(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
4 December 2014 (04.12.2014)

WIPO | PCT

(10) International Publication Number
WO 2014/191521 A2

(51) International Patent Classification: *C12N 9/22* (2006.01)

(21) International Application Number: PCT/EP2014/061181

(22) International Filing Date: 28 May 2014 (28.05.2014)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
PA201370297 29 May 2013 (29.05.2013) DK
PA201370771 13 December 2013 (13.12.2013) DK

(71) Applicant: **CELLECTIS** [FR/FR]; 8 Rue de la Croix Jarry, F-75013 Paris (FR).

(72) Inventors: **DUCHATEAU, Philippe**; Bateau Fawen, Quai des Dames, F-91210 Draveil (FR). **BERTONATI, Claudia**; 18, rue des Gravilliers, F-75003 Paris (FR).

(74) Agent: **ZACCO DENMARK A/S**; Hans Bekkevolds Allé 7, DK-2900 Hellerup (DK).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— *without international search report and to be republished upon receipt of that report (Rule 48.2(g))*

— *with sequence listing part of description (Rule 5.2(a))*

(54) Title: NEW COMPACT SCAFFOLD OF CAS9 IN THE TYPE II CRISPR SYSTEM

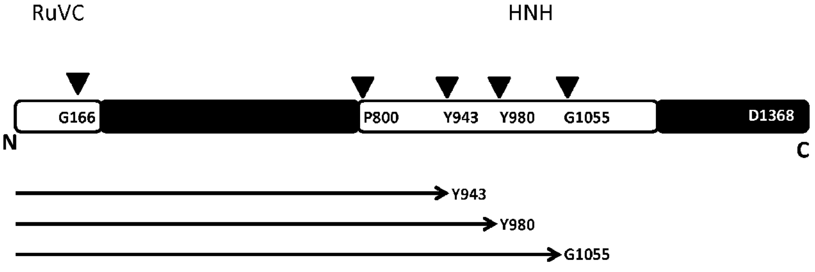

Figure 1

(57) **Abstract**: The present invention is in the field of CRISPR-Cas system for genome targeting. The present invention relates to new engineered Cas9 scaffolds and uses thereof. More particularly, the present invention relates to methods for genome targeting, cell engineering and therapeutic application. The present invention also relates to vectors, compositions and kits in which the new Cas9 scaffolds of the present invention are used.

NEW COMPACT SCAFFOLD OF CAS9 IN THE TYPE II CRISPR SYSTEM

FIELD OF THE INVENTION

The present invention is in the field of CRISPR-Cas system for genome targeting. The present invention relates to new engineered Cas9 scaffolds and uses thereof. More particularly, the present invention relates to methods for genome targeting, cell engineering and therapeutic application. The present invention also relates to vectors, compositions and kits in which the new Cas9 scaffolds of the present invention are used.

BACKGROUND OF THE INVENTION

Site-specific nucleases are powerful reagents for specifically and efficiently targeting and modifying a DNA sequence within a complex genome. There are numerous applications of genome engineering by site-specific nucleases extending from basic research to bioindustrial applications and human therapeutics. Re-engineering a DNA-binding protein for this purpose has been mainly limited to the design and production of proteins such as the naturally occurring LADLIDADG homing endonucleases (LHE), artificial zinc finger proteins (ZFP), and Transcription Activator-Like Effectors nucleases (TALE-nucleases).

Recently, a new genome engineering tool has been developed based on the RNA-guided Cas9 nuclease (Gasiunas, Barrangou et al. 2012; Jinek, Chylinski et al. 2012) from the type II prokaryotic CRISPR (Clustered Regularly Interspaced Short palindromic Repeats) adaptive immune system. The CRISPR Associated (Cas) system was first discovered in bacteria and functions as a defense against foreign DNA, either viral or plasmid. So far three distinct bacterial CRISPR systems have been identified, termed type I, II and III. The Type II system is the basis for the current genome engineering technology available and is often simply referred to as CRISPR. The type II CRISPR/Cas loci are composed of an operon of genes encoding the proteins Cas9, Cas1, Cas2 and/or Csn2, a CRISPR array consisting of a leader sequence followed by identical repeats interspersed with unique genome-targeting spacers and a sequence encoding the *trans*-activating tracrRNA.

CRISPR-mediated adaptative immunity proceeds in three distinct stages: acquisition of foreign DNA, CRISPR RNA (crRNA) biogenesis and target interference (see for review (Sorek, Lawrence et al. 2013)). First, the CRISPR/Cas machinery appears to target specific sequence for integration into the CRISPR locus. Sequences in foreign DNA selected for integration are called spacers and these sequences are often flanked by a short sequence motif, referred as the proto-spacer adjacent motif (PAM). crRNA biogenesis in type II systems is unique in that it requires a trans-activating crRNA (tracRNA). CRISPR locus is initially transcribed as long precursor crRNA (pre-crRNA) from a promoter sequence in the leader. Cas9 acts as a molecular anchor facilitating the base pairing of tracRNA with pre-cRNA for subsequent recognition and cleavage of pre-cRNA repeats by the host RNase III (Deltcheva, Chylinski et al. 2011). Following the processing events, tracrRNA remains paired to the crRNA and bound to the Cas9 protein. In this ternary complex, the dual tracrRNA:crRNA structure acts as guide RNA that directs the endonuclease Cas9 to the cognate target DNA. Target recognition by the Cas9-tracrRNA:crRNA complex is initiated by scanning the invading DNA molecule for homology between the protospacer sequence in the target DNA and the spacer-derived sequence in the crRNA. In addition to the DNA protospacer-crRNA spacer complementarity, DNA targeting requires the presence of a short motif adjacent to the protospacer (protospacer adjacent motif - PAM). Following pairing between the dual-RNA and the protospacer sequence, Cas9 subsequently introduces a blunt double strand break 3 bases upstream of the PAM motif (Garneau, Dupuis et al. 2010).

Cas9 is a large endonuclease capable of recognizing any potential target of 12 to 20 nucleotides and a specific PAM motif currently restricted to 2 nucleotides (NGG; (Mali, Yang et al. 2013)). The potential target is enough for ensuring unique cleavage site in prokaryotic genomes on a statistical basis, but is critical for larger genomes, like in eukaryotic cells, where potential target sequences may be found several times. There is therefore a need to develop strategies for improving specificity and reducing potential off-site using type II CRISPR system. Moreover, the large size of the natural Cas 9 (>1200 amino acids) is a disadvantage in gene delivery for genome engineering CRISPR system.

In order to improve gene delivery of Cas9 into cells, the present inventors have designed new Cas9 scaffolds including RuvC motif as defined by (D-[I/L]-G-X-X-S-X-G-W-A) (SEQ ID NO: 1) and/or HNH motif as defined by (Y-X-X-D-H-X-X-P-X-S-X-X-X-D-X-S) (SEQ ID NO: 2), wherein X represents any one of 20 natural amino acids and [I/L] represents isoleucine or leucine. These compact scaffolds were obtained by searching for the presence of the above putative motifs in genome databases and identifying those present on separate ORFs. The inventors made the presumption that if such motifs were found on separate subunit proteins, shorter proteins could be identified and fused together to obtain shorter functional fusion proteins.

By pursuing this strategy, the inventors have been able to determine the boundaries of the RuvC and HNH domains and to design new shorter Cas9 derived from the *S. pyogenes* or homologues thereof. Their Cas9 homologues analysis further allowed the identification of previously uncharacterized Cas9 residues involved in the binding of the guide RNA and the PAM motif. By engineering these domains, the inventors increase the number of target nucleotides specifically recognized by type II CRISPR system to avoid off-site target.

SUMMARY OF THE INVENTION

The present invention provides with new RuvC and HNH sequence motifs to be combined with each other to result into more compact and/or more specific recombinant Cas9 scaffolds (i.e. artificial fusion proteins of less than 1100 amino acids). Cas9 protein can be divided into two separate split Cas9 RuvC and HNH domains which can process target nucleic acid sequence together or separately with guideRNA. These scaffolds are used in methods for gene targeting, in particular as specific nucleases for gene editing. Expression vectors encoding these new scaffolds and the cells transformed and engineered with these vectors are also the subject-matter of the invention.

BRIEF DESCRIPTION OF THE TABLES AND THE FIGURES

Table 1: Multiple sequence alignment of RuvC domain of Cas9 homologues

Table 2: Secondary structure predictions for the RuvC domain and amino acids sequence of the RuvC domain of the *S. pyogenes* Cas9 (SEQ ID NO: 12).

Table 3: Multiple sequence alignment of HNH domains of Cas9 homologues.

Table 4: Secondary structure predictions for the HNH domain and related HNH domain sequence of the *S. pyogenes* Cas9 (SEQ ID NO: 23).

Table 5: Multiple sequence alignment of shorter Cas9 homologues

Table 6: Secondary structure predictions of shorter Cas9 versions and related shorter *S. pyogenes* Cas9 sequence.

Table 7: List of DNA/RNA binding regions of *S.pyogenes* Cas9.

Table 8: Multiple sequence alignment between Cas9 of *S. pyogenes* (SEQ ID NO: 61) and *S.thermophilus* (SEQ ID NO: 64) and the sequence of two pdb structures of RuvC domain of *E.coli* and *T. thermophilus* (SEQ ID NO: 62 and SEQ ID NO: 63).

Table 9: Multiple sequence alignment of the eight select sequences with Cas9 wild type of S. Pyogenes and Cas9 of *S. Thermophilus* and 4EP4 pdbcode. The position of the G247 is marked by a black arrow.

Table 10: Secondary structure elements prediction for the Cas9 wild type of *S. Pyogenes* sequence using PSIPRED. The sequence has been divided into the two split domains: N-terminal and C-terminal domain. In bold is marked the Leucine 248 which has been mutated to Valine in the sequence of the C-terminal domain.

Figure 1. The original sequence of *S. pyogenes* Cas9 and the proposed truncation Y943, Y980, G1055.

Figure 2 and 3: Fifteen DNA/RNA binding regions mapped in the 3D model of the sequence of *S. pyogenes.*

Figure 4: Nuclease activity of the split Cas9 domains measured as a reduction in GFP by flow cytometry (Macsquant) at day 4 and day 7 post-transfection. The values are reported for each single split domains or for the two co-transfected split domains.

Figure 5: Nuclease activity of the split Cas9 domains on GFP target tested using EndoT7 assay.

Figure 6: Nuclease activity of the split Cas9 domains on CD52 target gene tested using EndoT7 assay.

DISCLOSURE OF THE INVENTION

Unless specifically defined herein, all technical and scientific terms used have the same meaning as commonly understood by a skilled artisan in the fields of gene therapy, biochemistry, genetics, molecular biology and immunology.

All methods and materials similar or equivalent to those described herein can be used in the practice or testing of the present invention, with suitable methods and materials being described herein. All publications, patent applications, patents, and other references mentioned herein are incorporated by reference in their entirety. In case of conflict, the present specification, including definitions, will prevail. Further, the materials, methods, and examples are illustrative only and are not intended to be limiting, unless otherwise specified.

The practice of the present invention will employ, unless otherwise indicated, conventional techniques of cell biology, cell culture, molecular biology, transgenic biology, microbiology, recombinant DNA, and immunology, which are within the skill of the art. Such techniques are explained fully in the literature. See, for example, Current Protocols in Molecular Biology (Frederick M. AUSUBEL, 2000, Wiley and son Inc, Library of Congress, USA); Molecular Cloning: A Laboratory Manual, Third Edition, (Sambrook et al, 2001, Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press); Oligonucleotide Synthesis (M. J. Gait ed., 1984); Mullis et al. U.S. Pat. No. 4,683,195; Nucleic Acid Hybridization (B. D. Harries & S. J. Higgins eds. 1984); Transcription And Translation (B. D. Hames & S. J. Higgins eds. 1984); Culture Of Animal Cells (R. I. Freshney, Alan R. Liss, Inc., 1987); Immobilized Cells And Enzymes (IRL Press, 1986); B. Perbal, A Practical Guide To Molecular Cloning (1984); the series, Methods In ENZYMOLOGY (J. Abelson and M. Simon, eds.-in-chief, Academic Press, Inc., New York), specifically, Vols.154 and 155 (Wu et al. eds.) and Vol. 185, "Gene Expression Technology" (D. Goeddel, ed.); Gene Transfer Vectors For Mammalian Cells (J. H. Miller and M. P. Calos eds., 1987, Cold Spring Harbor Laboratory); Immunochemical Methods In Cell And Molecular Biology (Mayer and Walker, eds., Academic Press, London, 1987); Handbook Of Experimental Immunology, Volumes I-IV (D. M. Weir and C. C. Blackwell, eds., 1986); and Manipulating the Mouse Embryo, (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y., 1986).

New Cas 9 variants

Cas9, also named Csn1 (COG3513) is a large protein that participates in both crRNA biogenesis and in the destruction of invading DNA. Cas9 has been described in different bacterial species such as *S. thermophilus* (Sapranauskas NAR 2011), *listeria innocua* (jinek Science 2012) and *S. pyogenes* (Deltcheva, Chylinski et al. 2011). The large Cas9 protein (>1200 amino acids) contains two predicted nuclease domains, namely HNH (McrA-like) nuclease domain that is located in the middle of the protein and a split RuvC-like nuclease domain (RNase H fold) (Haft, Selengut et al. 2005; Makarova, Grishin et al. 2006). The insertion of the HNH nuclease domain into the RNAse H fold suggests that the two nuclease activities are closely coupled. Recently, it has been demonstrated that HNH domain is responsible for nicking of one strand of the target double-stranded DNA and the RuvC-like RNase H fold domain is involved in cleavage of the other strand of the double-stranded DNA target (Jinek, Chylinski et al. 2012). Together, these two domains each nick a strand of the target DNA within the proto-spacer in the immediate vicinity of the PAM, which results in blunt cleavage of the invasive DNA (Jinek, Chylinski et al. 2012). According to the present invention, a compact Cas9 variant is an endonuclease comprising less than 1100, preferably less than 1000, more preferably less than 900 amino acids, again more preferably less than 800 amino acids encoding RuvC and HNH domains.

By “Cas 9 variant” is meant an engineering endonuclease or a homologue of Cas9 which is capable of binding dual crRNA:tracRNA (or a single guide RNA) which acts as a guide RNA that directs the Cas9 to the nucleic acid target. In particular embodiment, Cas9 variants can induce a cleavage in the nucleic acid target sequence which can correspond to either a double-stranded break or a single-stranded break. Cas9 variant can be a Cas9 endonuclease that does not naturally exist in nature and that is obtained by genetic engineering or by random mutagenesis. Cas9 variants according to the invention can for example be obtained by mutations i.e. deletions from, or insertions or substitutions of at least one residue in the amino acid sequence of a *S. pyogenes* Cas9 endonuclease (SEQ ID NO: 3). In the frame aspects of the present invention, such Cas9 variants remain functional, i.e. they retain the capacity of binding dual crRNA:tracRNA (or a single guide RNA). Cas9 variant can also be homologues of *S. pyogenes* Cas9 which can comprise deletions from, or insertions or substitutions of, at least one residue within the amino acid sequence of *S. pyogenes* Cas9 (SEQ ID NO: 3). Any combination of deletion, insertion, and substitution may also be made to arrive at the final construct, provided that the final construct possesses the desired activity, in particular the capacity of binding dual crRNa:tracRNA (or a single guide RNA) or nucleic acid target sequence.

RuvC/RNaseH motif includes proteins that show wide spectra of nucleolytic functions, acting both on RNA and DNA (RNaseH, RuvC, DNA transposases and retroviral integrases and PIWI domain of Argonaut proteins). In the present invention the RuvC catalytic domain of the Cas9 protein can be characterized by the sequence motif: D-[I/L]-G-X-X-S-X-G-W-A, wherein X represents any one of the natural 20 amino acids and [I/L] represents isoleucine or leucine (SEQ ID NO: 1). In other terms, the present invention relates to Cas9 variant which comprises at least D-[I/L]-G-X-X-S-X-G-W-A sequence, wherein X represents any one of the natural 20 amino acids and [I/L] represents isoleucine or leucine (SEQ ID NO: 1).

The characterization of the RuvC motif mentioned above allows to extract different homologues of Cas9 RuvC domain. The comparison of smaller RuvC homologues domains (SEQ ID NO: 5 to SEQ ID NO: 12, and SEQ ID NO: 51) with *S. pyogenes* Cas9 allows to determine the boundaries of the ruvC domain in *S. pyogenes* Cas9 (SEQ ID NO: 4). Thus, in a particular embodiment, the Cas9 variant comprises a RuvC domain which comprises the amino acid sequence selected from the group consisting of: SEQ ID NO: 4 to SEQ ID NO: 12 and SEQ ID NO: 51. The multiple sequence alignment of Cas9 homologues allow to determine the optimal breaking position (G247) for the *S. pyogenes* Cas9 sequence. Thus, the RuvC domain can correspond to the amino acid sequence comprising residues from position 1 to position 247 (SEQ ID NO: 52) or aligned positions using CLUSTALW method on homologues of Cas family members.

HNH motif is characteristic of many nucleases that act on double-stranded DNA including colicins, restriction enzymes and homing endonucleases. The domain HNH (SMART ID: SM00507, SCOP nomenclature:HNH family) is associated with a range of DNA binding proteins, performing a variety of binding and cutting functions (Gorbalenya 1994; Shub, Goodrich-Blair et al. 1994). Several of the proteins are hypothetical or putative proteins of no well-defined function. The ones with known function are involved in a range of cellular processes including bacterial toxicity, homing functions in groups I and II introns and inteins, recombination, developmentally controlled DNA rearrangement, phage packaging, and restriction endonuclease activity (Dalgaard, Klar et al. 1997). These proteins are found in viruses, archaebacteria, eubacteria, and eukaryotes. Interestingly, as with the LAGLI-DADG and the GIY-YIG motifs, the HNH motif is often associated with endonuclease domains of self-propagating elements like inteins, Group I, and Group II introns (Gorbalenya 1994; Dalgaard, Klar et al. 1997). The HNH domain can be characterized by the presence of a conserved Asp/His residue flanked by conserved His (amino-terminal) and His/Asp/Glu (carboxy-terminal) residues at some distance. A substantial number of these proteins can also have a CX2C motif on either side of the central Asp/His residue. Structurally, the HNH motif appears as a central hairpin of twisted β-strands, which are flanked on each side by an α helix (Kleanthous, Kuhlmann et al. 1999). In the present invention, the HNH motif can be characterized by the sequence motif: Y-X-X-D-H-X-X-P-X-S-X-X-X-D-X-S, wherein X represents any one of the natural 20 amino acids (SEQ ID NO: 2). The present invention relates to a Cas9 variant which comprises at least Y-X-X-D-H-X-X-P-X-S-X-X-X-D-X-S sequence wherein X represents any one of the natural 20 amino acids (SEQ ID NO: 2).

The minimal region of the HNH domain and the different homologues of HNH domain characterized in this study can be used to engineer a Cas9 variant. Thus, the present invention relates to a Cas9 variant which comprises a HNH domain comprising amino acid sequences selected from SEQ ID NO: 13 to SEQ ID NO: 22. The multiple sequence alignment of Cas9 homologues allow to determine the optimal breaking position (G247) for the *S. pyogenes* Cas9 sequence. Thus, the HNH domain can correspond to the amino acid sequence comprising residues from position 248 to position 1368 (SEQ ID NO: 53) or aligned positions using CLUSTALW method on homologues of Cas family members.

The alignment of *S. pyogenes* Cas9 and homologues members suggests that C-terminal region of Cas9 are dispensable. Thus, C-terminal domain of Cas9 is truncated after the HNH motif Y-X-X-D-H-X-X-P-X-S-X-X-X-D-X-S, preferably between 1 to 1000 amino acid residues after the HNH motif, more preferably between 1 to 500, more preferably between 1 to 250 amino acids after the HNH motif. More particularly, Cas9 variant comprises a HNH domain comprising the amino acid sequence selected from the group consisting of: SEQ ID NO: 23 to 25.

In another approach, the inventors identified four natural Cas9 homologues with shorter sequence and determined shorter version of *S. pyogenes* Cas9. Thus, the present invention also relates to Cas 9 which comprises amino acid sequences selected from the group consisting of SEQ ID NO: 26 to SEQ ID NO: 33.

In a particular embodiment, the Cas9 of the present invention comprises Y-X-X-D-H-X-X-P-X-S-X-X-X-D-X-S sequence and D-[I/L]-G-X-X-S-X-G-W-A wherein X represents any one of the natural 20 amino acids. More particularly, the Cas9 comprises a RuvC domain comprising an amino acid sequence selected from the group consisting of: SEQ ID NO: 4 to SEQ ID NO: 12 and SEQ ID NO:

51, and a HNH domain comprising an amino acid sequence selected from the group consisting of: SEQ ID NO: 13 to SEQ ID NO: 25.

In a more particular embodiment, said RuvC domain and HNH domain as described above is separated by a peptide domain. This peptide domain can be as non limiting example a non-specific linker ((GS)n) as well as small domains (i.e. Immonuglobulin domain, TPR, pumilo, RRM fold). In a particular embodiment, said peptide domain comprises an amino acid sequence selected from the group consisting of SEQ ID NO: 49 and SEQ ID NO: 50.

The above characterization of the RuvC and HNH domains prompted the inventors to engineer Cas9 protein to create split Cas9 protein. Cas9 protein has been divided into two separate RuvC and HNH domains. Surprisingly, the inventors showed that these two split Cas9 could process together or separately the nucleic acid target (see example 4). This observation allows developing a new Cas9 system using split Cas9 proteins. Each Cas9 domains as described above can be prepared and used separately. Thus, this split system displays several advantages for vectorization, allowing to deliver shorter protein than the entire Cas9, protein purification and protein engineering, particularly to engineer region responsible of PAM recognition, DNA binding.

By “Split Cas9” is meant here a reduced or truncated form of a Cas9 protein or Cas9 variant, which comprises either a RuvC or HNH domain, but not both of these domains. Such “Split Cas9” can be used independently with guide RNA or in a complementary fashion, like for instance, one Split Cas9 providing a RuvC domain and another providing the HNH domain. Different split Cas9 may be used together having either RuvC and/or NHN domains. Split Cas9 are preferably less than 1000 amino acids long, more preferably less than 800, even more preferably less than 500 amino acids long.

