THERMOSTABLE CAS9 NUCLEASES

FIELD OF THE INVENTION

The present invention relates to the field of genetic engineering and more particularly to nucleic acid editing and genome modification. The present invention concerns genetic engineering tools in the form of nucleases which can be configured for sequence-directed site-specific nicking or cutting of genetic material; also ribonucleoproteins which exert nuclease activity on sequence specific sites of genetic material, and modified nucleases and ribonucleoproteins for use as markers. The invention therefore also concerns associated expression constructs for delivery and expression of nucleases and guide RNAs within cells. Further, the invention concerns the sequence-specific editing of nucleic acids *in vitro* or *in vivo* and methods used to achieve that. A particular area to which the invention relates is the genetic manipulation of thermophilic organisms, particularly microorganisms.

BACKGROUND TO THE INVENTION

It was first demonstrated in 2007 that CRISPR-Cas is an adaptive immune system in many bacteria and most archaea (Barrangou *et al.*, 2007, Science 315: 1709-1712), Brouns *et al.*, 2008, Science 321: 960-964). Based on functional and structural criteria, three types of CRISPR-Cas systems have so far been characterized, most of which use small RNA molecules as guide to target complementary DNA sequences (Makarova *et al.*, 2011, Nat Rev Microbiol 9: 467-477; Van der Oost *et al.*, 2014, Nat Rev Microbiol 12: 479-492).

In a recent study by the Doudna/Charpentier labs, a thorough characterization of the effector enzyme of the type II CRISPR-Cas system (Cas9) was performed, including demonstration that the introduction of designed CRISPR RNA guides (with specific spacer sequences) targets complementary sequences (protospacers) on a plasmid, causing double strand breaks of this plasmid (Jinek *et al.*, 2012, Science 337: 816-821). Following Jinek *et al.*, 2012, Cas9 is used as a tool for genome editing.

Cas9 has been used to engineer the genomes of a range of eukaryotic cells (e.g. fish, plant, man) (Charpentier and Doudna, 2013, Nature 495: 50-51).

In addition, Cas9 has been used to improve yields of homologous recombination in bacteria by selecting for dedicated recombination events (Jiang *et al.*, 2013, Nature Biotechnol 31: 233-239). To achieve this, a toxic fragment (Targeting construct) is co-transfected with a rescuing fragment carrying the desired alteration (Editing construct, carrying point mutation or deletions). The Targeting construct consists of Cas9 in combination with a design CRISPR and an antibiotic resistance marker, defining the site of the desired recombination on the host chromosome; in the presence of the corresponding antibiotic, integration of the Targeting construct in the host chromosome is selected for. Only when the additional recombination occurs of the Editing construct with the CRISPR target site elsewhere on the host chromosome, the host can escape from the auto-immunity problem. Hence, in the presence of the antibiotic, only the desired (marker-free) mutants are able to survive and grow. A related strategy to select for subsequent removal of the integrated Targeting construct from the chromosome is presented as well, generating a genuine marker free mutant.

It has been established in recent years that CRISPR-Cas mediated genome editing constitutes a useful tool for genetic engineering. It has been established that the prokaryotic CRISPR systems serve their hosts as adaptive immune systems (Jinek *et al.*, 2012, Science 337: 816-821) and can be used for quick and effective genetic engineering (Mali *et al.*, 2013, Nat Methods 10:957-963, for example), requiring only modification of the guide sequence in order to target sequences of interest.

However, there is a continuing need for the development of agents with improved sequence-specific nucleic acid detection, cleavage and manipulation under a variety of experimental conditions for application in the area of genetic research and genome editing. In particular, currently available sequence-specific genome editing tools, including Cas9, are not applicable for use in all conditions or organisms, for example, sequence-specific nucleases are relatively thermo-sensitive and therefore not applicable for use in strictly thermophilic microorganisms (which are capable of growth between 41°C and 122°C and grow optimally in the ranges of temperatures

from >60°C to 80°C with hyperthermophiles capable of optimal growth above 80°C), for example, microorganisms that are used in industrial fermentations or for *in vitro* laboratory processes conducted at elevated temperatures.

