunas, T., Gasiunas, G., Waghmare, S.P., Dickman, M.J., Barrangou, R., Horvath, P., and Siksnys, V. (2013). *In vitro* reconstitution of Cascade-mediated CRISPR immunity in *Streptococcus thermophilus*. EMBO J. *32*, 385–394.
187. Slaymaker, I.M., Gao, L., Zetsche, B., Scott, D.A., Yan, W.X., and Zhang, F. (2016). Rationally engineered Cas9 nucleases with improved specificity. Science *351*, 84–88.
188. Smith, J. (2000). Requirements for double-strand cleavage by chimeric restriction enzymes with zinc finger DNA-recognition domains. Nucleic Acids Res. *28*, 3361–3369.
189. Smith, C., Gore, A., Yan, W., Abalde-Atristain, L., Li, Z., He, C., Wang, Y., Brodsky, R.A., Zhang, K., Cheng, L., et al. (2014). Whole-genome sequencing analysis reveals high specificity of CRISPR/Cas9 and TALEN-based genome editing in human iPSCs. Cell Stem Cell *15*, 12–13.
190. Smith, J., Grizot, S., Arnould, S., Duclert, A., Epinat, J.-C., Chames, P., Prieto, J., Redondo, P., Blanco, F.J., Bravo, J., et al. (2006). A combinatorial approach to create artificial homing endonucleases cleaving chosen sequences. Nucleic Acids Res. *34*, e149.
191. Smithies, O., Gregg, R.G., Boggs, S.S., Koralewski, M.A., and Kucherlapati, R.S. (1985). Insertion of DNA sequences into the human chromosomal beta-globin locus by homologous recombination. Nature *317*, 230–234.
192. 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.
193. Sternberg, S.H., LaFrance, B., Kaplan, M., and Doudna, J.A. (2015). Conformational control of DNA target cleavage by CRISPR–Cas9. Nature *527*, 110–113.
194. Steuer, S., Pingoud, V., Pingoud, A., and Wende, W. (2004). Chimeras of the homing endonuclease PI-SceI and the homologous Candida tropicalis intein: a study to explore the possibility of exchanging DNA-binding modules to obtain highly specific endonucleases with altered specificity. Chembiochem *5*, 206–213.
195. Stoddard, B.L. (2006). Homing endonuclease structure and function. Q. Rev. Biophys. *38*, 49.
196. Stoddard, B.L. (2011). Homing Endonucleases: From Microbial Genetic Invaders to Reagents for Targeted DNA Modification. Structure *19*, 7–15.
197. Stoddard, B.L. (2014). Homing endonucleases from mobile group I introns: discovery to genome engineering. Mob. DNA *5*, 7.
198. Stormo, G.D. (2013). Modeling the specificity of protein-DNA interactions. Quant. Biol. *1*, 115–130.
199. Sussman, D., Chadsey, M., Fauce, S., Engel, A., Bruett, A., Monnat, R.,

Stoddard, B.L., and Seligman, L.M. (2004). Isolation and characterization of new homing endonuclease specificities at individual target site positions. J. Mol. Biol. *342*, 31–41.
200. Svitashev, S., Young, J.K., Schwartz, C., Gao, H., Falco, S.C., and Cigan, A.M. (2015). Targeted Mutagenesis, Precise Gene Editing, and Site-Specific Gene Insertion in Maize Using Cas9 and Guide RNA. Plant Physiol. *169*, 931–945.
201. Szczelkun, M.D., Tikhomirova, M.S., Sinkunas, T., Gasiunas, G., Karvelis, T., Pschera, P., Siksnys, V., and Seidel, R. (2014). Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes. Proc. Natl. Acad. Sci. *111*, 9798–9803.
202. Szczepek, M., Brondani, V., Büchel, J., Serrano, L., Segal, D.J., and Cathomen, T. (2007). Structure-based redesign of the dimerization interface reduces the toxicity of zinc-finger nucleases. Nat. Biotechnol. *25*, 786–793.
203. Tamulaitis, G., Kazlauskiene, M., Manakova, E., Venclovas, Č., Nwokeoji, A.O., Dickman, M.J., Horvath, P., and Siksnys, V. (2014). Programmable RNA Shredding by the Type III-A CRISPR-Cas System of *Streptococcus thermophilus*. Mol. Cell *56*, 506–517.
204. Tsai, S.Q., and Joung, J.K. (2016). Defining and improving the genome-wide specificities of CRISPR–Cas9 nucleases. Nat. Rev. Genet. *17*, 300–312.
205. Tsai, S.Q., Zheng, Z., Nguyen, N.T., Liebers, M., Topkar, V. V, Thapar, V., Wyvekens, N., Khayter, C., Iafrate, A.J., Le, L.P., et al. (2014a). GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. Nat. Biotechnol. *33*, 187–197.