RuvC domain generally comprises at least an amino acid sequence D-[I/L]-G-X-X-S-X-G-W-A, wherein X represents any one of the natural 20 amino acids and [I/L] represents isoleucine or leucine (SEQ ID NO: 1). HNH domain generally comprises at least an amino acid sequence Y-X-X-D-H-X-X-P-X-S-X-X-X-D-X-S sequence, wherein X represents any one of the natural 20 amino acids (SEQ ID NO: 2).

In a preferred embodiment said split cas9 protein comprises a RuvC domain comprising an amino acid sequence selected from the group consisting of: SEQ ID NO: 4 to SEQ ID NO: 12 and SEQ ID NO: 51 and 53, and a HNH domain comprising an amino acid sequence selected from the group consisting of: SEQ ID NO: 13 to SEQ ID NO: 25 and 52, preferably a RuvC domain comprising an amino acid sequence SEQ ID NO: 52 and an HNH domain comprising an amino acid sequence SEQ ID NO: 53. In a preferred embodiment, said HNH domain comprises a first amino acid Leucine mutated in Valine in SEQ ID NO: 53 to have a better kozak consensus sequence.

Each Cas9 split domain can be derived from different Cas9 homologues, or can be derived from the same Cas9. Each split domain can be fused to at least one active domain in the N-terminal and/or C-terminal end, said active domain can be selected from the group consisting of: nuclease (e.g. endonuclease or exonuclease), polymerase, kinase, phosphatase, methylase, demethylase, acetylase, desacetylase, topoisomerase, integrase, transposase, ligase, helicase, recombinase, transcriptional activator(e.g. VP64, VP16), transcriptional inhibitor (e. g; KRAB), DNA end processing enzyme (e.g. Trex2, Tdt), reporter molecule (e.g. fluorescent proteins, LacZ, luciferase).

In a particular embodiment, said split domains can be fused to an energy acceptor and the complementary split domain to an energy donor such that the emission spectrum of the fluorescent molecule energy donor overlaps with the absorption spectrum of the energy acceptor the energy. When split Cas9 domains binds DNA together and when energy donor and acceptor are closed to each other, FRET (Fluorescence resonance energy transfer) occurs and results in reduction of the intensity of donor emission, as energy from the donor in its excited state is transferred to the acceptor.

In another particular embodiment, said Cas9 split domains are separated by a linker capable of inactivating the resulting protein. Addition of a specific small molecule changing the conformational structure of the split domains induces their activity. In another particular embodiment, said linker can comprise a protease cleavage site (e.g. HIV1 protease cleavage site). In the presence of a specific protease, the linker is cleaved and the resulting isolated RuvC and HNH domains can bind the target nucleic acid. Thus, the use of said RuvC and HNH domain linked together is particularly suitable to induce Cas9 activation at the desired time

In another aspect of the invention, to modulate Cas9 specificity, the inventors identified the residues involved in the binding of PAM motif and crRNA. Thus, the invention encompasses a Cas9 variant or split Cas9 domain which comprises at least one mutated amino acid residue in the nucleic acid binding region of *S. pyogenes* Cas9, preferably in amino acid sequence selected from the group consisting of SEQ ID NO: 34 to SEQ ID NO: 48.

Cas9 homologues domains identified in the present invention can also be engineered. The DNA/RNA binding region of Cas9 homologues can be determined by the multiple alignment sequences of example 1 and 2 (grey highlighted sequences in Tables 1, 3 and 5). Thus, the invention relates to a Cas9 variant, or split Cas9 domain which comprises at least one mutated amino acid residue in the nucleic acid binding region as described above. Said split Cas9 domains can be derived from different Cas9 homologues or variant according to the present invention.

In a particular aspect, this Cas9 variant can be able to bind a smaller or larger PAM motif which comprises combinations of any one of 20 natural amino acids (non natural PAM motif). Preferably, the Cas9 variant or split Cas9 domain according to the invention is capable of specifically recognizing a PAM motif which comprises at least 3, preferably 4, more preferably 5 nucleotides. The capacity of Cas9 to bind a PAM motif within the genomic DNA, in absence of crRNA (or guide RNA) can present a toxic effect when Cas9 is overexpressed in the cell. Thus, to avoid this potential toxic effect, the inventors sought to engineer Cas9 variant or split Cas9 domain which are not capable of binding a PAM motif. The Cas9 variant or split Cas9 domain according to the present invention comprises at least one amino acid residue in the PAM binding region, preferably in the region from residue T38 to E57 and/or from T146 to L169 of the SEQ ID NO: 3 or aligned positions using CLUSTALW method on homologues of Cas family members.

In another aspect, the Cas9 variant or split Cas9 domain may also be able to induce the binding of a smaller or larger complementary sequence of guide RNA on the nucleic acid target sequence.

Because some variability may arise from the genomic data from which these polypeptides derive, and also to take into account the possibility to substitute some of the amino acids present in these polypeptides without significant loss of activity (functional variants), the invention encompasses polypeptides variants of the above polypeptides that share at least 70%, preferably at least 80 %, more preferably at least 90 % and even more preferably at least 95 % identity with the sequences provided in this patent application. The present invention is thus drawn to polypeptides comprising a polypeptide sequence that has at least 70%, preferably at least 80%, more preferably at least 90 %, 95 % 97 % or 99 % sequence identity with amino acid sequence selected from the group consisting of SEQ ID NO: 3 to SEQ ID NO: 53.

Recently, it has been demonstrated that HNH domain is responsible for nicking of one strand of the target double-stranded DNA and the RuvC-like RNaseH fold domain is involved in cleavage of the other strand of the double-stranded DNA target (Jinek, Chylinski et al. 2012). Together, these two domains each nick a strand of the target DNA within the proto-spacer in the immediate vicinity of the PAM, which results in blunt cleavage of the invasive DNA (Jinek, Chylinski et al. 2012). In particular embodiment, Cas9 variant lacks one nickase activity. In particular, Cas9 variant or split Cas9 comprises inactivating mutation(s) in the catalytic residues of either the HNH or RuvC-like domains. This resulting Cas9 or split Cas9 is known to function as a nickase and induce a single-strand break in the target nucleic acid sequence. As non limiting example, the catalytic residues of the Cas9, protein or split Cas9 domain can be the D10, D31, H840, H868, N882 and N891 of SEQ ID NO: 3 or aligned positions using CLUSTALW method on homologues of Cas family members. The residues comprised in HNH or RuvC motifs can be those described in the above paragraph. Any one of these residues can be replaced by any other amino acids, preferably by alanine residue. Mutation in the catalytic residues means either substitution by another amino acids, or deletion or addition of amino acids that induce the inactivation of at least one of the catalytic domain of cas9 (Sapranauskas, Gasiunas et al. 2011; Jinek, Chylinski et al. 2012). In a particular embodiment, the Cas9 variant comprises only one of the two RuvC and HNH catalytic domains. In a particular embodiment, isolated RuvC and/or HNH domain can comprise inactivation mutation in the catalytic residues as described above.

In another aspect of the present invention, Cas9 lacks endonucleolytic activity. The resulting Cas9 is co-expressed with guide RNA designed to comprises a complementary sequence to a target nucleic acid sequence. Expression of Cas9 lacking endonucleolytic activity yields to specific silencing of the gene of interest. This system is named CRISPR interference (CRISPRi) (Qi, Larson et al. 2013). By silencing, it is meant that the gene of interest is not expressed in a functional protein form. The silencing may occur at the transcriptional or the translational step. According to the present invention, the silencing may occur by directly blocking transcription, more particularly by blocking transcription elongation or by targeting key cis-acting motifs within any promoter, sterically blocking the association of their cognate trans-acting transcription factors. The Cas9 lacking endonucleolytic activity comprises both non-functional HNH and RuvC domains. In particular, the Cas9 polypeptide comprises inactivating mutations in the catalytic residues of both the RuvC-like and HNH domains. For example, the catalytic residues required for cleavage Cas9 activity can be the D10, D31, H840, H865, H868, N882 and N891 of SEQ ID NO: 3 or aligned positions using CLUSTALW method on homologues of Cas Family members. The residues comprised in HNH or RuvC motifs can be those described in the above paragraph. Any of these residues can be replaced by any one of the other amino acids, preferably by alanine residue. Mutation in the catalytic residues means either substitution by another amino acids, or deletion or addition of amino acids that induce the inactivation of at least one of the catalytic domain of cas9.

The invention also concerns the polynucleotides, in particular DNA or RNA encoding the polypeptides and proteins previously described. These polynucleotides may be included in vectors, more particularly plasmids or virus, in view of being expressed in prokaryotic or eukaryotic cells.

The present invention contemplates modification of the Cas9, split Cas9 polynucleotide sequence such that the codon usage is optimized for the organism in which it is being introduced. Thus, for example Cas9 polynucleotide sequence derived from the *pyogenes or S. Thermophilus* codon optimized for use in human is set forth in (Cong, Ran et al. 2013; Mali, Yang et al. 2013).

In particular embodiments, the Cas9, split Cas9 polynucleotides according to the present invention can comprise at least one subcellular localization motif. A subcellular localization motif refers to a sequence that facilitates transporting or confining a protein to a defined subcellular location that includes at least one of the nucleus, cytoplasm, plasma membrane, endoplasmic reticulum, golgi apparatus, endosomes, peroxisomes and mitochondria. Subcellular localization motifs are well-known in the art. A subcellular localization motif requires a specific orientation, e.g., N- and/or C-terminal to the protein. As a non-limiting example, the nuclear localization signal (NLS) of the simian virus 40 large T-antigen can be oriented at the N and/or C-terminus. NLS is an amino acid sequence which acts to target the protein to the cell nucleus through Nuclear Pore Complex and to direct a newly synthesized protein into the nucleus via its recognition by cytosolic nuclear transport receptors. Typically, a NLS consists of one or more short sequences of positively charged amino acids such as lysines or arginines.

In particular embodiments, the polynucleotide encoding a cas9 variant or a split Cas9 according to the present invention is placed under the control of a promoter. Suitable promoters include tissue specific and/or inducible promoters. Tissue specific promoters control gene expression in a tissue-dependent manner and according to the developmental stage of the cell. The transgenes driven by these type of promoters will only be expressed in tissues where the transgene product is desired, leaving the rest of the tissues unmodified by transgene expression. Tissue-specific promoters may be induced by endogenous or exogenous factors, so they can be classified as inducible promoters as well. An inducible promoter is a promoter which initiates transcription only when it is exposed to some particular (typically external) stimulus. Particularly preferred for the present invention are: a light-regulated promoter, nitrate reductase promoter, eukaryotic metallothionine promoter, which is induced by increased levels of heavy metals, prokaryotic lacZ promoter which is induced in response to isopropyl-β-D-thiogalacto-pyranoside (IPTG), steroid-responsive promoter, tetracycline-dependent promoter and eukaryotic heat shock promoter which is induced by increased temperature.

Method of genome targeting

In another aspect, the present invention relates to a method for use of said polypeptides and/or polynucleotides according to the invention for various applications ranging from targeted nucleic acid cleavage to targeted gene regulation. In genome engineering experiments, the efficiency of Cas9/CRISPR system as referred to in the present patent application, e.g. their ability to induce a desired event (Homologous gene targeting, targeted mutagenesis, sequence removal or excision) at a locus, depends on several parameters, including the specific activity of the nuclease, probably the accessibility of the target, and the efficacy and outcome of the repair pathway(s) resulting in the desired event (homologous repair for gene targeting, NHEJ pathways for targeted mutagenesis).

The present invention relates to a method for gene targeting using the cas9 described above. The present invention relates to a method comprising one or several of the following steps:

(a) selecting a target nucleic acid sequence, optionally comprising a PAM motif in the cell;

(b) providing a guideRNA comprising a sequence complementary to the target nucleic acid sequence;

(c) introducing into the cell the guide RNA and said Cas9, such that Cas9 processes the target nucleic acid sequence in the cell.

In a particular embodiment, the method comprises:

(a) selecting a target nucleic acid sequence, optionally comprising a PAM motif in the cell;

(b) providing a crRNA comprising a sequence complementary to the target nucleic acid sequence;

(c) Providing a TracrRNA comprising a sequence complementary to a portion of the crRNA and a Cas9 as described above;

(d) introducing into the cell the crRNA, said TracrRNA and said Cas9, such that Cas9-tracrRNA:crRNA complex process the target nucleic acid sequence in the cell.

In another particular embodiment, said method comprises:

(a) selecting a target nucleic acid sequence, optionally comprising a PAM motif in the cell;

(b) providing a guide RNA comprising a sequence complementary to the target nucleic acid sequence;

(c) providing at least one split Cas9 domain as described above;

(d) introducing into the cell said split Cas9 domain, such that said split Cas9 domain processes the target nucleic acid sequence in the cell.

Said Cas9 split domains (RuvC and HNH domains) can be simultaneously or sequentially introduced into the cell such that said split Cas9 domain(s) process the target nucleic acid sequence. The Cas9 split system is particularly suitable for an inducible method of genome targeting. In a preferred embodiment, to avoid the potential toxic effect of the Cas9 overexpression within the cell, a non-functional split Cas9 domain is introduced into the cell, preferably by stably transforming said cell with a transgene encoding said split domain. Then, the complementary split part of Cas9 is introduced into the cell, such that the two split parts reassemble into the cell to reconstitute a functional Cas9 protein at the desired time. Said split Cas9 can derive from the same Cas9 protein or can derive from different Cas9 variants, particularly RuvC and HNH domains as described above.

In another particular embodiment, the method of gene targeting using the split cas9 protein can further comprise adding antibodies or small molecules which bind to the interface between the two split Cas9 protein and thus avoid split Cas9 reassembling to reconstitute a functional Cas9 protein. To induce Cas9 activation, the antibodies or small molecules have to be removed by a washing step.

In another aspect of the invention, only one split Cas9 domain is introduced into said cell. Indeed, surprisingly the inventors showed that the split Cas9 domain comprising the RuvC motif as described above is capable of cleaving a target nucleic acid sequence independently of split domain comprising the HNH motif. The guideRNA does not need the presence of the HNH domain to bind to the target nucleic acid sequence and is sufficiently stable to be bound by the RuvC split domain.

In a preferred embodiment, said split Cas9 domain alone is capable of nicking said target nucleic acid sequence.

This Cas9 split system is particularly suitable for an inducible method of genome targeting. In a preferred embodiment, to avoid the potential toxic effect of the Cas9 overexpression within the cell, a HNH split Cas9 domain can be introduced into the cell, preferably by stably transforming said cell with a transgene encoding said split domain. Then, the complementary split part of Cas9 (RuvC domain) is introduced into the cell, such that the two split parts reassemble into the cell to reconstitute a functional Cas9 protein at the desired time.

The term “process” as used herein means that sequence is considered modified simply by the binding of the Cas9. Depending of the Cas9 used, different processed event can be induced within the target nucleic acid sequence. As non limiting example, Cas9 can induce cleavage, nickase events or can yield to specific silencing of the gene of interest. Any target nucleic acid sequence can be processed by the present methods. The target nucleic acid sequence (or DNA target) can be present in a chromosome, an episome, an organellar genome such as mitochondrial or chloroplast genome or genetic material that can exist independently to the main body of genetic material such as an infecting viral genome, plasmids, episomes, transposons for example. A target nucleic acid sequence can be within the coding sequence of a gene, within transcribed non-coding sequence such as, for example, leader sequences, trailer sequence or introns, or within non-transcribed sequence, either upstream or downstream of the coding sequence. The nucleic acid target sequence is defined by the 5’ to 3’ sequence of one strand of said target.

Any potential selected target nucleic acid sequence in the present invention may have a specific sequence on its 3’ end, named the protospacer adjacent motif or protospacer associated motif (PAM). The PAM is present in the targeted nucleic acid sequence but not in the guide RNA that is produced to target it. Preferably, the proto-spacer adjacent motif (PAM) may correspond to 2 to 5 nucleotides starting immediately or in the vicinity of the proto-spacer at the leader distal end. The sequence and the location of the PAM vary among the different systems. PAM motif can be for examples NNAGAA, NAG, NGG, NGGNG, AWG, CC, CC, CCN, TCN, TTC as non limiting examples (shah SA, RNA biology 2013). Different Type II systems have differing PAM requirements. For example, the *S. pyogenes* system requires an NGG sequence, where N can be any nucleotides. *S. thermophilus* Type II systems require NGGNG (Horvath and Barrangou 2010) and NNAGAAW (Deveau, Barrangou et al. 2008), while different *S. mutant* systems tolerate NGG or NAAR (van der Ploeg 2009). PAM is not restricted to the region adjacent to the proto-spacer but can also be part of the proto-spacer (Mojica, Diez-Villasenor et al. 2009). In a particular embodiment, the Cas9 protein can be engineered to recognize a non natural PAM motif. In this case, the selected target sequence may comprise a smaller or a larger PAM motif with any combinations of amino acids. In a preferred embodiment, the selected target sequence comprise a PAM motif which comprises at least 3, preferably, 4, more preferably 5 nucleotides recognized by the Cas9 variant according to the present invention. Preferably, the Cas9 variant comprise at least one mutated residue in the DNA/RNA binding region, preferably in the amino acid sequence selected from the group consisting of SEQ ID NO: 34 to SEQ ID NO: 48 and recognizes a non natural PAM motif. The aligned region (see Table 1, 3 and 5) of the Cas9 homologues can also be mutated in the present invention to recognize a non natural PAM motif. The capacity of Cas9 to bind a PAM motif within the genomic DNA, in absence of crRNA (or guide RNA) can present a potential toxic effect when Cas9 is overexpressed in the cell. Thus, to avoid this potential toxic effect, the inventors sought to engineer Cas9 or split Cas9 domain which are not capable of binding a PAM motif. The Cas9 variant or split Cas9 domain according to the present invention comprises at least one amino acid residue in the PAM binding region to avoid PAM binding, preferably in the region from residue T38 to E57 and/or from T146 to L169 of the SEQ ID NO: 3 or aligned positions using CLUSTALW method on homologues of Cas family members.

The method of the present invention comprises providing an engineered guide RNA. Guide RNA corresponds to a nucleic acid comprising a complementary sequence to a target nucleic acid sequence. Preferably, guide RNA corresponds to a crRNA and tracrRNA which can be used separately or fused together. In natural type II CRISPR system, the CRISPR targeting RNA (crRNA) targeting sequences are transcribed from DNA sequences known as protospacers. Protospacers are clustered in the bacterial genome in a group called a CRISPR array. The protospacers are short sequences (~20bp) of known foreign DNA separated by a short palindromic repeat and kept like a record against future encounters. To create the crRNA, the CRISPR array is transcribed and the RNA is processed to separate the individual recognition sequences between the repeats. The spacer-containing CRISPR locus is transcribed in a long pre-crRNA. The processing of the CRISPR array transcript (pre-crRNA) into individual crRNAs is dependent on the presence of a trans-activating crRNA (tracrRNA) that has sequence complementary to the palindromic repeat. The tracrRNA hybridizes to the repeat regions separating the spacers of the pre-crRNA, initiating dsRNA cleavage by endogenous RNase III, which is followed by a second cleavage event within each spacer by Cas9, producing mature crRNAs that remain associated with the tracrRNA and Cas9 and form the Cas9-tracrRNA:crRNA complex. Engineered crRNA with tracrRNA is capable of targeting a selected nucleic acid sequence, obviating the need of RNase III and the crRNA processing in general (Jinek, Chylinski et al. 2012).

In the present invention, guide RNA is engineered to comprise a sequence complementary to a portion of a target nucleic acid such that it is capable of targeting, preferably cleaving the target nucleic acid sequence. In a particular embodiment, the guide RNA comprises a sequence of 5 to 50 nucleotides, preferably at least 12 nucleotides which is complementary to the target nucleic acid sequence. In a more particular embodiment, the guide RNA is a sequence of at least 30 nucleotides which comprises at least 10 nucleotides, preferably 12 nucleotides complementary to the target nucleic acid sequence.

In the present invention, RNA/DNA binding region of Cas9 can be engineered to allow the recognition of larger guide RNA sequence. In particular, said RNA/DNA binding region of Cas9 can be engineered to increase the number of nucleotides which specifically bind the nucleic acid target sequence. In a particular embodiment, at least 12 nucleotides specifically binds the nucleic acid target sequence, more preferably at least 15 nucleotides, more preferably again at least 20 nucleotides.

In another aspect, guide RNA can be engineered to comprise a larger sequence complementary to a target nucleic acid. Indeed, the inventors showed that the RuvC split Cas9 domain is able to cleave the target nucleic acid sequence only with a tracRNA:crRNA complex (guide RNA). Thus, the guide RNA can bind the target nucleic acid sequence in absence of the HNH split Cas9 domain. The guide RNA can be designed to comprise a larger complementary sequence, preferably more than 20 bp, to increase the annealing between DNA-RNA duplex without the need to have the stability effect of the HNH split domain binding. Thus, the guide RNA can comprise a complementary sequence to a target nucleic acid sequence of more than 20 bp. Such guide RNA allow increasing the specificity of the Cas9 activity.

The guideRNA does not need the presence of the HNH domain to bind to the target nucleic acid sequence and is sufficiently stable to be bound by the RuvC split domain. Thus, in another particular embodiment, said guide RNA comprises only a nucleic acid sequence, preferably a RNA sequence comprising a complementary sequence to said target nucleic acid sequence without a tracrRNA sequence. Said complementary sequence comprises at least 10 nucleotides, preferably at least 20 nucleotides.

The guide RNA may also comprise a complementary sequence followed by 4-10 nucleotides on the 5'end to improve the efficiency of targeting (Cong, Ran et al. 2013; Mali, Yang et al. 2013). In preferred embodiment, the complementary sequence of the guide RNA is followed in 3'end by a nucleic acid sequence named repeat sequences or 3'extension sequence. Coexpression of several guide RNA with distinct complementary regions to two different genes targeted both genes can be used simultaneously. Thus, in particular embodiment, the guide RNA can be engineered to recognize different target nucleic acid sequences simultaneously. In this case, same guide RNA comprises at least two distinct sequences complementary to a portion of the different target nucleic acid sequences. In a preferred embodiment, said complementary sequences are spaced by a repeat sequence.