To date there is no experimental evidence for active Cas9 proteins in thermophiles. Based on a comparative genome screening by Chylinski *et al*. (2014; Nucleic Acids Research 42: 6091-61-05) on the presence of Cas9 in bacteria it was found that the Type II-C CRISPR-Cas system is only present in approximately 3.3% of all bacterial genomes. Among thermophilic bacteria, the Type II system is underrepresented based on statistical analysis (P=0.0019). In addition, no Type II system has been found in archaea however, this could possibly be due to the absence of the RNase III protein (involved in the Type II system) in archaea. Chylinski, *et al*., (2014; Nucleic Acids Research 42: 6091-6105) did describe the classification and evolution of type II CRISPR-Cas systems, in particular, two species are identified which exhibit these systems, however these species grow maximally at 55°C and do not exhibit strictly thermophilic growth with optimum growth temperature 60-80°C, with hyperthermophiles capable of growing optimally above 80°C.

Despite the rarity of the CRISPR-Cas system in bacterial genomes and in particular the fact that Cas9 has been found only in bacteria (not archaea) with optimal growth temperatures below 45°C, the inventors have surprisingly discovered several thermostable Cas9 variants which enable genome editing to be carried out at elevated temperatures. These Cas9 nucleases provide novel tools for genetic engineering at elevated temperatures and are of particular value in the genetic manipulation of thermophilic organisms; particularly microorganisms.

SUMMARY OF THE INVENTION

Accordingly, the present invention provides an isolated clustered regularly interspaced short palindromic repeat (CRISPR)-associated (Cas) protein or polypeptide comprising;

a. the amino acid motif EKDGKYYC [SEQ ID NO: 2]; and/or
b. the amino acid motif $X_1X_2CTX_3X_4$ [SEQ ID NO: 3] wherein X_1 is independently selected from Isoleucine, Methionine or Proline, X_2 is

independently selected from Valine, Serine, Asparagine or Isoleucine, X_3 is independently selected from Glutamate or Lysine and X_4 is one of Alanine, Glutamate or Arginine; and/or

c. the amino acid motif X_5LKX_6IE [SEQ ID NO: 4] wherein X_5 is independently selected from Methionine or Phenylalanine and X_6 is independently selected from Histidine or Asparagine; and/or

d. the amino acid motif X_7VYSX_8K [SEQ ID NO: 5] wherein X_7 is Glutamate or Isoleucine and X_8 is one of Tryptophan, Serine or Lysine; and/or

e. the amino acid motif $X_9FYX_{10}X_{11}REQX_{12}KEX_{13}$ [SEQ ID NO: 6] wherein X_9 is Alanine or Glutamate, X_{10} is Glutamine or Lysine, X_{11} is Arginine or Alanine, X_{12} is Asparagine or Alanine and X_{13} is Lysine or Serine.

A polypeptide in the context of this invention may be viewed as a fragment of the full length Cas protein. Such fragments may be inactive and used in ways and for purposes not associated directly with editing or cutting of genetic material, for example for standards in assays or raising antibodies or the like.

In preferred embodiments however, the Cas protein or polypeptide is functional and capable of DNA cleavage at a temperature in the range 50°C and 100°C, inclusive, when associated with at least one targeting RNA molecule, and a polynucleotide comprising a target nucleic acid sequence recognised by the targeting RNA molecule.

In particular embodiments, the invention may provide a Cas protein or polypeptide comprising the amino acid motif EKDGKYYC [SEQ ID NO: 2]. In other embodiments, the Cas proteins or polypeptides may further comprise the amino acid motif $X_1X_2CTX_3X_4$ [SEQ ID NO: 3] wherein X_1 is independently selected from Isoleucine, Methionine or Proline, X_2 is independently selected from Valine, Serine, Asparagine or Isoleucine, X_3 is independently selected from Glutamate or Lysine and X_4 is one of Alanine, Glutamate or Arginine.

In other embodiments the Cas proteins or polypeptides defined herein may additionally further comprise the amino acid motif X_5LKX_6IE [SEQ ID NO: 4] wherein

X_5 is independently selected from Methionine or Phenylalanine and X_6 is independently selected from Histidine or Asparagine.

In other embodiments, the Cas proteins or polypeptides defined herein may additionally further comprise the amino acid motif X_7VYSX_8K [SEQ ID NO: 5] wherein X_7 is Glutamate or Isoleucine and X_8 is one of Tryptophan, Serine or Lysine.