206. Tsai, S.Q., Wyvekens, N., Khayter, C., Foden, J.A., Thapar, V., Reyon, D., Goodwin, M.J., Aryee, M.J., and Joung, J.K. (2014b). Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. Nat. Biotechnol. *32*, 569–576.
207. Urnov, F.D., Rebar, E.J., Holmes, M.C., Zhang, H.S., and Gregory, P.D. (2010). Genome editing with engineered zinc finger nucleases. Nat. Rev. Genet. *11*, 636–646.
208. Veres, A., Gosis, B.S., Ding, Q., Collins, R., Ragavendran, A., Brand, H., Erdin, S., Cowan, C.A., Talkowski, M.E., and Musunuru, K. (2014). Low incidence of off-target mutations in individual CRISPR-Cas9 and TALEN targeted human stem cell clones detected by whole-genome sequencing. Cell Stem Cell *15*, 27–30.
209. Vojta, A., Dobrinić, P., Tadić, V., Bočkor, L., Korać, P., Julg, B., Klasić, M., and Zoldoš, V. (2016). Repurposing the CRISPR-Cas9 system for targeted DNA methylation. Nucleic Acids Res. *44*, 5615–5628.
210. Wah, D.A., Bitinaite, J., Schildkraut, I., and Aggarwal, A.K. (1998). Structure of FokI has implications for DNA cleavage. Proc. Natl. Acad. Sci. *95*, 10564–10569.
211. Wang, H., Yang, H., Shivalila, C.S., Dawlaty, M.M., Cheng, A.W., Zhang, F., and Jaenisch, R. (2013). One-step generation of mice carrying

mutations in multiple genes by CRISPR/cas-mediated genome engineering. Cell *153*, 910–918.
212. Wang, T., Wei, J.J., Sabatini, D.M., and Lander, E.S. (2014). Genetic Screens in Human Cells Using the CRISPR-Cas9 System. Science *343*, 80–84.
213. Wang, X., Wang, Y., Wu, X., Wang, J., Wang, Y., Qiu, Z., Chang, T., Huang, H., Lin, R.-J., and Yee, J.-K. (2015). Unbiased detection of off-target cleavage by CRISPR-Cas9 and TALENs using integrase-defective lentiviral vectors. Nat. Biotechnol. *33*, 175–178.
214. Wei, Y., Terns, R.M., and Terns, M.P. (2015). Cas9 function and host genome sampling in Type II-A CRISPR–Cas adaptation. Genes Dev. *29*, 356–361.
215. Westra, E.R.R., van Erp, P.B.G., Künne, T., Wong, S.P.P., Staals, R.H.J.H.J., Seegers, C.L.C.L.C., Bollen, S., Jore, M.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.
216. Wyman, C., and Kanaar, R. (2006). DNA Double-Strand Break Repair: All's Well that Ends Well. Annu. Rev. Genet. *40*, 363–383.
217. Wright, A. V, Nunez, J.K., and Doudna, J.A. (2016). Review Biology and Applications of CRISPR Systems: Harnessing Nature's Toolbox for Genome Engineering. Cell *164*, 29–44.
218. Xie, K., Zhang, J., and Yang, Y. (2014). Genome-wide prediction of highly specific guide RNA spacers for CRISPR-Cas9-mediated genome editing in model plants and major crops. Mol. Plant *7*, 923–926.
219. Zaremba, M., Sasnauskas, G., Urbanke, C., and Siksnys, V. (2006). Allosteric Communication Network in the Tetrameric Restriction Endonuclease Bse634I. J. Mol. Biol. *363*, 800–812.
220. Zetsche, B., Gootenberg, J.S., Abudayyeh, O.O., Slaymaker, I.M., Makarova, K.S., Essletzbichler, P., Volz, S.E., Joung, J., van der Oost, J., Regev, A., et al. (2015a). Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. Cell *163*, 759–771.
221. Zetsche, B., Volz, S.E., and Zhang, F. (2015b). A split-Cas9 architecture for inducible genome editing and transcription modulation. Nat. Biotechnol. *33*, 139–142.
222. Zhou, Y., Zhu, S., Cai, C., Yuan, P., Li, C., Huang, Y., and Wei, W. (2014). High-throughput screening of a CRISPR/Cas9 library for functional genomics in human cells. Nature *509*, 487–491.
223. Zuris, J.A., Thompson, D.B., Shu, Y., Guilinger, J.P., Bessen, J.L., Hu, J.H., Maeder, M.L., Joung, J.K., Chen, Z.-Y., and Liu, D.R. (2014). Cationic lipid-mediated delivery of proteins enables efficient protein-based genome editing *in vitro* and *in vivo*. Nat. Biotechnol. *33*, 73–80.
224.