The guide RNA according to the present invention can also be modified to increase its stability of the secondary structure and/or its binding affinity for Cas9. In a particular embodiment, the guide RNA can comprise a 2', 3'-cyclic phosphate. The 2', 3'- cyclic phosphate terminus seems to be involved in many cellular processes i.e. tRNA splicing, endonucleolytic cleavage by several ribonucleases, in self-cleavage by RNA ribozyme and in response to various cellular stress including accumulation of unfolded protein in the endoplasmatic reticulum and oxidative stress (Schutz, Hesselberth et al. 2010). The inventors have speculated that the 2', 3'-cyclic phosphate enhances the guide RNA stability or its affinity/specificity for Cas9. Thus, the present invention relates to the modified guide RNA comprising a 2', 3'-cyclic phosphate, and the methods for genome engineering based on the CRISPR/cas system (Jinek, Chylinski et al. 2012; Cong, Ran et al. 2013; Mali, Yang et al. 2013) using the modified guide RNA.

The guide RNA may also comprise a Trans-activating CRISPR RNA (TracrRNA). TracrRNA according to the present invention are characterized by an anti-repeat sequence capable of base-pairing with at least a part of the 3' extension sequence of crRNA to form a tracrRNA:crRNA also named guideRNA (gRNA). TracrRNA comprises a sequence complementary to a region of the crRNA. A synthetic single guide RNA (sgRNA) comprising a fusion of crRNA and tracrRNA that forms a hairpin that mimics the tracrRNA-crRNA complex (Jinek, Chylinski et al. 2012; Cong, Ran et al. 2013; Mali, Yang et al. 2013) can be used to direct Cas9 endonuclease-mediated cleavage of target nucleic acid. This system has been shown to function in a variety of eukaryotic cells, including human, zebra fish and yeast. The sgRNA may comprise two distinct sequences complementary to a portion of the two target nucleic acid sequences, preferably spaced by a repeat sequence.

The methods of the invention involve introducing guide RNA, split Cas9 or Cas9 into a cell. Guide RNA , Cas9 or split Cas9 domain may be synthesized *in situ* in the cell as a result of the introduction of polynucleotide encoding RNA or polypeptides into the cell. Alternatively, the guide RNA, split Cas9, Cas9 RNA or Cas9 polypeptides could be produced outside the cell and then introduced thereto. Methods for introducing a polynucleotide construct into bacteria, plants, fungi and animals are known in the art and including as non-limiting examples stable transformation methods wherein the polynucleotide construct is integrated into the genome of the cell, transient transformation methods wherein the polynucleotide construct is not integrated into the genome of the cell and virus mediated methods. Said polynucleotides may be introduced into a cell by for example, recombinant viral vectors (e.g. retroviruses, adenoviruses), liposomes and the like. For example, transient transformation methods include for example microinjection, electroporation or particle bombardment. Said polynucleotides may be included in vectors, more particularly plasmids or virus, in view of being expressed in prokaryotic or eukaryotic cells.

cas9 according to the present invention can induce genetic modification resulting from a cleavage event in the target nucleic acid sequence that is commonly repaired through non-homologous end joining (NHEJ). NHEJ comprises at least two different processes. Mechanisms involve rejoining of what remains of the two DNA ends through direct re-ligation (Critchlow and Jackson 1998) or via the so-called microhomology-mediated end joining (Ma, Kim et al. 2003). Repair via non-homologous end joining (NHEJ) often results in small insertions or deletions and can be used for the creation of specific gene knockouts. By “cleavage event” is intended a double-strand break or a single-strand break event. Said modification may be a deletion of the genetic material, insertion of nucleotides in the genetic material or a combination of both deletion and insertion of nucleotides.

The present invention also relates to a method for modifying target nucleic acid sequence further comprising the step of expressing an additional catalytic domain into a host cell. In a more preferred embodiment, the present invention relates to a method to increase mutagenesis wherein said additional catalytic domain is a DNA end-processing enzyme. Non limiting examples of DNA end-processing enzymes include 5-3' exonucleases, 3-5' exonucleases, 5-3' alkaline exonucleases, 5' flap endonucleases, helicases, hosphatase, hydrolases and template-independent DNA polymerases. Non limiting examples of such catalytic domain comprise of a protein domain or catalytically active derivate of the protein domain seleced from the group consisting of hExoI (EXO1_HUMAN), Yeast ExoI (EXO1_YEAST), E.coli ExoI, Human TREX2, Mouse TREX1, Human TREX1, Bovine TREX1, Rat TREX1, TdT (terminal deoxynucleotidyl transferase) Human DNA2, Yeast DNA2 (DNA2_YEAST). In a preferred embodiment, said additional catalytic domain has a 3'-5'-exonuclease activity, and in a more preferred embodiment, said additional catalytic domain has TREX exonuclease activity, more preferably TREX2 activity. In another preferred embodiment, said catalytic domain is encoded by a single chain TREX polypeptide. Said additional catalytic domain may be fused to a nuclease fusion protein or chimeric protein according to the invention optionally by a peptide linker.

Endonucleolytic breaks are known to stimulate the rate of homologous recombination. Therefore, in another preferred embodiment, the present invention relates to a method for inducing homologous gene targeting in the nucleic acid target sequence further comprising providing to the cell an exogeneous nucleic acid comprising at least a sequence homologous to a portion of the target nucleic acid sequence, such that homologous recombination occurs between the target nucleic acid sequence and the exogeneous nucleic acid.

In particular embodiments, said exogenous nucleic acid comprises first and second portions which are homologous to region 5' and 3' of the target nucleic acid sequence, respectively. Said exogenous nucleic acid in these embodiments also comprises a third portion positioned between the first and the second portion which comprises no homology with the regions 5' and 3' of the target nucleic acid sequence. Following cleavage of the target nucleic acid sequence, a homologous recombination event is stimulated between the target nucleic acid sequence and the exogenous nucleic acid. Preferably, homologous sequences of at least 50 bp, preferably more than 100 bp and more preferably more than 200 bp are used within said donor matrix. Therefore, the exogenous nucleic acid is preferably from 200 bp to 6000 bp, more preferably from 1000 bp to 2000 bp. Indeed, shared nucleic acid homologies are located in regions flanking upstream and downstream the site of the break and the nucleic acid sequence to be introduced should be located between the two arms.

Depending on the location of the target nucleic acid sequence wherein break event has occurred, such exogenous nucleic acid can be used to knock-out a gene, e.g. when exogenous nucleic acid is located within the open reading frame of said gene, or to introduce new sequences or genes of interest. Sequence insertions by using such exogenous nucleic acid can be used to modify a targeted existing gene, by correction or replacement of said gene (allele swap as a non-limiting example), or to up- or down-regulate the expression of the targeted gene (promoter swap as non-limiting example), said targeted gene correction or replacement.

Modified cells and kits

A variety of cells are suitable for use in the method according to the invention. Cells can be any prokaryotic or eukaryotic living cells, cell lines derived from these organisms for *in vitro* cultures, primary cells from animal or plant origin.

By “primary cell” or “primary cells” are intended cells taken directly from living tissue (i.e. biopsy material) and established for growth in vitro, that have undergone very few population doublings and are therefore more representative of the main functional components and characteristics of tissues from which they are derived from, in comparison to continuous tumorigenic or artificially immortalized cell lines. These cells thus represent a more valuable model to the *in vivo* state they refer to.

In the frame of the present invention, “eukaryotic cells” refer to a fungal, plant, algal or animal cell or a cell line derived from the organisms listed below and established for in vitro culture. More preferably, the fungus is of the genus *Aspergillus, Penicillium, Acremonium, Trichoderma, Chrysoporium, Mortierella, Kluyveromyces or Pichia*; More preferably, the fungus is of the species *Aspergillus niger, Aspergillus nidulans, Aspergillus oryzae, Aspergillus terreus, Penicillium chrysogenum, Penicillium citrinum, Acremonium Chrysogenum, Trichoderma reesei, Mortierella alpine, Chrysosporium lucknowense, Kluyveromyceslactis, Pichia pastoris or Pichia ciferrii*. More preferably the plant is of the genus *Arabidospis, Nicotiana, Solanum, lactuca, Brassica, Oryza, Asparagus, Pisum, Medicago, Zea, Hordeum, Secale, Triticum, Capsicum, Cucumis, Cucurbita, Citrullis, Citrus, Sorghum*; More preferably, the plant is of the species *Arabidospis thaliana, Nicotiana tabaccum, Solanum lycopersicum, Solanum tuberosum, Solanum melongena, Solanum esculentum, Lactuca saliva, Brassica napus, Brassica oleracea, Brassica rapa, Oryza glaberrima, Oryza sativa, Asparagus officinalis, Pisumsativum, Medicago sativa, zea mays, Hordeum vulgare, Secale cereal, Triticuma estivum, Triticum durum, Capsicum sativus, Cucurbitapepo, Citrullus* *lanatus, Cucumis melo, Citrus aurantifolia, Citrus maxima, Citrus medica, Citrus reticulata*. More preferably the animal cell is of the genus *Homo, Rattus, Mus, Sus, Bos, Danio, Canis, Felis, Equus, Salmo, Oncorhynchus, Gallus, Meleagris, Drosophila, Caenorhabditis*; more preferably, the animal cell is of the species *Homo sapiens, Rattus norvegicus, Mus musculus, Sus scrofa, Bos taurus, Danio rerio, Canis lupus, Felis catus, Equus caballus, Salmo salar, Oncorhynchus mykiss, Gallus gallus, Meleagris gallopavo, Drosophila melanogaster, Caenorhabditis elegans*.

In the present invention, the cell is preferably a plant cell, a mammalian cell, a fish cell, an insect cell or cell lines derived from these organisms for *in vitro* cultures or primary cells taken directly from living tissue and established for *in vitro* culture. As non limiting examples cell lines can be selected from the group consisting of CHO-K1 cells; HEK293 cells; Caco2 cells; U2-OS cells; NIH 3T3 cells; NSO cells; SP2 cells; CHO-S cells; DG44 cells; K-562 cells, U-937 cells; MRC5 cells; IMR90 cells; Jurkat cells; HepG2 cells; HeLa cells; HT-1080 cells; HCT-116 cells; Hu-h7 cells; Huvec cells; Molt 4 cells. Are also encompassed in the scope of the present invention stem cells, embryonic stem cells and induced Pluripotent Stem cells (iPS).

All these cell lines can be modified by the method of the present invention to provide cell line models to produce, express, quantify, detect, study a gene or a protein of interest; these models can also be used to screen biologically active molecules of interest in research and production and various fields such as chemical, biofuels, therapeutics and agronomy as non-limiting examples.

A particular aspect of the present invention relates to an isolated cell as previously described obtained by the method according to the invention. Typically, said isolated cell comprises at least a Cas9 variant, or a split cas9 domain as described above, optionally with guide RNA. Resulting isolated cell comprises a modified target nucleic acid sequence**.** The resulting modified cell can be used as a cell line for a diversity of applications ranging from bioproduction, animal transgenesis (by using for instance stem cells), plant transgenesis (by using for instance protoplasts), to cell therapy (by using for instance T-cells). The methods of the invention are useful to engineer genomes and to reprogram cells, especially iPS cells and ES cells. Another aspect of the invention is a kit for cell transformation comprising a Cas9 variant or a split Cas9 protein as previously described. This kit more particularly comprise a Cas9 variant or a split Cas9 protein comprising no more than 1100 amino acids encoding for RuvC and/or HNH domains comprising at least one RuvC motif sequence D-[I/L]-G-X-X-S-X-G-W-A or one HNH motif sequence Y-X-X-D-H-X-X-P-X-S-X-X-X-D-X-S, wherein X is anyone of the 20 natural amino acids and [I/L] represents isoleucine or leucine. The kit may also comprise a Cas9 variant or split Cas9 domaincomprising at least one residue mutated in the DNA/RNA binding region, preferably in amino acid sequence SEQ ID NO: 34 to SEQ ID NO: 48. The kit may further comprise one or several components of the type II CRISPR system as described above, such as guide RNA, or crRNA comprising a sequence complementary to a nucleic acid target and at least one tracrRNA.

Method for generating an animal/ a plant

Animals may be generated by introducing Cas9, a split Cas9 protein, guide RNA into a cell or an embryo. In particular, the present invention relates to a method for generating an animal, comprising providing an eukaryotic cell comprising a target nucleic acid sequence into which it is desired to introduce a genetic modification; generating a cleavage within the target nucleic acid sequence by introducing a cas9 according to the present invention; and generating an animal from the cell or progeny thereof, in which cleavage has occurred. Typically, the embryo is a fertilized one cell stage embryo. Polynucleotides may be introduced into the cell by any of the methods known in the art including micro injection into the nucleus or cytoplasm of the embryo. In a particular embodiment, the method for generating an animal, further comprises introducing an exogenous nucleic acid as desired. The exogenous nucleic acid can include for example a nucleic acid sequence that disrupts a gene after homologous recombination, a nucleic acid sequence that replaces a gene after homologous recombination, a nucleic acid sequence that introduces a mutation into a gene after homologous recombination or a nucleic acid sequence that introduce a regulatory site after homologous recombination. The embryos are then cultured to develop an animal. In one aspect of the invention, an animal in which at least a target nucleic acid sequence of interest has been engineered is provided. For example, an engineered gene may become inactivated such that it is not transcribed or properly translated, or an alternate form of the gene is expressed. The animal may be homozygous or heterozygous for the engineered gene.

The present invention also relates to a method for generating a plant comprising providing a plant cell comprising a target nucleic acid sequence into which it is desired to introduce a genetic modification; generating a cleavage within the target nucleic acid sequence by introducing a Cas9 or a split Cas9 protein according to the present invention; and generating a plant from the cell or progeny thereof, in which cleavage has occurred. Progeny includes descendants of a particular plant or plant line. In a particular embodiment, the method for generating a plant, further comprise introducing an exogenous nucleic acid as desired. Said exogenous nucleic acid comprises a sequence homologous to at least a portion of the target nucleic acid sequence, such that homologous recombination occurs between said exogenous nucleic acid and the target nucleic acid sequence in the cell or progeny thereof. Plant cells produced using methods can be grown to generate plants having in their genome a modified target nucleic acid sequence. Seeds from such plants can be used to generate plants having a phenotype such as, for example, an altered growth characteristic, altered appearance, or altered compositions with respect to unmodified plants.

In a particular embodiment, an animal or a plant may be generated by introducing only one split Cas9 protein. Another animal or plant may be generated by introducing the complementary split Cas9 protein. The resulting animals or plants can be crossed together, to generate descendants expressing both split Cas9 proteins which can cleave target nucleic acid sequence.

The polypeptides of the invention are useful to engineer genomes and to reprogram cells, especially iPS cells and ES cells.

Therapeutic applications

The method disclosed herein can have a variety of applications. In one embodiment, the method can be used for clinical or therapeutic applications. The method can be used to repair or correct disease-causing genes, as for example a single nucleotide change in sickle-cell disease. The method can be used to correct splice junction mutations, deletions, insertions, and the like in other genes or chromosomal sequences that play a role in a particular disease or disease state.

From the above, the polypeptides according to the invention can be used as a medicament, especially for modulating, activating or inhibiting gene transcription, at the promoter level or through their catalytic domains.

Cas9 or split Cas9 proteins according to the present invention can be used for the treatment of a genetic disease to correct a mutation at a specific locus or to inactivate a gene the expression of which is deleterious. Such proteins can also be used to genetically modify iPS or primary cells, for instance T-cells, in view of injected such cells into a patient for treating a disease or infection. Such cell therapy schemes are more particularly developed for treating cancer, viral infection such as caused by CMV or HIV or self-immune diseases.

General definitions

In the description above, a number of terms are used extensively. The following definitions are provided to facilitate understanding of the present embodiments.

Amino acid residues in a polypeptide sequence are designated herein according to the one-letter code, in which, for example, Q means Gln or Glutamine residue, R means Arg or Arginine residue and D means Asp or Aspartic acid residue.

Amino acid substitution means the replacement of one amino acid residue with another, for instance the replacement of an Arginine residue with a Glutamine residue in a peptide sequence is an amino acid substitution.

Nucleotides are designated as follows: one-letter code is used for designating the base of a nucleoside: a is adenine, t is thymine, c is cytosine, and g is guanine. For the degenerated nucleotides, r represents g or a (purine nucleotides), k represents g or t, s represents g or c, w represents a or t, m represents a or c, y represents t or c (pyrimidine nucleotides), d represents g, a or t, v represents g, a or c, b represents g, t or c, h represents a, t or c, and n represents g, a, t or c.

As used herein, “nucleic acid” or polynucleotide” refers to nucleotides and/or polynucleotides, such as deoxyribonucleic acid (DNA) or ribonucleic acid (RNA), oligonucleotides, fragments generated by the polymerase chain reaction (PCR), and fragments generated by any of ligation, scission, endonuclease action, and exonuclease action. Nucleic acid molecules can be composed of monomers that are naturally-occurring nucleotides (such as DNA and RNA), or analogs of naturally-occurring nucleotides (e.g., enantiomeric forms of naturally-occurring nucleotides), or a combination of both. Modified nucleotides can have alterations in sugar moieties and/or in pyrimidine or purine base moieties. Sugar modifications include, for example, replacement of one or more hydroxyl groups with halogens, alkyl groups, amines, and azido groups, or sugars can be functionalized as ethers or esters. Moreover, the entire sugar moiety can be replaced with sterically and electronically similar structures, such as aza-sugars and carbocyclic sugar analogs. Examples of modifications in a base moiety include alkylated purines and pyrimidines, acylated purines or pyrimidines, or other well-known heterocyclic substitutes. Nucleic acid monomers can be linked by phosphodiester bonds or analogs of such linkages. Nucleic acids can be either single stranded or double stranded.

By “complementary sequence” is meant the sequence part of polynucleotide (e.g. part of crRNa or tracRNA) that can hybridize to another part of polynucleotides (e.g. the target nucleic acid sequence or the crRNA respectively) under standard low stringent conditions. Such conditions can be for instance at room temperature for 2 hours by using a buffer containing 25% formamide, 4x SSC, 50 mM NaH2PO4 / Na2HPO4 buffer; pH 7.0,5x Denhardt’s, 1 mM EDTA,1 mg/ml DNA + 20 to 200 ng/ml probe to be tested (approx. 20 - 200 ng/ml)). This can be also predicted by standard calculation of hybridization using the number of complementary bases within the sequence and the content in G-C at room temperature as provided in the literature. Preferentially, the sequences are complementary to each other pursuant to the complementarity between two nucleic acid strands relying on Watson-Crick base pairing between the strands, i.e. the inherent base pairing between adenine and thymine (A-T) nucleotides and guanine and cytosine (G-C) nucleotides. Accurate base pairing equates with Watson-Crick base pairing includes base pairing between standard and modified nucleosides and base pairing between modified nucleosides, where the modified nucleosides are capable of substituting for the appropriate standard nucleosides according to the Watson-Crick pairing. The complementary sequence of the single-strand oligonucleotide can be any length that supports specific and stable hybridization between the two single-strand oligonucleotides under the reaction conditions. The complementary sequence generally authorizes a partial double stranded overlap between the two hybridized oligonucleotides over more than 3bp, preferably more than 5 bp, preferably more than to 10 bp. The complementary sequence is advantageously selected not to be homologous to any sequence in the genome to avoid off-target recombination or recombination not involving the whole donor matrix (i.e. only one oligonucleotide).

By “nucleic acid homologous sequence” it is meant a nucleic acid sequence with enough identity to another one to lead to homologous recombination between sequences, more particularly having at least 80% identity, preferably at least 90% identity and more preferably at least 95%, and even more preferably 98 % identity. “Identity" refers to sequence identity between two nucleic acid molecules or polypeptides. Identity can be determined by comparing a position in each sequence which may be aligned for purposes of comparison. When a position in the compared sequence is occupied by the same base, then the molecules are identical at that position. A degree of similarity or identity between nucleic acid or amino acid sequences is a function of the number of identical or matching nucleotides at positions shared by the nucleic acid sequences. Various alignment algorithms and/or programs may be used to calculate the identity between two sequences, including FASTA, or BLAST which are available as a part of the

GCG sequence analysis package (University of Wisconsin, Madison, Wis.), and can be used with, e.g., default setting.

The terms "vector" or “vectors” refer to a nucleic acid molecule capable of transporting another nucleic acid to which it has been linked. A “vector” in the present invention includes, but is not limited to, a viral vector, a plasmid, a RNA vector or a linear or circular DNA or RNA molecule which may consists of a chromosomal, non-chromosomal, semi-synthetic or synthetic nucleic acids. Preferred vectors are those capable of autonomous replication (episomal vector) and/or expression of nucleic acids to which they are linked (expression vectors). Large numbers of suitable vectors are known to those of skill in the art and commercially available. Viral vectors include retrovirus, adenovirus, parvovirus (e. g. adenoassociated viruses), coronavirus, negative strand RNA viruses such as orthomyxovirus (e. g., influenza virus), rhabdovirus (e. g., rabies and vesicular stomatitis virus), paramyxovirus (e. g. measles and Sendai), positive strand RNA viruses such as picornavirus and alphavirus, and double-stranded DNA viruses including adenovirus, herpesvirus (e. g., Herpes Simplex virus types 1 and 2, Epstein-Barr virus, cytomegalovirus), and poxvirus (e. g., vaccinia, fowlpox and canarypox). Other viruses include Norwalk virus, togavirus, flavivirus, reoviruses, papovavirus, hepadnavirus, and hepatitis virus, for example. Examples of retroviruses include: avian leukosis-sarcoma, mammalian C-type, B-type viruses, D type viruses, HTLV-BLV group, lentivirus, spumavirus (Coffin, J. M., Retroviridae: The viruses and their replication, In Fundamental Virology, Third Edition, B. N. Fields, et al., Eds., Lippincott-Raven Publishers, Philadelphia, 1996).

Having generally described this invention, a further understanding can be obtained by reference to certain specific examples, which are provided herein for purposes of illustration only, and are not intended to be limiting unless otherwise specified.