In other embodiments, the Cas proteins or polypeptides defined herein may additionally further comprise the amino acid motif $X_9FYX_{10}X_{11}REQX_{12}KEX_{13}$ [SEQ ID NO: 6] wherein X_9 is Alanine or Glutamate, X_{10} is Glutamine or Lysine, X_{11} is Arginine or Alanine, X_{12} is Asparagine or Alanine and X_{13} is Lysine or Serine.

In accordance with the present invention, it may be appreciated that a Cas protein or polypeptide of the invention may comprise any of the motifs of SEQ ID NOs 2 to 6, either alone or in combination. The following summarises each of the combinations of motifs which may characterize Cas proteins or polypeptides of the invention:

EKDGKYYC [SEQ ID NO: 2].

EKDGKYYC [SEQ ID NO: 2]; and $X_1X_2CTX_3X_4$ [SEQ ID NO: 3] wherein X_1 is independently selected from Isoleucine, Methionine or Proline, X_2 is independently selected from Valine, Serine, Asparagine or Isoleucine, X_3 is independently selected from Glutamate or Lysine and X_4 is one of Alanine, Glutamate or Arginine.

EKDGKYYC [SEQ ID NO: 2]; and $X_1X_2CTX_3X_4$ [SEQ ID NO: 3] wherein X_1 is independently selected from Isoleucine, Methionine or Proline, X_2 is independently selected from Valine, Serine, Asparagine or Isoleucine, X_3 is independently selected from Glutamate or Lysine and X_4 is one of Alanine, Glutamate or Arginine; and X_5LKX_6IE [SEQ ID NO: 4] wherein X_5 is independently selected from Methionine or Phenylalanine and X_6 is independently selected from Histidine or Asparagine.

EKDGKYYC [SEQ ID NO: 2]; and $X_1X_2CTX_3X_4$ [SEQ ID NO: 3] wherein X_1 is independently selected from Isoleucine, Methionine or Proline, X_2 is independently selected from Valine, Serine, Asparagine or Isoleucine, X_3 is independently selected

from Glutamate or Lysine and X_4 is one of Alanine, Glutamate or Arginine; and X_5LKX_6IE [SEQ ID NO: 4] wherein X_5 is independently selected from Methionine or Phenylalanine and X_6 is independently selected from Histidine or Asparagine; and X_7VYSX_8K [SEQ ID NO: 5] wherein X_7 is Glutamate or Isoleucine and X_8 is one of Tryptophan, Serine or Lysine.

EKDGKYYC [SEQ ID NO: 2]; and $X_1X_2CTX_3X_4$ [SEQ ID NO: 3] wherein X_1 is independently selected from Isoleucine, Methionine or Proline, X_2 is independently selected from Valine, Serine, Asparagine or Isoleucine, X_3 is independently selected from Glutamate or Lysine and X_4 is one of Alanine, Glutamate or Arginine; and X_5LKX_6IE [SEQ ID NO: 4] wherein X_5 is independently selected from Methionine or Phenylalanine and X_6 is independently selected from Histidine or Asparagine; and X_7VYSX_8K [SEQ ID NO: 5] wherein X_7 is Glutamate or Isoleucine and X_8 is one of Tryptophan, Serine or Lysine; and $X_9FYX_{10}X_{11}REQX_{12}KEX_{13}$ [SEQ ID NO: 6] wherein X_9 is Alanine or Glutamate, X_{10} is Glutamine or Lysine, X_{11} is Arginine or Alanine, X_{12} is Asparagine or Alanine and X_{13} is Lysine or Serine.

EKDGKYYC [SEQ ID NO: 2]; and $X_1X_2CTX_3X_4$ [SEQ ID NO: 3] wherein X_1 is independently selected from Isoleucine, Methionine or Proline, X_2 is independently selected from Valine, Serine, Asparagine or Isoleucine, X_3 is independently selected from Glutamate or Lysine and X_4 is one of Alanine, Glutamate or Arginine; and X_5LKX_6IE [SEQ ID NO: 4] wherein X_5 is independently selected from Methionine or Phenylalanine and X_6 is independently selected from Histidine or Asparagine; and $X_9FYX_{10}X_{11}REQX_{12}KEX_{13}$ [SEQ ID NO: 6] wherein X_9 is Alanine or Glutamate, X_{10} is Glutamine or Lysine, X_{11} is Arginine or Alanine, X_{12} is Asparagine or Alanine and X_{13} is Lysine or Serine.