EXAMPLES

Example 1: Identification of conserved sequence segments of Cas9 homologues

In order to increase the efficacy of transfection and vectorization the inventors perform truncations of the protein of Cas9 of *S. pyogenes* (gi|15675041|). The truncated forms of Cas9 will be tested in mammalian cells for efficiency of NHEJ and HR. A first strategy implies a semi rational approach based on the identification of conserved sequence segments of homologues of Cas9 Pyogenes. The strategy is based on the use of data derived from sequence features of Cas9 of pyogenes i.e. sequence homologues as well as secondary structure predictions and protein domain boundaries predictions.

The sequence of *S. pyogenes* Cas9 belongs to the COG3513 (Predicted CRISPR-associated nuclease, contains McrA/HNH-nuclease and RuvC-like nuclease domain). The alignment of sequence members of COG3513 has been used to build two sequence motifs, each one next to one of the two known catalytic domains RuvC and HNH. The two designed motifs, RuvC motif (D-[I/L]-G-X-X-S-X-G-W-A) (SEQ ID NO: 1) and HNH motif (Y-X-X-D-H-X-X-P-X-S-X-X-X-D-X-S) (SEQ ID NO: 2) have been used to extract all the protein sequences presented in UniProtKB database using the ScanProsite tool (de Castro, Sigrist et al. 2006).

The use of the RuvC motif (D-[I/L]-G-X-X-S-X-G-W-A) (SEQ ID NO: 1) allows the extraction of 358 sequences and the use of HNH motif (Y-X-X-D-H-X-X-P-X-S-X-X-X-D-X-S) (SEQ ID NO: 2) allows the extraction of 187 sequences. All the extracted sequenced have been inspected looking for putative cas9 homologues with interesting features as smaller size, different origins and/or different organization of the locus. The homologues for each domain have been analysed separately and a few of them have been extracted and aligned. The boundaries of each domain have been identified.

- **RuvC-like domain**

Among the 358 sequences found using the RuvC sequence motif, eight sequences (SEQ ID NO: 5 to SEQ ID NO: 12) have been extracted and aligned to the original sequence of *S. pyogenes* Cas9 (SEQ ID NO: 3). The alignments have been made using standard multiple sequence alignment software (DIALIGN 2.2.1 software) (Morgenstern 2004). The alignments of the Cas9 homologues are presented in Table 1 as follows:

1) ***S. pyogenes* Cas9** (SEQ ID NO: 4) 1368 amino acids (AA)

2)	D8IJI3_LACSC	(SEQ ID NO: 5)	183 AA
3)	F0K1W4_LACD2	(SEQ ID NO: 6)	669 AA
4)	E1NX15_9LACO	(SEQ ID NO: 7)	142 AA
5)	C5F1Z4_9HELI	(SEQ ID NO: 8)	344 AA
6)	F3ZS86_9BACE	(SEQ ID NO: 9)	349 AA
7)	H1D479_9FUSO	(SEQ ID NO: 10)	198 AA
8)	K1M766_9LACO	(SEQ ID NO: 11)	857 AA
9)	Q7VG48_HELHP	(SEQ ID NO: 12)	131 AA

The protein secondary structure of RuvC-like domain of Cas9 has been predicted using the PSIPRED secondary structure prediction method (Jones 1999; Buchan, Ward et al. 2010) (See Table 2).

Using this multiple sequence alignment and the prediction derived from DoBo server (Eickholt, Deng et al. 2011) together with the secondary structures prediction we can assume that the RuvC-like domain of *S. pyogenes* Cas9 extends until position 166G (SEQ ID NO: 2).

- **HNH Domain**

Among the 187 sequences found using the HNH sequences motif (Y-X-X-D-H-X-X-P-X-S-X-X-X-D-X-S), nine sequences (SEQ ID NO: 14 to 22) have been extracted and aligned to the original sequences of *S. pyogenes* Cas9 (SEQ ID NO: 3) using DIALIGN 2.2.1 software as described above (see Table 3). The alignments of the Cas9 homologues are presented in Table 3 as follows:

1)	D8IJI4_LACSC	(SEQ ID NO: 14)	897 AA
2)	F0K1W6_LACD2	(SEQ ID NO: 15)	544 AA
3)	D4FGK2_9LACO	(SEQ ID NO: 16)	534 AA
4)	E1NX12_9LACO	(SEQ ID NO: 17)	667 AA
5)	E7NSW3_TREPH	(SEQ ID NO: 18)	591 AA
6)	H1D477_9FUSO	(SEQ ID NO: 19)	387 AA
7)	C2KFJ4_9LACO	(SEQ ID NO: 20)	544 AA
8)	K1MRU9_9LACO	(SEQ ID NO: 21)	206 AA
9)	E3ZTQ9_LISSE	(SEQ ID NO: 22)	874 AA
10)	***S. pyogenes*** **Cas9**	(SEQ ID NO: 3)	1368 AA

The protein secondary structure of HNH domain of Cas9 has been predicted using the PSIPRED secondary structure prediction method (Jones 1999; Buchan, Ward et al. 2010) (See Table 4).

The boundaries of the HNH domain has been identified using the multiple sequence alignment of the *S. pyogenes* Cas9 homologues and DoBO server (Eickholt, Deng et al. 2011) and secondary structure prediction server (psipred). Two versions of Cas9 HNH domains have been predicted. The N-terminus of each HNH domain version corresponds to P800, while the C-terminus corresponds to the Y981 for the shorter version (SEQ ID NO: 23) or G1055 for the longer version (SEQ ID NO: 13).

A Cas9 comprising the new RuvC domain identified (SEQ ID NO: 4) and one of the two versions of the HNH domains (SEQ ID NO: 13 and SEQ ID NO: 23) will be engineered and its activity will be tested.

Example 2: Identification of shorter Cas9 homologues and digestion of the C-terminal domain of Cas9

The present study further allows identifying four putative natural Cas9 homologues with shorter sequence (SEQ ID NO: 26 to SEQ ID NO: 29). These natural shorter Cas9 versions have been aligned with the *S. pyogenes* Cas9 using DIALIGN 2.1.1 software as described above. The alignments of the shorter Cas9 homologues are presented in Table 5 as follows:

1)	D4IZM9_BUTFI	(SEQ ID NO: 26)	765 AA
2)	Q9CLT2_PASMU	(SEQ ID NO: 27)	1056 AA
3)	E0G5X6_ENTFL	(SEQ ID NO: 28)	936 AA
4)	E0XXB7_9DELT	(SEQ ID NO: 29)	1011 AA
5)	***S. pyogenes* Cas9**	(SEQ ID NO: 3)	1368 AA

The protein secondary structure of shorter Cas9 has been predicted using the PSIPRED secondary structure prediction method (Jones 1999; Buchan, Ward et al. 2010) (See Table 6).

For all the shorter sequence homologs (D4IZM9_BUTFI, Q9CLT2_PASMU, E0G5X6_ENTFL, E0G5X6_ENTFL; SEQ ID NO: 26 to SEQ ID NO: 29) position T956 seems to be quite conserved anyway looking at the secondary structure prediction of Cas9 in this zone Y943 seems to be more a better position to cut the C-terminus of Cas9 of Pyogenes.

To perform a progressive enzymatic digestion we will use a modified protocol described by (Lutz, Ostermeier et al. 2001). Through the use of exonuclease and heat inactivation we will create an incremental truncation library of the C-terminal of Cas9. Approximately we will create libraries of fragments of Cas9 starting from 1364 up to ~ 900 aa.

Three shorter version of the entire sequence of Cas9: Cas9_delta943 (SEQ ID NO: 31), Cas9_delta980 (SEQ ID NO: 32) and Cas9_delta1055 (SEQ ID NO: 33) will be engineered (see Figure 1).

The new cas9 scaffolds obtained with the two different strategies will be tested in mammalian cells using the sgRNA chimera and the PAM specific for *S. pyogenes* already described in (Mali, Yang et al. 2013)

Example 3: Identification of the residues involved in the DNA/RNA binding specificity of *S.pyogenes* Cas9 and homologues thereof.

In the present study, the identification of Cas9 residues involved in the binding of the guide RNA and the PAM motif will allow to engineer new Cas9 scaffolds and thus modulate affinity for the selected target.

Using the multiple sequence alignment of Cas9 homologues as described above, the inventors identified the most conserved regions in terms of primary sequence of S. *pyogenes* Cas9. The inventors matched this data with results derived from servers capable of predicting DNA and RNA binding residues from sequence features (i.e. BindN) (Wang and Brown 2006). Contemporaneously the solvent accessibility and secondary structure prediction of the primary sequence of Cas9 has been used to identify the most exposed residues on the surface of the protein. The predicted DNA and RNA binding region on *S. pyogenes* Cas9 are listed in Table 7. The predicted DNA and RNA binding regions on Cas9 homologues are represented in Table 1, 3 and 5 (grey highlighted sequences).

Structure of *S. pyogenes* Cas9 was predicted using automated software. In Figure 2 and Figure 3, the inventors mapped the 15 predicted DNA/RNA binding regions described above on the best 3-dimensional model output to determine the residues susceptible to be in contact with DNA or RNA.

A multiple sequence alignment between Cas9 of S. pyogenes (SEQ ID NO: 61) and S. Thermophilus (SEQ ID NO: 64) and the sequence of two pdb structures of RuvC domain of E. Coli and T. Thermophilus (SEQ ID NO: 62 and SEQ ID NO: 63) has also been built using clustalw (see Table 8).

The multiple sequence alignment of the two RuvC domains with the two sequences of Cas9 can point out the stretch of residues not presented in the RuvC domains that could be responsible of the specificity of the PAM. The RuvC domains have a specificity of cleavage which is not present in Cas9, on the contrary the stretch of residues 38T-57E and T146-L169 (which are not conserved in the RuvC domains) could represents the zone responsible of the specificity of the PAM. In particular the differences of sequences between Cas9 of S.pyogens and S.Thermophilus in these two zones could hint to the specificity of each PAM. The residues 39-DRHS-42 and E57 and D147 and I154 are the principal differences between the Cas9 of S.pyogenes and S. Thermophilus and finally they could be the positions responsible for the PAM specificity. Also interestingly are the positions 173D-174L and 177D which are highly exposed and on the same loop of two key residues for the activity of RuvC domain of E. Coli: lys 107 and lys 118 (both positions are not conserved in Cas9).

Collecting all these different sources of data allows the inventors to pinpoint the most probable DNA/RNA binding segments of *S. pyogenes* Cas9. In a first approach, the inventors will create independent libraries (from 3 to 5 amino acids) for each DNA/RNA binding region. In parallel, the inventors will select cluster of amino acids, based on their 3D localization, belonging to different zones but lying on possible patch of charge surface.

In particular to decipher Cas9 residues involved in the recognition of the PAM motif, the inventors will create Cas9 variants libraries comprising randomized residues at each protein seed positions (i.e. with a NVK degenerate codes). The Cas9 variants libraries will be further screened against artificial synthesized targets. As a first set up, the synthesized targets will comprise 20 constant nucleotides necessary for the complex sgRNA::DNA while the base responsible for the recognition of the PAM motif will be modified. Currently the number of PAM nucleotides specifically recognized by *S. pyogenes* Cas9 were restricted to 2 (NGG; (Mali, Yang et al. 2013)). Here, the inventors plan to increase or suppress the number of nucleotides specifically recognized by Cas9 as a way to modulate its specificity.

The “non natural PAM” will be constituted of at least 5 bases; they will be treated as 3 sliding windows of three bases each starting from position 1 to 5. Finally the cas9 libraries will be screened against each set of 64 targets constituting the 3 different sliding windows.

All the Cas9 constructs will be analysed in yeast with a high-throughput screening platform. Once identified a “non natural PAM”, different rounds of refinements will be performed in order to assess the synergistic effects of the contemporary mutation of more than one protein seeds.

Each “non natural PAM” will be also tested on sets of new targets harboring the most dissimilar 20 bases RNA target recognitions. Complementary to this in vivo approach we will also set up experiments of high throughput in vitro protein-DNA interaction using methodology as i.e. Bind-n-Seq. The best combinations of “non natural PAM” and protein constructs will be tested in mammalian cell. Once identified on Cas9 of pyogenes the zone responsible for the recognition of the nucleic acid they will be also plotted on the sequences of chosen homologues and tested in eukaryotic cells.

Example 4: Creation of a split cas9 RNA guided nuclease

The sequence of Cas9 of S. Pyogenes belongs to the COG3513 (Predicted CRISPR-associated nuclease, contains McrA/HNH-nuclease and RuvC-like nuclease domain. The alignment of sequence members of COG3513 has been used to build two sequence motives (each one next to one of the two known catalytic domains: RuvC and HNH). The two sequence motives have been used to extract (using PROSITE) all the protein sequences presented in Uniprot bearing each of the two domains. The use of the RuvC motif (D-[IL]-G-x(2)-S-x-G-W-A) allows the extractions of 358 sequences while the use of HNH motif (Y-2x-D-H-2x-P-x-S-3x-D-x-S) allows the extraction of 187 sequences.

Between the sequences extracted using the RuvC motif eight sequences derived from different organisms were selected. These RuvC-like sequences share interesting features as such as to be present in a short truncated form (if they are compared to the Cas9 of S. Pyogenes composed of 1368 aa) and also to be related to a putative independent HNH domains.

Six of these eight proteins are annotated as uncharacterized proteins: D8IJI3_LACSC from *Lactobacillus salivaris* (SEQ ID NO: 5), F0K1W4 from *Lactobacillus Delbrueckii* (SEQ ID NO: 6), Q7VG48 from Helicobacter Hepaticus (SEQ ID NO: 12) and E9S0G6 from *Treponema Denticola* (SEQ ID NO: 51) and C5F1Z4 from *Helicobacter Pullorum* (SEQ ID NO: 8). Two RuvC-like domains are annotated as Crispr related proteins: H1D479 from *Fusobacterium Necrophorum* (SEQ ID NO: 10) and K1M766 from *Lactobacillus Crispatus* (SEQ ID NO: 11).

The finding of these naturally occurring independent RuvC / HNH like domains has prompted us to engineer the wild type sequence of *S. Pyogenes* Cas9 to create split cas9 proteins. The wild type sequence of *S. Pyogenes* Cas9 has been divided into two separate polypeptide chains (RuvC and

HNH like domains) that co-transfected could assemble to reconstitute the entire wild type sequence of *S. Pyogenes* Cas9. In order to predict the optimal breaking position for the *S. Pyogenes* Cas9 sequence we have built a multiple sequence alignment between the above described eight sequences and wild type sequences of Cas9 of *S. Pyogenes* and *S. Thermophilus* together with the PDB structure of the RuVC domain of *E. coli* (Pdbcode 4EPA) (Table 9). We have integrated these informations with the prediction of secondary structure elements (using PSIPRED) for the sequence of Cas9 *S. Pyogenes* (Table 10).

We have chosen to create a split Cas9 dividing the sequence of S. Pyogenes Cas9 in two independent polypeptide chains using as possible breaking point the position: G247. Specifically we have created two separated domains of cas9 of S. Pyogenes. The domain N-terminal consists of the residues from position 1 to position 247 (SEQ ID NO: 52) and the C-terminal comprehends the residues from amino acid 248 to 1368 (SEQ ID NO: 53).

A S-Tag plus one NLS was fused to the 5' terminus of the split RuvC domain using standard biological tools yielding pCLS24814 plasmid (SEQ ID NO: 54). A 2NLS-BFP-HA-Tag was fused to the 3' terminus of the split HNH domain, then the first amino acid of the split HNH domain was mutated from Leu to Val to have a better Kozak consensus sequence yielding pCLS24813 (SEQ ID NO: 55; pCLS24813).

The nuclease activity of these two split domains with the guide RNA was tested on endogenous GFP_C9_T01 target (SEQ ID NO: 56) in CHO-KI (π10) cell. pCLS24814 and pCLS24813 were co-transfected at three different doses. Positive control corresponds to the transfection of the wild type Cas9 of *S.Pyogenes* with guide RNA (SEQ ID NO: 57; pCLS22972) and control corresponds to the transfection of each split domain separately in presence of the guide RNA.

Nuclease activity of the split cas9 domains was measured as a reduction in GFP fluorescence via flow cytometry using MACSQuant Analyzer (Myltenyi Biotec.) at four and seven days post transfection. The results clearly show that the co-transfection of the two split domains induce a reduction of the percentage of GFP positive cells which is stable over the time (from D4 to D7)(Figure 4).

The nuclease activity of the split cas9 was also tested using a T7 Endo assay (Figure 5). As shown in figure 5, co-transfection of both split domains (at the three different doses) induces cleavage of the DNA,. Our results show that co-transfection of both split domains efficiently cleave the DNA target with no evident toxicity over the time.

The nuclease activity of these two split domains together or each split separately with the guide RNA was also tested on endogenous CD52 target in CHO-KI (π10) cell. The nuclease activity of the split cas9 was tested using a T7 Endo assay (Figure 6). As shown in figure 6, co-transfection of both split domains induces cleavage of the DNA. Surprisingly, the transfection of RuvC split Cas9 domain (N-terminal domain) alone shows the same cleavage profile. Our results show that the N-terminal split domain is active independently of the C-terminal split domain and can cleave the target nucleic acid sequence.

Material and Methods

CHO-KI (π10) cells containing the chromosomally integrated GFP reporter gene including the guide RNA recognition sequence (SEQ ID NO: 56), were cultured at 37°C with 5% CO_2 in F12-K complete medium supplemented with 2 mM l-glutamine, penicillin (100 IU/ml), streptomycin (100 µg/ml), amphotericin B (Fongizone: 0.25 µg/ml, Life Technologies,) and 10% FBS. Cell transfection was performed according to the manufacturer's instructions using the Nucleofector apparatus (Amaxa, Cologne, Germany). Adherent CHO-KI cells were harvested at day 0 of culture, washed twice in phosphate-buffered saline (PBS), trypsinized, and resuspended in T nucleofection solution to a concentration of 1×10^6 cells/100 µL.

We performed the co-transfection of the two split domains at three different doses (we keep constant the quantity of the quide RNA encoding plasmide at 4ug). As first dose we used 1ug for the N-terminal split and 2ug for the C-terminal plasmid; we also double the dose of the two split domains at 2ug and 4ug and as third dose we used an equal quantity for the two split domains plasmids at 4ug.

For each point of transfection we mixed the chosen quantity of the vectors for the two splits domain with the 4µg of guide RNA plasmid GFP_C9_T01 (SEQ ID NO: 58) with 0.1 mL of the CHO-KI (π10) cell suspension (T Nucleofection solution). We transferred the mix to a 2.0-mm electroporation cuvette and nucleofected using program U23 of Amaxa Nucleofector apparatus.

250 ng of BFP expression plasmid have been added to the samples (besides to the one with the C-terminal split domain) in order to estimate the transfection efficiency. Maximum 20 min after nucleofection, 0.5 mL of prewarmed CHO-K1 medium was added to the electroporation cuvette. For each sample cells were then divided into two parts to seed two Petri dish (10ml F12-K) and cultured at 37°C under 5% CO_2 as previously described.

On day 4 post-transfection, cells were washed twice in phosphate-buffered saline (PBS), trypsinized, resuspended in 5 mL medium and percentage of GFP negative cells (200 µl at 2x105 cells/mL). The percentage of GFP negative cells was monitored at D4 and D7 (Figure 4) by flow cytometry MACSQuant Analyzer (Myltenyi Biotec.). Four days post-transfection (day 4), genomic DNA was extracted and the locus of interest was amplified with locus primers 1 and 2 (SEQ ID NO: 59 and 60). Amplicons were analyzed by EndoT7 assay according to the protocol described in (Reyon, Tsai et al. 2012) see Figure 5.