EKDGKYYC [SEQ ID NO: 2]; and $X_1X_2CTX_3X_4$ [SEQ ID NO: 3] wherein X_1 is independently selected from Isoleucine, Methionine or Proline, X_2 is independently selected from Valine, Serine, Asparagine or Isoleucine, X_3 is independently selected from Glutamate or Lysine and X_4 is one of Alanine, Glutamate or Arginine; and X_7VYSX_8K [SEQ ID NO: 5] wherein X_7 is Glutamate or Isoleucine and X_8 is one of Tryptophan, Serine or Lysine; and $X_9FYX_{10}X_{11}REQX_{12}KEX_{13}$ [SEQ ID NO: 6]

wherein X_9 is Alanine or Glutamate, X_{10} is Glutamine or Lysine, X_{11} is Arginine or Alanine, X_{12} is Asparagine or Alanine and X_{13} is Lysine or Serine.

EKDGKYYC [SEQ ID NO: 2]; and X_5LKX_6IE [SEQ ID NO: 4] wherein X_5 is independently selected from Methionine or Phenylalanine and X_6 is independently selected from Histidine or Asparagine; and X_7VYSX_8K [SEQ ID NO: 5] wherein X_7 is Glutamate or Isoleucine and X_8 is one of Tryptophan, Serine or Lysine; and $X_9FYX_{10}X_{11}REQX_{12}KEX_{13}$ [SEQ ID NO: 6] wherein X_9 is Alanine or Glutamate, X_{10} is Glutamine or Lysine, X_{11} is Arginine or Alanine, X_{12} is Asparagine or Alanine and X_{13} is Lysine or Serine.

EKDGKYYC [SEQ ID NO: 2]; and X_5LKX_6IE [SEQ ID NO: 4] wherein X_5 is independently selected from Methionine or Phenylalanine and X_6 is independently selected from Histidine or Asparagine.

EKDGKYYC [SEQ ID NO: 2]; and X_7VYSX_8K [SEQ ID NO: 5] wherein X_7 is Glutamate or Isoleucine and X_8 is one of Tryptophan, Serine or Lysine.

EKDGKYYC [SEQ ID NO: 2]; and $X_9FYX_{10}X_{11}REQX_{12}KEX_{13}$ [SEQ ID NO: 6] wherein X_9 is Alanine or Glutamate, X_{10} is Glutamine or Lysine, X_{11} is Arginine or Alanine, X_{12} is Asparagine or Alanine and X_{13} is Lysine or Serine.

EKDGKYYC [SEQ ID NO: 2]; and X_5LKX_6IE [SEQ ID NO: 4] wherein X_5 is independently selected from Methionine or Phenylalanine and X_6 is independently selected from Histidine or Asparagine; and X_7VYSX_8K [SEQ ID NO: 5] wherein X_7 is Glutamate or Isoleucine and X_8 is one of Tryptophan, Serine or Lysine.

EKDGKYYC [SEQ ID NO: 2]; and X_5LKX_6IE [SEQ ID NO: 4] wherein X_5 is independently selected from Methionine or Phenylalanine and X_6 is independently selected from Histidine or Asparagine; and $X_9FYX_{10}X_{11}REQX_{12}KEX_{13}$ [SEQ ID NO: 6] wherein X_9 is Alanine or Glutamate, X_{10} is Glutamine or Lysine, X_{11} is Arginine or Alanine, X_{12} is Asparagine or Alanine and X_{13} is Lysine or Serine.

EKDGKYYC [SEQ ID NO: 2]; and X_7VYSX$_8$K [SEQ ID NO: 5] wherein X_7 is Glutamate or Isoleucine and X_8 is one of Tryptophan, Serine or Lysine; and X_9FYX$_{10}$X$_{11}$REQX$_{12}$KEX$_{13}$ [SEQ ID NO: 6] wherein X_9 is Alanine or Glutamate, X_{10} is Glutamine or Lysine, X_{11} is Arginine or Alanine, X_{12} is Asparagine or Alanine and X_{13} is Lysine or Serine.

X_1X_2CTX$_3X_4$ [SEQ ID NO: 3] wherein X_1 is independently selected from Isoleucine, Methionine or Proline, X_2 is independently selected from Valine, Serine, Asparagine or Isoleucine, X_3 is independently selected from Glutamate or Lysine and X_4 is one of Alanine, Glutamate or Arginine; and X_5LKX$_6$IE [SEQ ID NO: 4] wherein X_5 is independently selected from Methionine or Phenylalanine and X_6 is independently selected from Histidine or Asparagine.