Table 1: Multiple sequence alignment of RuvC domain of Cas9 homologues: D8IJI3_LACSC (SEQ ID NO: 4), F0K1W4_LACD2 (SEQ ID NO: 5), E1NX15_9LACO (SEQ ID NO: 6), C5F1Z4_9HELI (SEQ ID NO: 7), F3ZS86_9BACE (SEQ ID NO: 8), H1D479_9FUSO (SEQ ID NO: 9), K1M766_9LACO (SEQ ID NO: 10), Q7VG48_HELHP (SEQ ID NO: 11) with *S.pyogenes* Cas9 (SEQ ID NO: 3). * corresponds to the predicted 3'-end amino acid (G166) of the *S. pyogenes* Cas9 RuvC-like domain. Grey highlighted sequence: predicted DNA/RNA biding region (see example 3).

```
Cas9 pyogenes       1     ---MDKKYSI GLDIGTNSVG WAVITDEYKV PSKKfkvlgn tdrhsikKNL
D8IJI3_LACSC        1     m----ERYHI GLDIGTSSIG WAVIGDDFKI KRKKG----- -------KNL
F0K1W4_LACD2        1     MAKP-KDYTI GLDIGTNSVG WVVTDDQNNI LRIKG----- -------KKA
E1NX15_9LACO        1     ---MNNNYYL GLDLGTNSVG WAVTDDHYNI IKFHG----- -------KHM
C5F1Z4_9HELI        1     M-K-----IL GFDIGIASIG WAFVENGE-- ----L----- --------KD
F3ZS86_9BACE        1     mkK-----IL GLDIGTNSVG WAVVNTNQeg epsqI----- --------EK
H1D479_9FUSO        1     MKKF-ENYYL GLDIGTSSIG WAVTNSQYDI LKFNG----- -------KYM
K1M766_9LACO        1     mtkLNNEYMV GLDIGTNSCG WVATDFDNNI LKMHG----- -------KRA
Q7VG48_HELHP        1     M-R-----IL GFDIGITSIG WAYVESNE-- ----L----- --------KD

Cas9 pyogenes      48     IGALLF---- ---------D SGETAEATRL KRTARRRYTR RKNRICYLQE
D8IJI3_LACSC       35     IGVRLF---- ---------K EGDTAAERRS FRTQRRRLNR RKWRLKLLEE
F0K1W4_LACD2       38     IGARLF---- ---------T EGKVAAERRS FRTTRRRLSR RRWRIKMLEE
E1NX15_9LACO       36     WGMRLF---- ---------E EAETAKDRRL HRQARRRRQR LVERINLLEE
C5F1Z4_9HELI       26     CGVRIFTKAE NPK------T GDSLAMPRRE ARSVRRRLAR RKGRLETLKR
F3ZS86_9BACE       33     LGSRIIPMSQ DildkfgqgQ TVSSTASRTD YRGIRRLRER SLLRRERLHR
H1D479_9FUSO       38     WGTRLF---- ---------P EANTAQERRI HRSSRRRLKR RKERIQILQM
K1M766_9LACO       39     LGSHLF---- ---------D EGVSAADRRA FRTTRRRIKR RKWRLKLLEE
Q7VG48_HELHP       26     CGVRIFTKAE NPK------N GDSLAAPRRE ARGARRRLAR RKARLNAIKR

Cas9 pyogenes      85     IFSNE----- -------MAK VD-------- ---------- ----------
D8IJI3_LACSC       72     IFDPY----- -------MAE VD-------- ---------- ----------
F0K1W4_LACD2       75     LFDEE----- -------IAK VD-------- ---------- ----------
E1NX15_9LACO       73     LFDKE----- -------ISK VD-------- ---------- ----------
C5F1Z4_9HELI       70     LLAKE----- -------WDL CY-------- ---------- ----------
F3ZS86_9BACE       83     VLhildflpk hyadsigWDp rnsktygkfl pgtevklawv ptadghqflf
H1D479_9FUSO       75     LFDKE----- -------IAK ID-------- ---------- ----------
K1M766_9LACO       76     IFDEE----- -------MAK VD-------- ---------- ----------
Q7VG48_HELHP       70     LLCKE----- -------FEL nln------- ---------- ----------

Cas9 pyogenes      95     -DSFFHRLEE S-FLVEEDKK herhpifgni vdeva--YHE KYPTIYHLRK
D8IJI3_LACSC       82     -EYFFARLKE S-NLSPKDSN KKYLGSLlfp -DISDSNFYD KYPTIYHLRR
F0K1W4_LACD2       85     -PSFFARLHE S-WISPKDKR KRYSAIVFPS PEE-DKKFHE SYPTIYHLRD
E1NX15_9LACO       83     -QGFFARKKE S-DLHFEDKT TKSEYALFND KSYTDRDYYK QYPTIFHLIM
C5F1Z4_9HELI       80     -EDYIAADGE LPKAFmgknl tnp------- ---------- -----YVLRY
F3ZS86_9BACE      133     ySTYLEMLED L-KQTQAQLF ETSQTPVPLD w--------- ---TIYYLRK
H1D479_9FUSO       85     -SGFFQRLKD S-KYYKEDKT EKQTNSIFHD KDYSDKEYHQ DFPTIYHLRK
K1M766_9LACO       86     -PNFFARLKE S-GLSPLDTR KNVSSIVFPT KKM-DKQFYK KFPTIYHLRN
Q7VG48_HELHP       81     --DYLANDGE LPKAYQTSKD TKSPYELY-- ---------- ---TAFHWII

                                                *
Cas9 pyogenes     141     KLVDSTDKAD LRLIYLALAH MIKFRGHFLI EGDLNpdn-- ----------
```

```
D8IJI3_LACSC     129  DLMEKDKKFD LREIYLAIHH IVKYRGNFL- ---------- ----------
F0K1W4_LACD2     132  KLMKDDQKHD IREIYIAVHQ MIKARGNFL- ---------- ----------
E1NX15_9LACO     131  DLIENDKKgi yv-------- ---------- ---------- ----------
C5F1Z4_9HELI     107  EALQRLLSK- -EELVRVVLH IAKHRGYGN- ---------- ----------
F3ZS86_9BACE     170  KALTQPITK- -HELAWLLLH FNTKRGYYQR RGELEdtptd klveyhalkv
H1D479_9FUSO     133  FLLEGNKPKD IRFVYLALHH ILTHRGHFLf pdm------- ----------
K1M766_9LACO     133  ALMKQDKKFD LRAIYIAIHH IVKYRGNFL- ---------- ----------
Q7VG48_HELHP     114  fa-------- ---------- ---------- ---------- ----------
```

Table 2: Secondary structure predictions for the RuvC domain and amino acids sequence of the RuvC domain of the *S. pyogenes* Cas9 (SEQ ID NO: 12). H represents helix, S represents sheet and C represents coil.

Sequence of *S. pyogenes* Cas9 (SEQ ID NO: 12)

```
MDKKYSIGLDIGTNSVGWAVITDEYKVPSKKFKVLGNTDRHSIKKNLIGALLFDSGETAEATRLKRTARRRYTRRKNRICYLQEIFSNEMAKVD
DSFFHRLEESFLVEEDKKHERHPIFGNIVDEVAYHEKYPTIYHLRKKLVDSTDKADLRLIYLALAHMIKFRG
```

Secondary structure Cas9 of Pyogenes

```
CCCCSSSSSSSCCCCSSSSSSCCCCCCCCCCCCCCCCCCCCCCCCCCSSSSSSCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHCCCC
CHHHHHHCCCCCCCCCCCCCCCCCCCCHHHHHHHHHCCCHHHHHHHHHHCCCCCCCHHHHHHHHHHHHHHHCC
```

Table 3: Multiple sequence alignment of HNH domains of Cas9 homologues: D8IJI4_LACSC (SEQ ID NO: 13), FOK1W6_LACD2 (SEQ ID NO: 14), D4FGK2_9LACO (SEQ ID NO: 15), E1NX12_9LACO (SEQ ID NO: 16), E7NSW3_TREPH (SEQ ID NO: 17), H1D477_9FUSO (SEQ ID NO: 18), C2KFJ4_9LACO (SEQ ID NO: 19), K1MRU9_9LACO (SEQ ID NO: 20), E3ZTQ9_LISSE (SEQ ID NO: 21) with Cas9 Pyogenes (SEQ ID NO: 3). * corresponds to the predicted 5'-first and 3'-end positions of the HNH domain of *S. pyogenes* Cas9. Grey highlighted sequence: predicted DNA/RNA biding region (see example 3).

```
                                                                  *
D8IJI4_LACSC     510  -RKGKSKLTN TRYKKISETY EKITDELISE YELGKLQSKL DSKANNmr--
F0K1W6_LACD2       5  ---------- ---------- ---------- ---------- ----------
D4FGK2_9LACO       1  ---------- ---------- ---------- ---------- ----------
E1NX12_9LACO      82  SKEDHPKRKL SRKADLKQVY KDSKKQIISI IGKDKYQDLS NELDNK---D
E7NSW3_TREPH       4  GKEAEKGRTS SRYASIKALY ENCKQDLADY DA-------- -VLEQFkseE
H1D477_9FUSO     258  QEDMKKERKE SRKSTFLTLY KSIKEEGRDW IK-------- -EIENW---S
C2KFJ4_9LACO       1  ---------- ---------- ---------- ---------- ----------
K1MRU9_9LACO       1  ---------- ---------- ---------- ---------- ----------
E3ZTQ9_LISSE     309  ENQTTGKGKN NSKPRFTSLE KAIKELGSQI LK-------- -EHPT----D
Cas9 Pyogenes    766  ENQTTQKGQK NSRERMKRIE EGIKELGSQI LK-------- -EHPV----E

D8IJI4_LACSC     557  ------DIYY LYFMQLGRDM YTGEKINIDE L----HQYYD IDHIFPRSFI
F0K1W6_LACD2       5  ---------- ---------- ---------- -----AD-YD VDHIMPQSFV
D4FGK2_9LACO       1  ---------- -------RDA YTDKPINIDE V----SQYYD IDHILPQSFI
E1NX12_9LACO     129  DRDLRWDNLY LYYTQLGRSM YSLKPIDISE LMNKNL--YD QDHIFPKSKK
E7NSW3_TREPH      45  PLRLRSDKLY LYYTQLGRCM YTGRVIDIDR LMSDNSA-YD IDHIYPRSKI
H1D477_9FUSO     296  DSEFRSKKLY LYYTQMGKCM YTGEKISLDQ LFNKNI--YD IDHIYPRSKI
C2KFJ4_9LACO       1  -------KYY LYFMQLGRDA YTGKPINIDE V----SQYYD IDHILPQSFI
K1MRU9_9LACO       1  ---------- ---MQLGRDA YTGKPINIDE V----SQYYD IDHILPQSFI
E3ZTQ9_LISSE     346  NQGLKNDRLY LYYLQNGKDM YTGQELDIHN L----SN-YD IDHVVPQSFI
Cas9 Pyogenes    803  NTQLQNEKLY LYYLQNGRDM YVDQELDINR L----SD-YD VDHIVPQSFL

D8IJI4_LACSC     597  KDNSLNNRVL TRKEINNNEK adrtaadlya v-------KM GDFWRKLRKQ
F0K1W6_LACD2      19  KDDSLDNRVL VARAVNNQKS DKVPALLFGN KVVADLGITV REMWDKWQKL
```

```
D4FGK2_9LACO       30     KDDSLNNRVL VAKPINNGKS DGVPLKLFGD NLATGLGITV KQMWNNWADK
E1NX12_9LACO      177     YDDSIENRVL VEKELNVKKS DIYPIsd-AN IIPQKIKGQV ESFWKMLYDH
E7NSW3_TREPH       94     KDDSLTNRVL VVKDANQDKR DEplSP-QIQ D-------KQ KGFWDFLKHN
H1D477_9FUSO      344     KDDSIENIVL VKRNINAKKT DEYPLErNIQ Q-------KQ HDFWKMLHSK
C2KFJ4_9LACO       40     KDDSLNNRVL VAKPINNGKS DGVPLKLFGD NLATGLGITV KQMWNNWADK
K1MRU9_9LACO       34     KDDSLNNRVL VAKPINNGKS DGVPLKLFGD NLATGLGITV KQMWNNWADK
E3ZTQ9_LISSE      391     TDNSIDNRVL ASSAANREKG DNVPSL-EVV R-------KR KVYWEKLYQA
Cas9 Pyogenes     848     KDDSIDNKVL TRSDKNRGKS DNVPSE-EVV K-------KM KNYWRQLLNA

D8IJI4_LACSC      640     GLITEKKYKN LLT--RTDSI DKYTKQSFIK RQLVETSQVV KMAANILQDK
F0K1W6_LACD2       69     GMISKRKLSN LLT--DPDAL TEYRAQGFIR RQLVETSQVI KLTATILQSE
D4FGK2_9LACO       80     GLINKAKQNN LFL--DPENI NKHQASGFIR KQLVETSQII KLATTILQAE
E1NX12_9LACO      226     KLIGDKKYAR LIR--SK-AF TDDELAGFIA RQLVETRQAT KETADLLKRL
E7NSW3_TREPH      136     NFISIEKYER LTY--RG-YF TEEMLSGFIA RQLVETRQGT KTAGQILEQL
H1D477_9FUSO      387     N--------- ---------- ---------- ---------- ----------
C2KFJ4_9LACO       90     GLINKAKQNN LFL--DPENI NKHQASGFIR KQLVETSQII KLATTILQAE
K1MRU9_9LACO       84     GLINKAKQNN LFL--DPENI NKHQASGFIR KQLVETSQII KLATTILQAE
E3ZTQ9_LISSE      433     KLMSKRKFDY LTKAERG-GL TEADKARFIH RQLVETRQIT KNVANILHQR
Cas9 Pyogenes     890     KLITQRKFDN LTKAERG-GL SELDKAGFIK RQLVETRQIT KHVAQILDSR
                                                                     *
D8IJI4_LACSC      688     YS-------- ---NTKIIEV RARLNSDLRK EYELIKNREV NDYHHAIDGY
F0K1W6_LACD2      117     FP-------- ---DSKIIEV PAKYNSIVRK QFDLYKSREV NDFHHAIDAY
D4FGK2_9LACO      128     YP-------- ---KTKIIVV KASSNHYLRN EFDLYKSREV NDYHHAIDAY
E1NX12_9LACO      273     CP-------- ---KSRIVYA KAQNASIFRQ KFDIPKSRTI NDLHHAQDAY
E7NSW3_TREPH      183     YP-------- ---DSTVVYC KAANTSEFRQ KFNLIKCREI NDLHHAHDAY
H1D477_9FUSO      388     ---------- ---------- ---------- ---------- ----------
C2KFJ4_9LACO      138     YP-------- ---KTKIIVV KASSNHYLRN EFDLYKSREV NDYHHAIDAY
K1MRU9_9LACO      132     YP-------- ---KTKIIVV KASSNHYLRN EFDLYKSREV NDYHHAIDAY
E3ZTQ9_LISSE      482     FNCKKDESGN VIEQVRIVTL KAALVSQFRK QFQLYKVREV NDYHHAHDAY
Cas9 Pyogenes     939     MNTKYDENDK LIREVKVITL KSKLVSDFRK DFQFYKVREI NNYHHAHDAY


D8IJI4_LACSC      727     LTIFIGQYLY KTYPKLRSYF VYDDFKKL-- -----D---- -----SNYLK
F0K1W6_LACD2      156     LSTIVGNYLY QVYPNLRRMF VYGEFKKFSS NaeESA---- -----HDVAR
D4FGK2_9LACO      167     LTTICGNLLY QAYPKLRPFF VYGQFKKFSS DP-KKE---- -----NEILK
E1NX12_9LACO      312     LNIVVGNIFD T--------- ------KFTQ DP-RNF---- -----IKNTK
E7NSW3_TREPH      222     LNIAVGNVYY T--------- ------KFTS NP-RNF---- -----MKl--
H1D477_9FUSO      388     ---------- ---------- ---------- ---------- ----------
C2KFJ4_9LACO      177     LTTICGNLLY QAYPKLRPFF VYGQFKKFSS DP-KKE---- -----NEILK
K1MRU9_9LACO      171     LTTICGNLLY QAYPKLRPFF VYGQFKKFSS DP-KKrk--- ----------
E3ZTQ9_LISSE      532     LNCVVANTLL KVYPQLEPEF VYGDYHQF-- -----Dwfka n--------K
Cas9 Pyogenes     989     LNAVVGTALI KKYPKLESEF VYGDYKVY-- -----Dvrkm iakseQEIGK
                                                 *
D8IJI4_LACSC      761     HMDKFNFIWK LEDKKAE-D- ---------- --VYDN-VNN EFILNVPKMK
F0K1W6_LACD2      197     RVKSMNFLDD LLRGTHG-D- ---------- --NIycrSTG EIVFNRNDII
D4FGK2_9LACO      207     KTKNFDFVAK LLGSKAP-N- ---------- --EIRS-QQG KVLFEKNKIR
E1NX12_9LACO      337     DSRNYNLe-- ---------- ---------- -----K-IYD YNVERNNYVA
E7NSW3_TREPH      245     ---------- ----KEP-Y- ---------- --NLRE-LFD RDVERNNTIA
H1D477_9FUSO      388     ---------- ---------- ---------- ---------- ----------
C2KFJ4_9LACO      217     KTKNFDFVAK LLGSKAP-N- ---------- --EIRS-QQG KVLFEKNKIR
K1MRU9_9LACO      207     ---------- ---------- ---------- ---------- ----------
E3ZTQ9_LISSE      567     ATAKKQFYTN IMLFFakkD- ---------- --RIID-ENG EILWDK-KYL
Cas9 Pyogenes    1032     ATAKYFFYSN IMNFFkteit langeirkrp liETNG-ETG EIVWDKGRDF
```

Table 4: Secondary structure predictions for the HNH domain and related HNH domain sequence of the *S. pyogenes* Cas9 (SEQ ID NO: 23). H represents helix, S represents sheet and C represents coil.

Sequence of Cas9 Pyogenes

```
PVENTQLQNEKLYLYYLQNGRDMYVDQELDINRLSDYDVDHIVPQSFLKDDSIDNKVLTRSDKNRGKSDNVPSEEVVKKMKNYWRQLLNAKLIT
QRKFDNLTKAERGGLSELDKAGFIKRQLVETRQITKHVAQILDSRMNTKYDENDKLIREVKVITLKSKLVSDFRKDFQFYKVREINNYHHAHDA
YLNAVVGTALIKKYPKLESEFVYGDYKVYDVRKMIAKSEQEIGKATAKYFFYSNIMNFFKTEITLANG
```

Secondary structure Prediction (Psipred)

```
CCCCCHHHHHHHHHHHHHHCCCCCCCCCCCCHCHCCCCCCCSSSCCCCCCCCCHHHHHHCCHHHHHHHCCCCHHHHHHHHHHHHHHHHHCCCCC
HHHHHHHHHHCCCCCCCHCHHHHHHHCCHHHHHHHHHHHHHHHHHHHHHHCCCCCCCCCCCSSSSSCCHHHHHHHHHCCCCCCCCCCCCHHHHHH
HHHHHCCHHHHHHHHHHCHHHHHCCCHHHHHHHHHCCCCCCHHHHHHCCCCCCCCHHCCHHHHHHHHHCC
```

Table 5: Multiple sequence alignment of shorter Cas9 homologues: D4IZM9_BUTFI (SEQ ID NO: 24), Q9CLT2_PASMU (SEQ ID NO: 25), E0G5X6_ENTFL (SEQ ID NO: 26), E0XXB7_9DELT (SEQ ID NO: 27) with Cas9 Pyogenes (SEQ ID NO: 3). * corresponds to the predicted 3'-end positions of the shorter Cas9 versions. Grey highlighted sequence: predicted DNA/RNA biding region (see example 3).

```
D4IZM9_BUTFI       1   mgi------- ---------- TIGLDLGVAS VGWAVVNDDY EILESCSNIF
Q9CLT2_PASMU       1   mqttnls--- ---------Y ILGLDLGIAS VGWAVVeine nedpigliDV
E0G5X6_ENTFL       1   MK-------- -------KDY VIGLDIGTNS VGWAVMTEDY QLVKKKMPIY
E0XXB7_9DELT       1   msskaidsle qldlfkpQEY TLGLDLGIKS IGWAILSGEr ia------NA
Cas9 Pyogenes      1   MD-------- -------KKY SIGLDIGTNS VGWAVITDEY KVPSKKFKVL

D4IZM9_BUTFI      34   ---------- PSADASK--- ---------- ---NSERRGF RQGRRLTRRR
Q9CLT2_PASMU      39   GVRIFERAEV PKTGESL--- ---------- ---ALSRRLA RSTRRLIRRR
E0G5X6_ENTFL      36   ---------- GNTEKKKIKK NFWGVRLFEE GHTAEDRRLK RTARRRISRR
E0XXB7_9DELT      45   GVYLFETAEE LNSTGNK--- ---------- ---LISKAAE RGRKRRIRRM
Cas9 Pyogenes     36   ---------- GNTDRhsIKK NLIGALLFDS GETAEATRLK RTARRRYTRR

D4IZM9_BUTFI      58   KNRIHDFQKL WEDKgf---- ---------- ---------- ----------
Q9CLT2_PASMU      73   AHRLLLAKRF LKREgilsti dlekglpnqa ---------- ----------
E0G5X6_ENTFL      76   RNRLRYLQAF FEEAMTDLDE NFFARLQESF LVPEDKKWHR HPIFakleDE
E0XXB7_9DELT      79   LDRkarrgrh iryll----- ---------- --------ER EGLPTDELEE
Cas9 Pyogenes     76   KNRICYLQEI FSNEMAKVDD SFFHRLEESF LVEEDKKHER HPIFGNIVDE

D4IZM9_BUTFI      74   VIPSQGTEDV LAIKIKGLS- -EKLSVDEVY WVLLNSLKHR GIsy----LD
Q9CLT2_PASMU     103   ---------- WELRVAGLE- -RRLSAIEWG AVLLHLIKHR GYLSKRKNES
E0G5X6_ENTFL     126   VAYHETYPTI YHLRKKLADS SEQADLRLIY LALAHIVKYR GHFLIEGKLS
E0XXB7_9DELT     106   VVVHQSNRTL WDVRAEAVE- -RKLTKQELA AVLFHLVRHR GYFPNTKKLP
Cas9 Pyogenes    126   VAYHEKYPTI YHLRKKLVDS TDKADLRLIY LALAHMIKFR GHFLIEGDLN

D4IZM9_BUTFI     118   DADSGDNSSD YAKSISRNEE ELKEKLpcei qwerlqkyga yrgnisived
Q9CLT2_PASMU     141   QTNNKELGAL LSGVAQNHQL LQsddyrtpa elalkkfake eghi-RNQRG
E0G5X6_ENTFL     176   TENISVKEQF QQFMIIYNQT Fvngesrlvs ap-------- ----------
E0XXB7_9DELT     154   PDDESDSADE EQGKINRATS RLREELkasd cktigqflaq nrdrqRNREG
Cas9 Pyogenes    176   PDNSDVDKLF IQLVQTYNQL FEenpinasg vdakailsar lsksrrlen-

D4IZM9_BUTFI     168   gepitlrnvf ttsaykKEVE QFIDTQAKYN AQYSGDFKAD YLEIFNRKRe
Q9CLT2_PASMU     190   AYTHTFNRLD LL----AELN LLFAQQHQFG NPHCKEhiqq ymtellmwqk
E0G5X6_ENTFL     208   -----LPESV LI----EEEL TEKASRTKKS EKVLQQFPQE KANGLFGQFL
E0XXB7_9DELT     204   DYSNLMARKL VF----EEAL QILAFQRKQG HELSKDFEKT YLDVLMGQRs
Cas9 Pyogenes    225   ---------- ---------- ---------- --LIAQLPGE KKNGLFGNLI

D4IZM9_BUTFI     218   ---------- ---------- ---------- ---------- ----------
Q9CLT2_PASMU     236   palsgeail- ---------- ---------- ---------- ----------
E0G5X6_ENTFL     249   KLMVGNKADF KKVFGLEEEA KItyasESYE EDLEGILAKV GDEYSDVFLA
E0XXB7_9DELT     250   grspk----- ---------- ---------- ---------- ----------
```

```
Cas9 Pyogenes     243   ALSLGLTPNF KSNFDLAEDA KLqlskDTYD DDLDNLLAQI GDQYADLFLA

D4IZM9_BUTFI      218   ---------- ---------- ---------- ---------- ----------
Q9CLT2_PASMU      245   ---------- ---------- ---------- ---------- ----------
E0G5X6_ENTFL      299   AKNVYDAVEL STILadsdkk shaklsssmi vRFTEHQEDL KKFKRFIREN
E0XXB7_9DELT      255   ---------- ---------- ---------- ---------- ----------
Cas9 Pyogenes     293   AKNLSDAILL SDILrvntei tkaplsasmi kRYDEHHQDL TLLKALVRQQ

D4IZM9_BUTFI      218   ---------- ---------- ---------- ---------- ----------
Q9CLT2_PASMU      245   ---------- ---------- ---------- ---------- ----------
E0G5X6_ENTFL      349   CPDEYDNLFK NEQKDGYAGY IahaGKVSQL KFYQYVKKII QDIAGAEYFL
E0XXB7_9DELT      255   ---------- ---------- ---------- ---------- ----------
Cas9 Pyogenes     343   LPEKYKEIFF DQSKNGYAGY Id--GGASQE EFYKFIKPIL EKMDGTEELL

D4IZM9_BUTFI      218   ---------- ---------- ---------- ---------- ----------
Q9CLT2_PASMU      245   ---------- ---------- ---------- ---------- ----------
E0G5X6_ENTFL      399   EKIaQENFLR KQRTFDNGVI PHQIHLAELQ AIIHRQaaYY PFLKENQEKI
E0XXB7_9DELT      255   ---------- ---------- ---------- ---------- ----------
Cas9 Pyogenes     391   VKLnREDLLR KQRTFDNGSI PHQIHLGELH AILRRQedFY PFLKDNREKI

D4IZM9_BUTFI      218   ---------Y YEGPgnelsr tdygkyttei nadgeyitvd nif-------
Q9CLT2_PASMU      245   ---------- ---------- ---------- ---------- ----------
E0G5X6_ENTFL      449   EQLVTFRIPY YVGPLSKGDa STFAWLKRQS EEPIRPWNLQ ETVDLDQSAT
E0XXB7_9DELT      255   ---------- ---------- ---------- ---------- ----------
Cas9 Pyogenes     441   EKILTFRIPY YVGPLARGN- SRFAWMTRKS EETITPWNFE EVVDKGASAQ

D4IZM9_BUTFI      252   ---DKLVGKC SVNPDERRAA GASYTAQEFN VLNDLNNLTI SSESsfi---
Q9CLT2_PASMU      245   ----KMLGKC THEKNEFKAA KHTYSAERFV WLTKLNNLRI LEDGAER-Al
E0G5X6_ENTFL      499   AFIERMTNFD TYLPSEKVLP KHSLLYEKFM VFNELTKISY TDDRGIK-AN
E0XXB7_9DELT      255   ------LGNC SLIPSELRAP SSAPSTEWFK FLQNLGNLQI SNAYREewsi
Cas9 Pyogenes     490   SFIERMTNFD knLPNEKVLP KHSLLYEYFT VYNELTKVKY VTEGMRKpAF

D4IZM9_BUTFI      296   ---------- ---EDGKLTE DAKRKIIeTI K----NAKTV NVKKIICdvi
Q9CLT2_PASMU      290   neeeRQLLIN HPYEKSKLTY AQVRKLLGLS EQAIFKHLRY SK--------
E0G5X6_ENTFL      548   FSGKEKEKIF DYLFKTRRKV --------KK K----DIIQF YR--------
E0XXB7_9DELT      299   daprRAQIID ACSQRSTSSY WQIRRDFQIP DEYRFNLVNY ER--------
Cas9 Pyogenes     540   LSGEQKKAIV DLLFKTNRKV --------TV K----QLKED YFKKIECfds

D4IZM9_BUTFI      329   gdkkcqisga riDKNEKEIF HSFE------ -----AYNKM RRALEEIGF-
Q9CLT2_PASMU      332   ---------- --ENAESATF MELK------ -----AWHAI RKALENQGLK
E0G5X6_ENTFL      578   ---------- --NEYNTEIV TLSGLEEDQF NASFSTYQDL LKc----GLT
E0XXB7_9DELT      341   ---------- --RDPDVDLQ EYLQQQERKT LANFRNWKQL EKiigtghpi
Cas9 Pyogenes     578   veisgv---- ---------- ------EDRF NASLGTYHDL LKIIKDKDF-

D4IZM9_BUTFI      367   ---DISSLSR ENLDLIGDIL TLNTDRESIL NAFNRKGIEL ADEAkdilvk
Q9CLT2_PASMU      359   DTWQDLAKKP DLLDEIGTAF SLYKTDEDIQ QYLTNKVPNS VINAL--LVS
E0G5X6_ENTFL      612   RAELDHPDNA EKLEDIIKIL TIFEDRQRIR TQLSTFKGQF SAEVLKKLER
E0XXB7_9DELT      379   ---------- QTLDEAARLI TLIKDDEKLS DQLADLLPEA SDKAITQLCE
Cas9 Pyogenes     607   ---LDNEENE DILEDIVLTL TLFEDREMIE ERLKTYAHLF DDKVMKQLKR

D4IZM9_BUTFI      414   vrktngsl-- ---------- FNKWQSFGLS IMNELIPELY AQPknqmell
Q9CLT2_PASMU      407   LNFDKFIELS LKSLRKILPL MEQGKRYDQA CREiyghhyg eanqktsqll
E0G5X6_ENTFL      662   KHYTGWGRL- ---------- ---------- ---------- ----------
E0XXB7_9DELT      419   LDFTTAAKIS LEAMYRILPH MNQGMGFFDA CQQESLPEIG VPPagdrvpp
Cas9 Pyogenes     654   RRYTGWGRL- ---------- ---------- ---------- ----------

D4IZM9_BUTFI      452   tamgvfksrg drfleckeip gdlivDDIYN PVVSKTVRIT VRILNALIKK
Q9CLT2_PASMU      457   paipaq---- ---------- ------EIRN PVVLRTLSQA RKVINAIIRQ
```

```
E0G5X6_ENTFL      671   ---------- ---------- ---------- ---------S KKLINGIYDK
E0XXB7_9DELT      469   f--------- ---------- -----DEMYN PVVNRVLSQS RKLINAVIDE
Cas9 Pyogenes     663   ---------- ---------- ---------- ---------S RKLINGIRDK

D4IZM9_BUTFI      502   YG-------- ---------- ---------- ---------- ----------
Q9CLT2_PASMU      487   YG-------- ---------- ---------- ---------- ----------
E0G5X6_ENTFL      682   ESGKTILGYL IKDdgvskhy NRNFMQLIND SQLSFKNAIQ KAQSSeheET
E0XXB7_9DELT      495   YG-------- ---------- ---------- ---------- ----------
Cas9 Pyogenes     674   QSGKTILDFL KSDgfa---- NRNFMQLIHD DSLTFKEDIQ KAQVSgqgDS

D4IZM9_BUTFI      504   ---------- ---------- ---------- ----Y-PDRV VIEMPRDK-N
Q9CLT2_PASMU      489   ---------- ---------- ---------- ----S-PARV HIETGRELGK
E0G5X6_ENTFL      732   LSETVNELAG SPAIKKGIYQ SLKIVDELVA IMGyA-PKRI VVEMAREN-Q
E0XXB7_9DELT      497   ---------- ---------- ---------- ----M-PAKI RVELARDLGK
Cas9 Pyogenes     720   LHEHIANLAG SPAIKKGILQ TVKVVDELVK VMGrhkPENI VIEMAREN-Q

D4IZM9_BUTFI      518   SDEEQQRLKK EQRDNENEIK DIKARVKTEY GREITEEDFR QHSKLSLKLK
Q9CLT2_PASMU      504   SFKERREIQK QQEDNRTKRE SAVQKFKELF SDFSSEPK-- --SKDILKFR
E0G5X6_ENTFL      780   TTSTGKRRSI QRLKIVEKAM AEIGSNL--- ---LKEQPTT NEQLRDTRLF
E0XXB7_9DELT      512   grELRERIKL DQLDKSKQnd ---QRAEDFR AEFQQAPR-- --GDQSLRYR
Cas9 Pyogenes     769   TTQKGQKNSR ERMKRIEEGI KELGSQI--- ---LKEHPVE NTQLQNEKLY

D4IZM9_BUTFI      568   LWNEQQGICP YSGKSIKIDD LL---Dnpnl FEVDHIIPLS ISFDDSRNNK
Q9CLT2_PASMU      550   LYEQQHGKCL YSGKEINIHR LNekgY---- VEIDHALPFS RTWDDSFNNK
E0G5X6_ENTFL      824   LYYMQNGKDM YTGDELSLHR LS---H---- YDIDHIIPQS FMKDDSLDNL
E0XXB7_9DELT      555   LWKEQNCTCP YSGRMIPVNS VLse-D---- TQIDHILPIS QSFDNSLSNK
Cas9 Pyogenes     813   LYYLQNGRDM YVDQELDINR LS---D---- YDVDHIVPQS FLKDDSIDNK

D4IZM9_BUTFI      615   VLVYSSENQD KGNRTPLAYL asvnrqwdih sfmdyvLKTY AGAQKRKKRD
Q9CLT2_PASMU      596   VLVLASENQN KGNQTPYEWL qgkinserwk nfvalvlgsq csaa------
E0G5X6_ENTFL      867   VLVGSTENRG KSDDVPSKEV VKDMKAYWek lyaAGLI--- ---SQRKFQR
E0XXB7_9DELT      600   VLCFTEENAQ KSNRTPFEYL daadfqr--- ------LEAI SGNWPEAKRN
Cas9 Pyogenes     856   VLTRSDKNRG KSDNVPSEEV VKKMKNYWrq llnAKLI--- ---TQRKFDN
                                                                *
D4IZM9_BUTFI      665   NLLNEQDITK VEVLQGFVNR NINDTRYASK VVLNSLQEYF SSK-------
Q9CLT2_PASMU      640   --KKQRLLTQ VIDDNKFIDR NLNDTRYIAR FLSNYIQENL llvgknkk--
E0G5X6_ENTFL      911   LTKGEQGGLT LEDKAHFIQR QLVETR---- ---------- ----------
E0XXB7_9DELT      641   KLLHKSfg-- -KVAEEWKSR ALNDTRYLTS ALADHLRHHL PDS-------
Cas9 Pyogenes     900   LTKAERGGLS ELDKAGFIKR QLVETRQITK HVAQILDSRM NTKydendkl
                             *
D4IZM9_BUTFI      708   ----ECSTkv kvirgsfthq mrvnlk---- ---------- ----------
Q9CLT2_PASMU      686   ----NVFTPN GQITALLRSR WGLIKARENN NRHHALDAIV VACATPSMQQ
E0G5X6_ENTFL      937   ---------- ---------- ---------- ---------- ----------
E0XXB7_9DELT      681   ----KIQTVN GRITGYLRKQ WGLEKDRDKH t-HHAVDAIV VACTTPAIVQ
Cas9 Pyogenes     950   irevKVITLK SKLVSDFRKD FQFYKVREIN NYHHAHDAYL NAVVGTALIK
```

Table 6: Secondary structure predictions of shorter Cas9 versions and related shorter *S. pyogenes* Cas9 sequence. H represents helix, S represents sheet and C represents coil.

Sequence of Cas9 Pyogenes

```
MDKKYSIGLDIGTNSVGWAVITDEYKVPSKKFKVLGNTDRHSIKKNLIGALLFDSGETAEATRLKRTARRRYTRRKNRICYLQEIFSNEMAKVD
DSFFHRLEESFLVEEDKKHERHPIFGNIVDEVAYHEKYPTIYHLRKKLVDSTDKADLRLIYLALAHMIKFRGHFLIEGDLNPDNSDVDKLFIQL
VQTYNQLFEENPINASGVDAKAILSARLSKSRRLENLIAQLPGEKKNGLFGNLIALSLGLTPNFKSNFDLAEDAKLQLSKDTYDDDLDNLLAQI
GDQYADLFLAAKNLSDAILLSDILRVNTEITKAPLSASMIKRYDEHHQDLTLLKALVRQQLPEKYKEIFFDQSKNGYAGYIDGGASQEEFYKFI
KPILEKMDGTEELLVKLNREDLLRKQRTFDNGSIPHQIHLGELHAILRRQEDFYPFLKDNREKIEKILTFRIPYYVGPLARGNSRFAWMTRKSE
ETITPWNFEEVVDKGASAQSFIERMTNFDKNLPNEKVLPKHSLLYEYFTVYNELTKVKYVTEGMRKPAFLSGEQKKAIVDLLFKTNRKVTVKQL
```

```
KEDYFKKIECFDSVEISGVEDRFNASLGTYHDLLKIIKDKDFLDNEENEDILEDIVLTLTLFEDREMIEERLKTYAHLFDDKVMKQLKRRRYTG
WGRLSRKLINGIRDKQSGKTILDFLKSDGFANRNFMQLIHDDSLTFKEDIQKAQVSGQGDSLHEHIANLAGSPAIKKGILQTVKVVDELVKVMG
RHKPENIVIEMARENQTTQKGQKNSRERMKRIEEGIKELGSQILKEHPVENTQLQNEKLYLYYLQNGRDMYVDQELDINRLSDYDVDHIVPQSF
LKDDSIDNKVLTRSDKNRGKSDNVPSEEVVKKMKNYWRQLLNAKLITQRKFDNLTKAERGGLSELDKAGFIKRQLVETRQITKHVAQILDSRMN
TKYDENDKLIREVKVITLKSKLVSDFRKDFQFYKVREINNY
```

Secondary structure Prediction (Psipred)

```
CCCCSSSSSSCCCCCSSSSSSCCCCCCCCCCCCCCCCCCCCCCCCCSSSSSCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHCCCC
CHHHHHHCCCCCCCCCCCCCCCCCCCCHHHHHHHHHHCCCHHHHHHHHHCCCCCCHHHHHHHHHHHHHHHCCCCCCCCCCCCCCCHHHHHHHHH
HHHHHHHCCCCCCCHHHHHHHHHHHCCCCHHHHHHHHHHCCCCCCHHHHHHHHHHHHHCCCCHHCCCCCCCCCCCCSCCCCHHHHHHHHHHHHH
CHHHHHHHHHHHHHHHHHHHHHHCCCCCCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHCHHHHHHHHHHHCCCCCCCCCCCCCCHHHHHHHHH
HHHHCCCCHHHHHHHHHHHCCCCCCCCCCCCCSSCHHHHHHHHHHHHHHHHHHCHHHCCHHHHHHHHHHSCCCCCCCCCCCCCCCCCCSSSCCC
CCCCCCCCCCHHCCCHHHHHHHHHCCCCCCCCCCCCCCCCCCHHHHHHHHHHHHHCSSSCCCCCCCCCCCCCHHHHHHHHHHHHHHCCCCHHHH
HHHHHHCCCCCCCCSCCCCCCCHCCCHHHHHHHHHHHCCCCCCCCCCHHHHHHHHHHHHHHCCHHHHHHHHHHHCCCCHHHHHHHHHHHCCCCC
HHHHHHHHHHHCCHHHCCCHHHHHHHHHHHHCCHHHHHHHCCCCHHHHHHHHCCCCCCCCHHHHHHHHCCCHHHHHHHHHHHHHHHHHHHHHHC
CCCCCSSSSSSCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHHHHCCCCCHHHHHHHHHHHHHHCCCCCCCCCCCCHCHCCCCCCCSSSCCCC
CCCCHHHHHHCHHHHHHHHHCCCHHHHHHHHHHHHHHHHHHHCCCCHHHHHHHHHHHCCCCCCCHCHHHHHHHHCCHHHHHHHHHHHHHHHHHH
HHCCCCCCCCCCCSSSSSCCHHHHHHHHHCCCCCCCCCCCCHHHHHHHHHHCHHHHHHHHHHCHHHHHCCHHHHHHHHHHCCCCCCHHHHHHHC
CCCCCCCHHCCHHHHHHHHCCCCCCCCCCCCCCCCCSSSSC
```

Table 7: List of DNA/RNA binding regions of *S.pyogenes* Cas9.

N°	SEQ ID	Cas 9 domain	amino acid positions	sequence of amino acids	secondary structure prediction	degree of solvent exposition
1	34	RuvC domain	S15-V20	SVGWAV	CEEEEE	C-terminal exposed
2	35	RuvC domain	S29-K33	SKKFK	CCCCC	not exposed
3	36	RuvC domain	T63-R78	TRLKRTARRRYTRRKNR	HHHHHHHHHHHHHHHH	not exposed
4	37	Interdomain	S213-R221	SARLSKSRR	HHCCCHHHH	not exposed
5	38	Interdomain	K314-Y325	KAPLSASMIKRY	CCCCCHHHHHHH	not exposed
6	39	Interdomain	F446-R467	FRIPYYVGPLARGNSRFAWMTR	HCCCCCCCCCCCCCCHHHHHHH	not exposed
7	40	Interdomain	T525-R535	TKVKYVTEGMR	HCEEEECCCCC	highly exposed
8	41	Interdomain	R557-K565	RKVTVKQLK	CCCCHHHHH	highly exposed
9	42	Interdomain	K652-K665	KRRRYTGWGRLSRK	HCCCCCCHHHHHHH	c- terminal highly exposed
10	43	Interdomain	Q768-R780	QTTQKGQKNSRER	CCCHHHHHHHHHH	highly exposed
11	44	HNH domain	R859-S867	RSDKNRGKS	CCCCCCCCC	average exposed
12	45	HNH domain	K878-A889	KKMKNYWRQLLNA	HHHHHHHHHHHH	not exposed
13	46	HNH domain	N979-Y988	NNYHHAHDAY	CCCHHHHHHH	not exposed
14	47	HNH domain	E1150-S1159	EKGKSKKLKS	EECCCCCCCEEH	not exposed
15	48	HNH domain	R1333-E1341	RKRYTSTKE	CCCCCCCCC	not exposed

Table 8: Multiple sequence alignment between Cas9 of S. pyogenes (SEQ ID NO: 61) and S.thermophilus (SEQ ID NO: 64) and the sequence of two pdb structures of RuvC domain of E.coli and T. thermophilus (SEQ ID NO: 62 and SEQ ID NO: 63).

```
               ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   5         15         25         35         45         55
4EP4:A|PDB     ---------- ---------- ----MVVAGI DPGITHLGLG VVAVEGKG-A LKARLLHG--
Cas9_sp        ---------- ---------- -MDKKYSIGL DIGTNSVGWA VITDEYKVPS KKFKVLGNTD
Cas9_S.The     MLFNKCIIIS INLDFSNKEK CMTKPYSIGL DIGTNSVGWA VITDNYKVPS KKMKVLGNTS
RuvC_E.Col     ---------- ---------- ---MAIILGI DPGSRVTGYG VIRQVGR--- -QLSYLGS--

               ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   65        75         85         95         105        115
4EP4:A|PDB     ---------- --------EV VKTSPQEPAK ERVGRIHARV LEVLHRFRPE AVAVEEQFFY
Cas9_sp        RHSIKKNLIG ALLFDSGETA EATRLKRTAR RRYTRRKNRI CYLQEIFSNE MAKVDDSFFH
Cas9_S.The     KKYIKKNLLG VLLFDSGITA EGRRLKRTAR RRYTRRRNRI LYLQEIFSTE MATLDDAFFQ
RuvC_E.Col     ---------- --------GC IRTKVDD-LP SRLKLIYAGV TEIITQFQPD YFAIEQVFMA

               ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  125        135        145        155        165        175
4EP4:A|PDB     RQNELAYKVG WALG------ --AVLVAAFE AGVPVYAYGP MQVKQA---- ----------
Cas9_sp        RLEESFLVEE DKKHERHPIF GNIVDEVAYH EKYPTIYHLR KKLVDSTDKA DLRLIYLALA
Cas9_S.The     RLDDSFLVPD DKRDSKYPIF GNLVEEKVYH DEFPTIYHLR KYLADSTKKA DLRLVYLALA
RuvC_E.Col     KNADSALKLG QARG------ --VAIVAAVN QELPVFEYAA RQVKQT---- ----------

               ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  185        195        205        215        225        235
4EP4:A|PDB     ---------- LAGHGHAAKE EVALMVRGIL G-----LKEA PRPSHLADAL AIALTHAFYA
Cas9_sp        HMIKFRGHFL IEGDLNPDNS DVDKLFIQLV QTYNQLFEEN PINASGVDAK AILSARLSKS
Cas9_S.The     HMIKYRGHFL IEGEFNSKNN DIQKNFQDFL DTYNAIFESD LSLENSKQLE EIVKDKISKL
RuvC_E.Col     ---------- VVGIGSAEKS QVQHMVRTLL K-----LPAN P-QADAADAL AIAITHCHVS

               ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  245        255        265        275        285        295
4EP4:A|PDB     R--MGTAKPL ---------- ---------- ---------- ---------- ----------
Cas9_sp        RRLENLIAQL PGEKKNGLFG NLIALSLGLT PNFKSNFDLA EDAKLQLSKD TYDDDLDNLL
Cas9_S.The     EKKDRILKLF PGEKNSGIFS EFLKLIVGNQ ADFRKCFNLD EKASLHFSKE SYDEDLETLL
RuvC_E.Col     QNAMQMSESR LNLARGRLR- ---------- ---------- ---------- ----------
```

Table 9: Multiple sequence alignment of the eight select sequences with Cas9 wild type of S. Pyogenes and Cas9 of S. Thermophilus and 4EP4 pdbcode. The position of the G247 is marked by a black arrow.

```
               ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   5         15         25         35         45         55         65
Cas9wt         ---------- ---------- -MDKKYSIGL DIGTNSVGWA VITDEYKVPS KKFKVLGNTD RHSIKKNLIG
D8IJI3         ---------- ---------- --MERYHIGL DIGTSSIGWA VIGDDFKIK- ---------- -RKKGKNLIG
F0K1W4         ---------- ---------M AKPKDYTIGL DIGTNSVGWV VTDDQNNIL- ---------- -RIKGKKAIG
C5F1Z4         ---------- ---------- ----MKILGF DIGIASIGWA FVENGELKD- ---------- -CGVRIFTKA
F3ZS86         ---------- ---------- ---MKKILGL DIGTNSVGWA VVNTNQEGEP SQIEKLGSRI IPMSQDILDK
H1D479         ---------- ---------M KKFENYYLGL DIGTSSIGWA VTNSQYDIL- ---------- -KFNGKYMWG
K1M766         ---------- --------MT KLNNEYMVGL DIGTNSCGWV ATDFDNNIL- ---------- -KMHGKRALG
Q7VG48         ---------- ---------- ----MRILGF DIGITSIGWA YVESNELKD- ---------- -CGVRIFTKA
E9S0G6         ---------- --------MK KEIKDYFLGL DVGTGSVGWA VTDTDYKLL- ---------- -KANRKDLWG
4EP4           ---------- ---------- ----MVVAGI DPGITHLGLG VVAVE----- ---------- ----GKGALK
G3ECR1         MLFNKCIIIS INLDFSNKEK CMTKPYSIGL DIGTNSVGWA VITDNYKVPS KKMKVLGNTS KKYIKKNLLG

               ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   75        85         95         105        115        125        135
Cas9wt         ALLFDSGETA EATRLKRTAR RRYTR---RK NRICYLQEIF SNEMAKVDDS FFHR-LEES- FLVEEDKKHE
D8IJI3         VRLFKEGDTA AERRSFRTQR RRLNR---RK WRLKLLEEIF DPYMAEVDEY FFAR-LKESN LSPKDSNKKY
F0K1W4         ARLFTEGKVA AERRSFRTTR RRLSR---RR WRIKMLEELF DEEIAKVDPS FFAR-LHESW ISPKDK-RKR
C5F1Z4         ENPKTGDSLA MPRREARSVR RRLAR---RK GRLETLKRLL AKEWDLCYED YIAADGELPK AFMG-KNLTN
F3ZS86         FGQGQTVSST ASRTDYRGIR RLRERSLLRR ERLHRVLHIL DFLPKHYADS IGWDPRNSKT YGKFLPGTEV
H1D479         TRLFPEANTA QERRIHRSSR RRLKR---RK ERIQILQMLF DKEIAKIDSG FFQR-LKDSK YYKEDKTEKQ
K1M766         SHLFDEGVSA ADRRAFRTTR RRIKR---RK WRLKLLEEIF DEEMAKVDPN FFAR-LKESG LSPLDT-RKN
```

```
Q7VG48        ENPKNGDSLA APRREARGAR RRLAR---RK ARLNAIKRLL CKEFELNLND YLANDGELPK AYQTSKDTKS
E9S0G6        MRCFETAETA EVRRLHRGAR RRIER---RK KRIKLLQELF SQEIAKTDEG FFQR-MKESP FYAEDKTILQ
4EP4          ARLLHGEVVK TSPQ--EPAK ERVGR---IH ARVLEVLHRF RPEAVAVEEQ FFYR------ --QNELAYKV
G3ECR1        VLLFDSGITA EGRRLKRTAR RRYTR---RR NRILYLQEIF STEMATLDDA FFQR-LDDS- FLVPDDKRDS

              ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                 145        155        165        175        185        195        205
Cas9wt        RHPIFGN-IV DEVAYHEKYP TIYHLRKKLV DSTDKADLRL IYLALAHMIK FRGHFLIEGD LNPDNSDVDK
D8IJI3        LGSLLFP-DI SDSNFYDKYP TIYHLRRDLM EKDKKFDLRE IYLAIHHIVK YRGNFLEKVP AKNYKNSGAS
F0K1W4        YSAIVFPSPE EDKKFHESYP TIYHLRDKLM KDDQKHDIRE IYIAVHQMIK ARGNFLHDES VETYRSGMSS
C5F1Z4        PYVLRYEALQ RLLSKEELVR VVLHIAKHRG YGNKNAKITK SEESKREQGK ILSALATNAS VIARYRTVGE
F3ZS86        KLAWVPTADG HQFLFYSTYL EMLEDLKQTQ AQLFETSQTP VPLDWTIYYL RKKALTQPIT KHELAWLLLH
H1D479        TNSIFHDKDY SDKEYHQDFP TIYHLRKFLL EGNKPKDIRF VYLALHHILT HRGHFLFPDM EVSNVTEFSN
K1M766        VSSIVFPTKK MDKQFYKKFP TIYHLRNALM KQDKKFDLRA IYIAIHHIVK YRGNFLSNSS ISNFSASKIE
Q7VG48        PYEL-YTAFH ---------- ---------- ---------- ---------W IIFAFCSIAS SLS-------
E9S0G6        ENALFNDRDF TDKTYHKAYP TINHLIKAWI ENKVKPDPRL LYLACHNIIK KRGHFLF-EG DFDSENQFDT
4EP4          GWALGAVLVA AFEAGVPVYA YGPMQVKQAL AGHGHAAKEE VALMVRGILG LKEAPRPSHL ADALAIALTH
G3ECR1        KYPIFGN-LV EEKVYHDEFP TIYHLRKYLA DSTKKADLRL VYLALAHMIK YRGHFLIEGE FNSKNNDIQK

              ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                 215        225        235        245        255        265        275
Cas9wt        LFI------- --QLVQTYNQ LFEENP---- ----INASGV DAKAILSARL SKSRRLENLI AQLP------
D8IJI3        IG-------- --FLLEEVNR FI-------- ---------- ---------- ---------- ----------
F0K1W4        LGGRSERNIL SVQTLEELND LFAENEGTEE VELNVASAEQ INDILTGGHL N-ADSQKEIS NLLLPSSFPS
C5F1Z4        YFYK------ --EFCEVIKN PQGLNT---- ---NENCTQP KVRVLKPIRN KGGEYTN--- ----------
F3ZS86        FNTKR----- --GYYQRRGE LEDTPT---- ----DKLVEY HALKVVDVEV DPEDQSK--- ----------
H1D479        IFS------- --ELKQYLYD EMDLDF---- ----EWKTEN ---------- ---------- ----------
K1M766        ID-------- --RFVNELND LYSIFLPESG VIFDAGNASK VEDIIRNEQM FKLDKIKEIA DVLP------
Q7VG48        ---------- ---------N RQ-------- ---------- ----MLPI-- ---------- ----------
E9S0G6        SIQ------- --AFFEYLRE DMEVDI---- ----DADSQK IKEILKDSSL KNSEKQSRLN KILG------
4EP4          AFY------- -----ARMGT AKPL------ ---------- ---------- ---------- ----------
G3ECR1        NFQ------- --DFLDTYNA IFESDL---- ----SLENSK QLEEIVKDKI SKLEKKDRIL KLFP------

              ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                 285        295        305        315        325        335        345
Cas9wt        ---------- GEKKNGLFGN LIALSLGLTP NFKSNFD--- --LAEDAKLQ LSK--DTYDD DLDNLLAQIG
D8IJI3        ---------- ---------- ---------- ---------- ---------- ---------- ----------
F0K1W4        FDDKAKEKQV KKLINNVATN ISKAWLGYKA DFSTILNLAK VDKDQKKIFA FALQGGDEED KVQELESLLE
C5F1Z4        ---------- -CILQEDLQR ELRCIFEHQK GFGFSITQEF QDKILKIAFY QRSLKDFSHL VGKCTFYPDE
F3ZS86        ---------- ---KPWYFVH LENGWIYKRQ SSEPLDNWKG LVKEFIVTTH LDKEGKPKLD KEGEVRRSFS
H1D479        ---------- ---------- ---------- ---------- ---------- ---------- ----------
K1M766        ---DTENKSG LKLSKKISKE ISKAILGYKA KFEIIL-QVN VDKTDSSIWN FKLNDENADV NLSEITSDLT
Q7VG48        ---------- ---------- ---------- ---------- ---------- ---------- ----------
E9S0G6        ---------- LKSSDKQKKA ITNLISGNKI NFADLYDNPD LKDAEKNSIS FSK--DDFDA LSDDLASILG
4EP4          ---------- ---------- ---------- ---------- ---------- ---------- ----------
G3ECR1        ---------- GEKNSGIFSE FLKLIVGNQA DFRKCFN--- --LDEKASLH FSK--ESYDE DLETLLGYIG
```

Table 10: Secondary structure elements prediction for the Cas9 wild type of S. Pyogenes sequence using PSIPRED. The sequence has been divided into the two split domains: N-terminal and C-terminal domain. In bold is marked the Leucine 248 which has been mutated to Valine in the sequence of the C-terminal domain.

Sequence of Cas9 Pyogenes

<u>N-terminal domain</u>

MDKKYSIGLDIGTNSVGWAVITDEYKVPSKKFKVLGNTDRHSIKKNLIGALLFDSGETAEATRLKRTARRRYTRRKNRI
CYLQEIFSNEMAKVDDSFFHRLEESFLVEEDKKHERHPIFGNIVDEVAYHEKYPTIYHLRKKLVDSTDKADLRLIYLALA
HMIKFRGHFLIEGDLNPDNSDVDKLFIQLVQTYNQLFEENPINASGVDAKAILSARLSKSRRLENLIAQLPGEKKNGLF
GNLIALSLG

<u>C-terminal domain</u>

LTPNFKSNFDLAEDAKLQLSKDTYDDDLDNLLAQIGDQYADLFLAAKNLSDAILLSDILRVNTEITKAPLSASMIKRYD
EHHQDLTLLKALVRQQLPEKYKEIFFDQSKNGYAGYIDGGASQEEFYKFIKPILEKMDGTEELLVKLNREDLLRKQRTF

DNGSIPHQIHLGELHAILRRQEDFYPFLKDNREKIEKILTFRIPYYVGPLARGNSRFAWMTRKSEETITPWNFEEVVDK
GASAQSFIERMTNFDKNLPNEKVLPKHSLLYEYFTVYNELTKVKYVTEGMRKPAFLSGEQKKAIVDLLFKTNRKVTVK
QLKEDYFKKIECFDSVEISGVEDRFNASLGTYHDLLKIIKDKDFLDNEENEDILEDIVLTLTLFEDREMIEERLKTYAHLFD
DKVMKQLKRRRYTGWGRLSRKLINGIRDKQSGKTILDFLKSDGFANRNFMQLIHDDSLTFKEDIQKAQVSGQGDSL
HEHIANLAGSPAIKKGILQTVKVVDELVKVMGRHKPENIVIEMARENQTTQKGQKNSRERMKRIEEGIKELGSQILK
EHPVENTQLQNEKLYLYYLQNGRDMYVDQELDINRLSDYDVDHIVPQSFLKDDSIDNKVLTRSDKNRGKSDNVPSE
EVVKKMKNYWRQLLNAKLITQRKFDNLTKAERGGLSELDKAGFIKRQLVETRQITKHVAQILDSRMNTKYDENDKLI
REVKVITLKSKLVSDFRKDFQFYKVREINNY

Secondary structure Prediction (Psipred)

N-terminal domain

CCCCSSSSSSSCCCCSSSSSSCCCCCCCCCCCCCCCCCCCCCCCCCSSSSSSCCCCCCHHHHHHHHHHHHHHHHHH
HHHHHHHHHHHHHHCCCCCHHHHHHCCCCCCCCCCCCCCCCCCCCHHHHHHHHHCCCHHHHHHHHHHCCC
CCCCHHHHHHHHHHHHHHCCCCCCCCCCCCCCCHHHHHHHHHHHHHHHHCCCCCCCHHHHHHHHHHHCCC
CHHHHHHHHHCCCCCCCHHHHHHHHHHHHHC

C-terminal domain

CCCCHHCCCCCCCCCCCSCCCCHHHHHHHHHHHHHCHHHHHHHHHHHHHHHHHHHHHHCCCCCCCCCCCCC
HHHHHHHHHHHHHHHHHHHHHHHCCHHHHHHHHHHCCCCCCCCCCCCCCCHHHHHHHHHHHHCCCCCHH
HHHHHHHHCCCCCCCCCCCCCSSCHHHHHHHHHHHHHHHHHHCHHHCCCHHHHHHHHHSCCCCCCCCCCCC
CCCCCCSSSCCCCCCCCCCCCCHHCCCHHHHHHHHCCCCCCCCCCCCCCCCCCHHHHHHHHHHHHHCSSSCCC
CCCCCCCCCHHHHHHHHHHHHHHCCCCCHHHHHHHHHHCCCCCCCCSCCCCCCCCHCCCHHHHHHHHHHCC
CCCCCCCCCHHHHHHHHHHHHHCCCHHHHHHHHHHHCCCCCHHHHHHHHCCCCCCHHHHHHHHHHHCCH
HHCCCCHHHHHHHHHHCCHHHHHHHHCCCCCHHHHHHHCCCCCCCCCHHHHHHHCCCCHHHHHHHHHHH
HHHHHHHHHHCCCCCCSSSSSSCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHHHHCCCCCCHHHHHHH
HHHHHHCCCCCCCCCCCCHCHCCCCCCCSSSCCCCCCCCCHHHHHHCCHHHHHHHCCCCHHHHHHHHHHHHH
HHHHHCCCCCHHHHHHHHHHCCCCCCCHCHHHHHHHCCHHHHHHHHHHHHHHHHHHHHHCCCCCCCCCC
CSSSSSCCHHHHHHHHHCCCCCCCCCCCCHHHHHHHHHHHCCHHHHHHHHHCHHHHHCCCHHHHHHHHHC
CCCCCHHHHHHCCCCCCCCCHHCCHHHHHHHCCCCCCCCCCCCCCCCCSSSSC

REFERENCES

Buchan, D. W., S. M. Ward, et al. (2010). "Protein annotation and modelling servers at University College London." Nucleic Acids Res **38**(Web Server issue): W563-8.

Cong, L., F. A. Ran, et al. (2013). "Multiplex genome engineering using CRISPR/Cas systems." Science **339**(6121): 819-23.

Critchlow, S. E. and S. P. Jackson (1998). "DNA end-joining: from yeast to man." Trends Biochem Sci **23**(10): 394-8.

Dalgaard, J. Z., A. J. Klar, et al. (1997). "Statistical modeling and analysis of the LAGLIDADG family of site-specific endonucleases and identification of an intein that encodes a site-specific endonuclease of the HNH family." Nucleic Acids Res **25**(22): 4626-38.

de Castro, E., C. J. Sigrist, et al. (2006). "ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins." Nucleic Acids Res **34**(Web Server issue): W362-5.

Deltcheva, E., K. Chylinski, et al. (2011). "CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III." Nature **471**(7340): 602-7.

Deveau, H., R. Barrangou, et al. (2008). "Phage response to CRISPR-encoded resistance in Streptococcus thermophilus." J Bacteriol **190**(4): 1390-400.

Eickholt, J., X. Deng, et al. (2011). "DoBo: Protein domain boundary prediction by integrating evolutionary signals and machine learning." BMC Bioinformatics **12**: 43.

Garneau, J. E., M. E. Dupuis, et al. (2010). "The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA." Nature **468**(7320): 67-71.

Gasiunas, G., R. Barrangou, et al. (2012). "Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria." Proc Natl Acad Sci U S A **109**(39): E2579-86.

Gorbalenya, A. E. (1994). "Self-splicing group I and group II introns encode homologous (putative) DNA endonucleases of a new family." Protein Sci **3**(7): 1117-20.

Haft, D. H., J. Selengut, et al. (2005). "A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes." PLoS Comput Biol **1**(6): e60.

Horvath, P. and R. Barrangou (2010). "CRISPR/Cas, the immune system of bacteria and archaea." Science **327**(5962): 167-70.

Jinek, M., K. Chylinski, et al. (2012). "A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity." Science **337**(6096): 816-21.

Jones, D. T. (1999). "Protein secondary structure prediction based on position-specific scoring matrices." J Mol Biol **292**(2): 195-202.

Kleanthous, C., U. C. Kuhlmann, et al. (1999). "Structural and mechanistic basis of immunity toward endonuclease colicins." Nat Struct Biol **6**(3): 243-52.

Lutz, S., M. Ostermeier, et al. (2001). "Rapid generation of incremental truncation libraries for protein engineering using alpha-phosphothioate nucleotides." Nucleic Acids Res **29**(4): E16.

Ma, J. L., E. M. Kim, et al. (2003). "Yeast Mre11 and Rad1 proteins define a Ku-independent mechanism to repair double-strand breaks lacking overlapping end sequences." Mol Cell Biol **23**(23): 8820-8.

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

Mali, P., L. Yang, et al. (2013). "RNA-guided human genome engineering via Cas9." Science **339**(6121): 823-6.

Mojica, F. J., C. Diez-Villasenor, et al. (2009). "Short motif sequences determine the targets of the prokaryotic CRISPR defence system." Microbiology **155**(Pt 3): 733-40.

Morgenstern, B. (2004). "DIALIGN: multiple DNA and protein sequence alignment at BiBiServ." Nucleic Acids Res **32**(Web Server issue): W33-6.

Qi, L. S., M. H. Larson, et al. (2013). "Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression." Cell **152**(5): 1173-83.

Reyon, D., S. Q. Tsai, et al. (2012). "FLASH assembly of TALENs for high-throughput genome editing." Nat Biotechnol **30**(5): 460-5.

Sapranauskas, R., G. Gasiunas, et al. (2011). "The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli." Nucleic Acids Res **39**(21): 9275-82.

Schutz, K., J. R. Hesselberth, et al. (2010). "Capture and sequence analysis of RNAs with terminal 2',3'-cyclic phosphates." Rna **16**(3): 621-31.

Shub, D. A., H. Goodrich-Blair, et al. (1994). "Amino acid sequence motif of group I intron endonucleases is conserved in open reading frames of group II introns." Trends Biochem Sci **19**(10): 402-4.

Sorek, R., C. M. Lawrence, et al. (2013). "CRISPR-mediated Adaptive Immune Systems in Bacteria and Archaea." Annu Rev Biochem.

van der Ploeg, J. R. (2009). "Analysis of CRISPR in Streptococcus mutans suggests frequent occurrence of acquired immunity against infection by M102-like bacteriophages." Microbiology **155**(Pt 6): 1966-76.

Wang, L. and S. J. Brown (2006). "BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences." Nucleic Acids Res **34**(Web Server issue): W243-8.

CLAIMS

1. A split Cas9 comprising a HNH domain but no RuvC domain, wherein said HNH domain comprises at least one HNH motif sequence Y-X-X-D-H-X-X-P-X-S-X-X-X-D-X-S, where X is anyone of the 20 natural amino acids.

2. A split Cas9 comprising a RuvC domain, but no HNH domain, wherein said RuvC domain comprises at least one motif sequence D-[I/L]-G-X-X-S-X-G-W-A, where X is anyone of the 20 natural amino acids.

3. A split Cas9 according to claim 1 or 2, which is less than 500 amino acids long.

4. A Cas9 variant of less than 1100 amino acids, comprising RuvC and HNH domains, said domains comprising at least one RuvC motif sequence D-[I/L]-G-X-X-S-X-G-W-A or one HNH motif sequence Y-X-X-D-H-X-X-P-X-S-X-X-X-D-X-S, where X is anyone of the 20 natural amino acids.

5. The Cas9 variant or split Cas9 according to any one of claim 1 to 4, wherein the C-terminal domain of the Cas9 variant is truncated after the HNH motif sequence Y-X-X-D-H-X-X-P-X-S-X-X-X-D-X-S.

6. The Cas9 variant or split Cas9 according to any one of claims 1 to 5 wherein said RuvC domain comprises an amino acid sequence selected from the group consisting of: SEQ ID NO: 4 to SEQ ID NO: 12

7. The Cas9 variant or split Cas9 of claim 6, wherein said RuvC domain comprises at least 80%, preferably at least 85%, more preferably 90% sequence identity with the amino acid sequence selected from the group consisting of: SEQ ID NO: 4 to SEQ ID NO: 12.

8. The Cas9 variant or split Cas9 according to any one of claims 1 to 7, wherein said HNH domain, comprises an amino acid sequence selected from the group consisting of: SEQ ID NO: 13 to SEQ ID NO: 25.

9. The Cas9 variant or split Cas9 of claim 8 wherein said HNH domain, comprises at least 80%, preferably at least 85%, more preferably 90% sequence identity with the amino acid sequence selected from the group consisting of: SEQ ID NO: 13 to SEQ ID NO: 25.

10. The Cas9 variant according to any one of claims 1 to 9, wherein the RuvC domain and the HNH domain are separated by a peptide linker.

11. The Cas9 variant of claim 10, wherein the peptide linker comprises the amino acid sequence selected from the group of: SEQ ID NO: 49 to SEQ ID NO: 50.

12. The Cas9 variant according to claims 1 to 11, wherein said variant comprises amino acid sequence selected from the group consisting of: SEQ ID NO: 26 to SEQ ID NO: 33.

13. The Cas9 variant or split Cas9 according to claims 1 to 12 comprising either a non-functional RuvC or HNH catalytic domain.

14. The Cas9 variant or split Cas9 according to any one of claims 1 to 13 comprising at least one residue mutated in the RNA/DNA binding region wherein the DNA/RNA binding region is selected from the amino acid sequence consisting of: SEQ ID NO: 34 to SEQ ID NO: 48.

15. The Cas9 variant or split Cas9 according to any one of claims 1 to 14 comprising at least one residue mutated in the RNA/DNA binding region wherein the DNA/RNA binding region is selected from the amino acid sequence consisting of: SEQ ID NO: 34 to SEQ ID NO: 48.

16. Use of a Cas9 variant of split Cas9 as defined in any one of claims 1 to 15, to form a complex with at least one guide RNA on a target nucleic acid sequence, in order to cleave said target nucleic acid sequence.

17. A method of genome targeting in a cell comprising:

(a) selecting a target nucleic acid sequence, optionally comprising a PAM motif,

(b) providing a guide RNA comprising a sequence complementary to the target nucleic acid sequence;

(c) providing a Cas9 variant or at least one split Cas9 according to any one of claims 1 to 15;

(d) introducing into the cell said guide RNA and said Cas9 variant or split Cas9; such that Cas9 or split Cas9 processes the target nucleic acid sequence in the cell.

18. The method of genome targeting in a cell of claim 17 comprising:

(a) selecting a target nucleic acid sequence, optionally comprising a PAM motif,

(b) providing a crRNA comprising a sequence complementary to the target nucleic acid sequence and having a 3' extension sequence;

(c) providing a TracrRNA comprising a sequence complementary to a part of the 3'extension of said crRNA;

(d) providing a Cas9 variant or at least one split Cas9 according to any one of claims 1 to 15;

(e) introducing into the cell said crRNA, said TracrRNA and said Cas9 variant or split Cas9; such that Cas9-tracrRNA:crRNA complex process the target nucleic acid sequence in the cell.

19. The method of claim 18, wherein the crRNA and the tracrRNA are fused to form a single guided RNA.

20. The method according to any one of claims 17 to 19, further comprising introducing an exogenous nucleic acid sequence comprising at least one sequence homologous to at least a portion of the target nucleic acid sequence.

21. The method of any one of claims 17 to 20, wherein the cell is a plant cell.

22. The method of any one of claims 17 to 20, wherein the cell is a mammalian cell.

23. An isolated cell comprising a Cas9 variant or a split Cas9 according to any one of claims 1 to 15.

24. A method for generating an animal comprising:

(a) providing a eukaryotic cell comprising a target nucleic acid sequence into which it is desired to introduce a genetic modification;

(b) processing said target nucleic acid sequence into said cell by the method according to any one of claims 17 to 20; and

(c) generating an animal from the cell or progeny thereof, in which a cleavage has occurred.

25. A method of claim 24, further comprising: introducing into the cell an exogenous nucleic acid comprising a sequence homologous to at least a portion of the target nucleic acid sequence

and generating an animal from the cell or progeny thereof in which homologous recombination has occurred.

26. A method for generating a plant comprising:

(a) providing a plant cell comprising a target nucleic acid sequence into which it is desired to introduce a genetic modification;

(b) processing said target nucleic acid sequence into said cell by the method according to any one of claims 17 to 20; and

(c) generating a plant from the cell or progeny thereof in which a cleavage has occurred.

27. The method of claim 26, further comprising: introducing into the plant cell an exogenous nucleic acid comprising a sequence homologous to at least a portion of the target nucleic acid sequence; and generating a plant from the cell or progeny thereof in which homologous recombination has occurred.

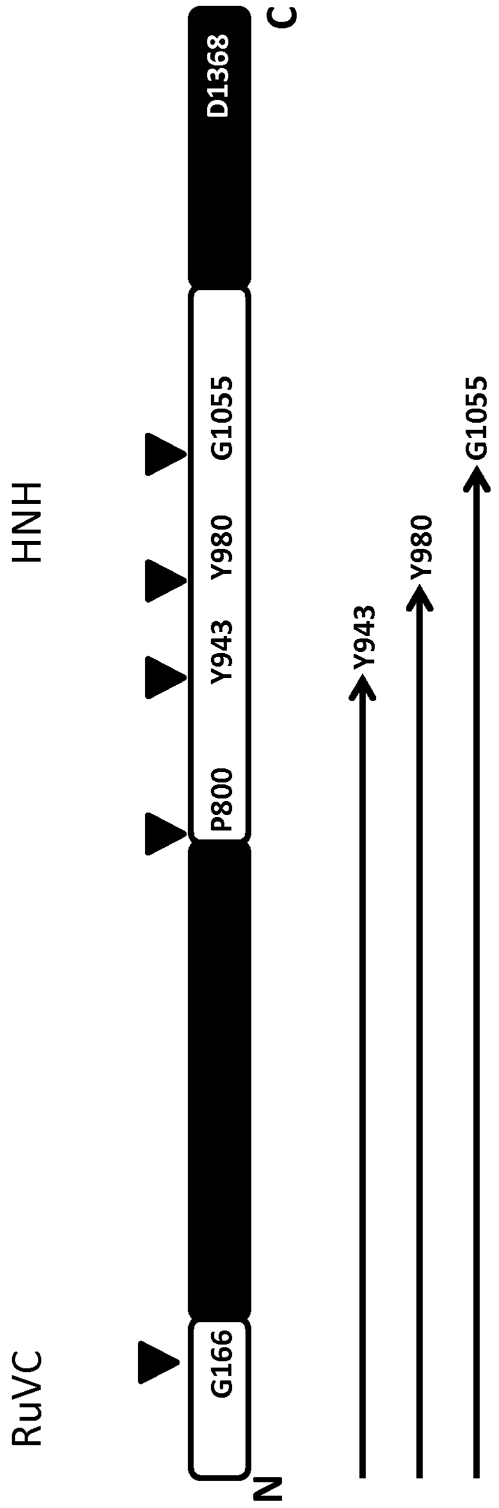

Figure 1

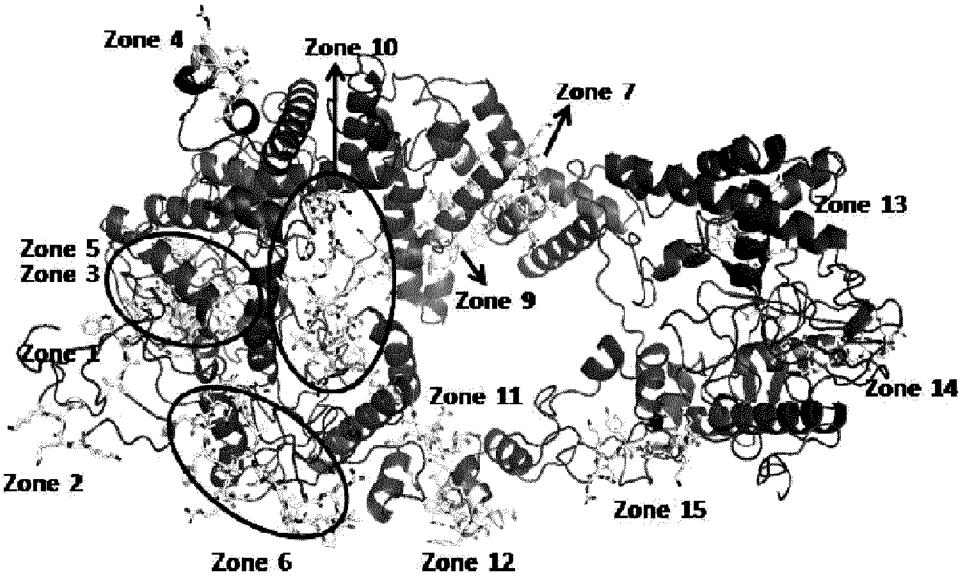
Zone 4
Zone 10
Zone 7
Zone 13
Zone 5
Zone 3
Zone 9
Zone 1
Zone 14
Zone 11
Zone 2
Zone 15
Zone 6
Zone 12

Figure 2

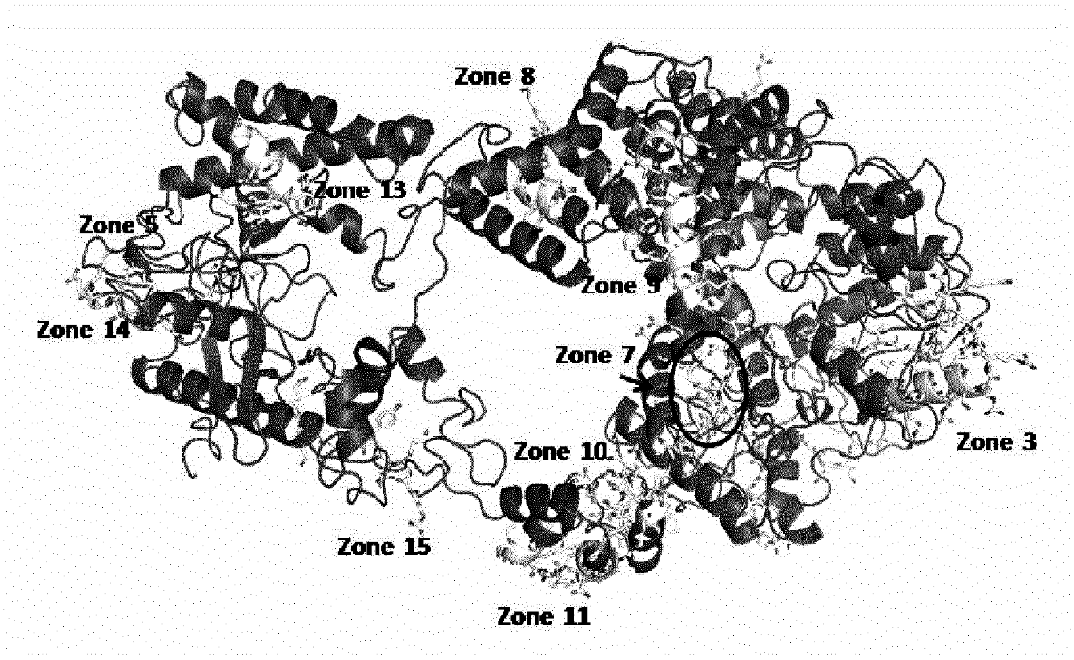
Zone 8
Zone 13
Zone 5
Zone 9
Zone 14
Zone 7
Zone 3
Zone 10
Zone 15
Zone 11

Figure 3

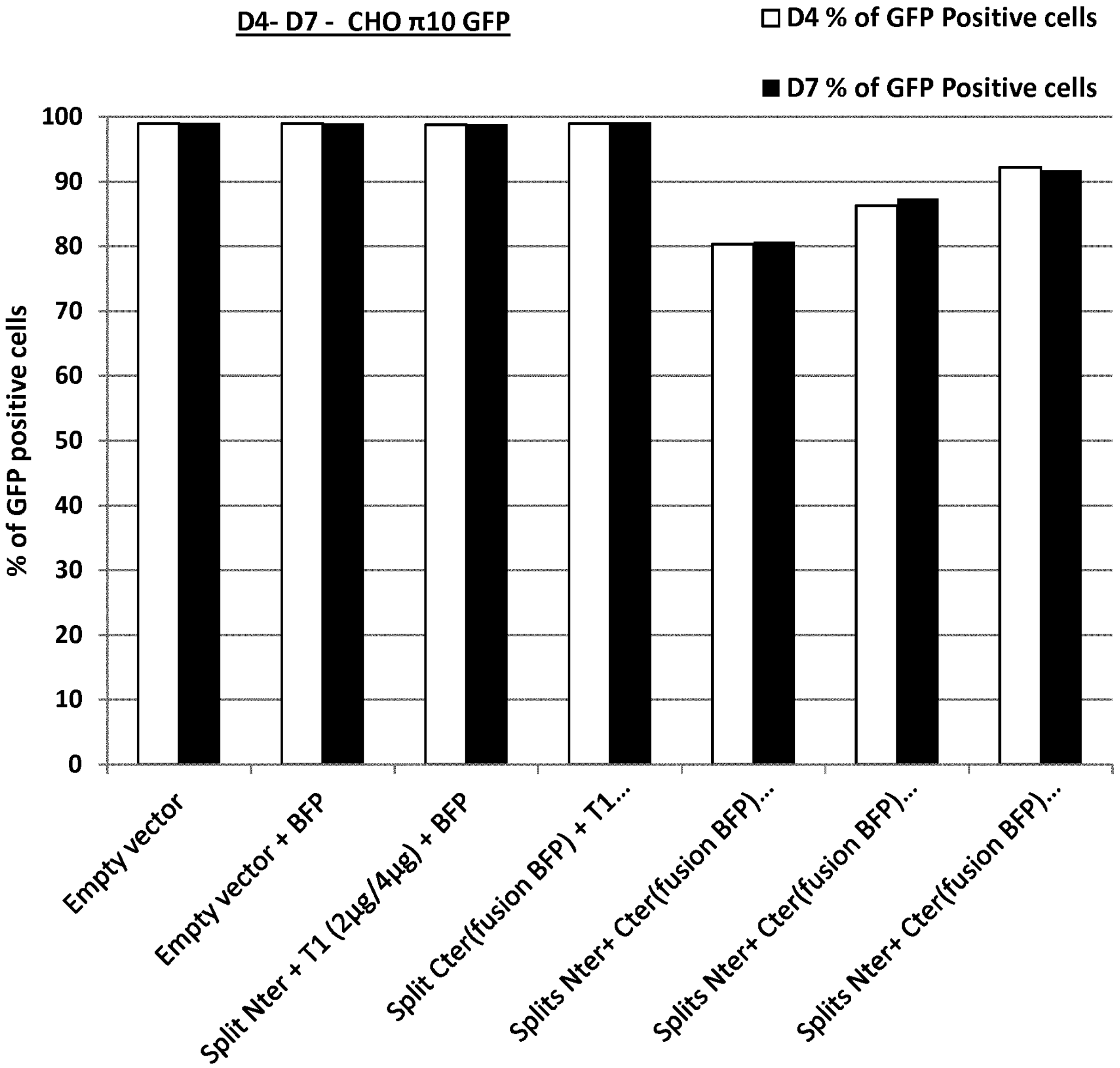
D4- D7 - CHO π10 GFP
□ D4 % of GFP Positive cells
■ D7 % of GFP Positive cells
% of GFP positive cells
100
90
80
70
60
50
40
30
20
10
0
Empty vector
Empty vector + BFP
Split Nter + T1 (2µg/4µg) + BFP
Split Cter(fusion BFP) + T1...
Splits Nter+ Cter(fusion BFP)...
Splits Nter+ Cter(fusion BFP)...
Splits Nter+ Cter(fusion BFP)...

Figure 4

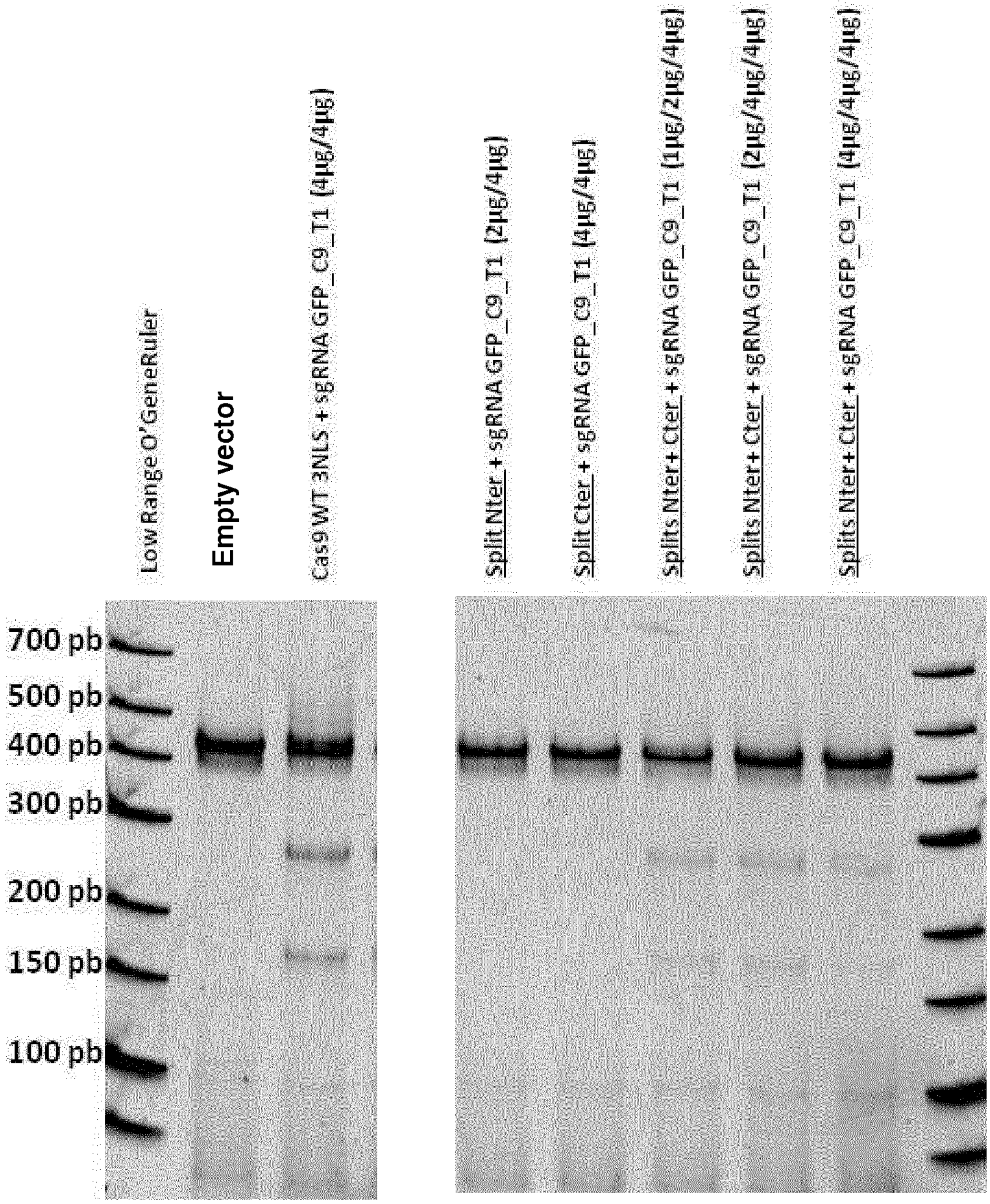
Low Range O'GeneRuler
Empty vector
Cas9 WT 3NLS + sgRNA GFP_C9_T1 (4µg/4µg)
Split Nter + sgRNA GFP_C9_T1 (2µg/4µg)
Split Cter + sgRNA GFP_C9_T1 (4µg/4µg)
Splits Nter+ Cter + sgRNA GFP_C9_T1 (1µg/2µg/4µg)
Splits Nter+ Cter + sgRNA GFP_C9_T1 (2µg/4µg/4µg)
Splits Nter+ Cter + sgRNA GFP_C9_T1 (4µg/4µg/4µg)
700 pb
500 pb
400 pb
300 pb
200 pb
150 pb
100 pb

Figure 5

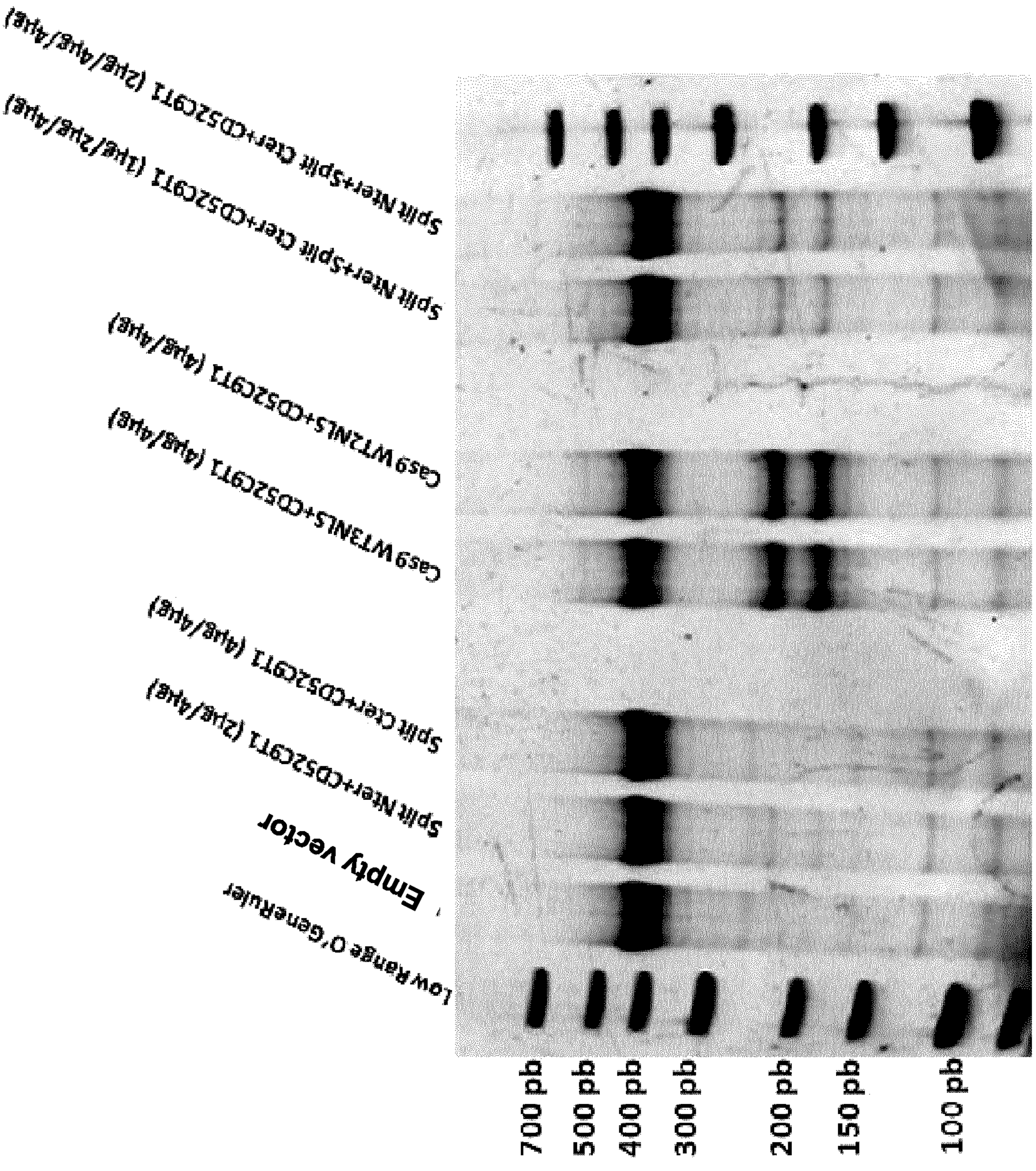
Low Range O'GeneRuler
Empty vector
Split Nter+CD52C9T1 (2µg/4µg)
Split Cter+CD52C9T1 (4µg/4µg)
Cas9 WT3NLS+CD52C9T1 (4µg/4µg)
Cas9 WT2NLS+CD52C9T1 (4µg/4µg)
Split Nter+Split Cter+CD52C9T1 (1µg/2µg/4µg)
Split Nter+Split Cter+CD52C9T1 (2µg/4µg/4µg)
700 pb
500 pb
400 pb
300 pb
200 pb
150 pb
100 pb

Figure 6