X_1X_2CTX$_3X_4$ [SEQ ID NO: 3] wherein X_1 is independently selected from Isoleucine, Methionine or Proline, X_2 is independently selected from Valine, Serine, Asparagine or Isoleucine, X_3 is independently selected from Glutamate or Lysine and X_4 is one of Alanine, Glutamate or Arginine; and X_5LKX$_6$IE [SEQ ID NO: 4] wherein X_5 is independently selected from Methionine or Phenylalanine and X_6 is independently selected from Histidine or Asparagine; and X_7VYSX$_8$K [SEQ ID NO: 5] wherein X_7 is Glutamate or Isoleucine and X_8 is one of Tryptophan, Serine or Lysine.

X_1X_2CTX$_3X_4$ [SEQ ID NO: 3] wherein X_1 is independently selected from Isoleucine, Methionine or Proline, X_2 is independently selected from Valine, Serine, Asparagine or Isoleucine, X_3 is independently selected from Glutamate or Lysine and X_4 is one of Alanine, Glutamate or Arginine; and X_5LKX$_6$IE [SEQ ID NO: 4] wherein X_5 is independently selected from Methionine or Phenylalanine and X_6 is independently selected from Histidine or Asparagine; and X_7VYSX$_8$K [SEQ ID NO: 5] wherein X_7 is Glutamate or Isoleucine and X_8 is one of Tryptophan, Serine or Lysine; and X_9FYX$_{10}$X$_{11}$REQX$_{12}$KEX$_{13}$ [SEQ ID NO: 6] wherein X_9 is Alanine or Glutamate, X_{10} is Glutamine or Lysine, X_{11} is Arginine or Alanine, X_{12} is Asparagine or Alanine and X_{13} is Lysine or Serine.

X_1X_2CTX$_3X_4$ [SEQ ID NO: 3] wherein X_1 is independently selected from Isoleucine, Methionine or Proline, X_2 is independently selected from Valine, Serine, Asparagine

or Isoleucine, X_3 is independently selected from Glutamate or Lysine and X_4 is one of Alanine, Glutamate or Arginine; and X_7VYSX_8K [SEQ ID NO: 5] wherein X_7 is Glutamate or Isoleucine and X_8 is one of Tryptophan, Serine or Lysine; and $X_9FYX_{10}X_{11}REQX_{12}KEX_{13}$ [SEQ ID NO: 6] wherein X_9 is Alanine or Glutamate, X_{10} is Glutamine or Lysine, X_{11} is Arginine or Alanine, X_{12} is Asparagine or Alanine and X_{13} is Lysine or Serine.

$X_1X_2CTX_3X_4$ [SEQ ID NO: 3] wherein X_1 is independently selected from Isoleucine, Methionine or Proline, X_2 is independently selected from Valine, Serine, Asparagine or Isoleucine, X_3 is independently selected from Glutamate or Lysine and X_4 is one of Alanine, Glutamate or Arginine; and X_7VYSX_8K [SEQ ID NO: 5] wherein X_7 is Glutamate or Isoleucine and X_8 is one of Tryptophan, Serine or Lysine.

$X_1X_2CTX_3X_4$ [SEQ ID NO: 3] wherein X_1 is independently selected from Isoleucine, Methionine or Proline, X_2 is independently selected from Valine, Serine, Asparagine or Isoleucine, X_3 is independently selected from Glutamate or Lysine and X_4 is one of Alanine, Glutamate or Arginine; and $X_9FYX_{10}X_{11}REQX_{12}KEX_{13}$ [SEQ ID NO: 6] wherein X_9 is Alanine or Glutamate, X_{10} is Glutamine or Lysine, X_{11} is Arginine or Alanine, X_{12} is Asparagine or Alanine and X_{13} is Lysine or Serine.

X_5LKX_6IE [SEQ ID NO: 4] wherein X_5 is independently selected from Methionine or Phenylalanine and X_6 is independently selected from Histidine or Asparagine; and X_7VYSX_8K [SEQ ID NO: 5] wherein X_7 is Glutamate or Isoleucine and X_8 is one of Tryptophan, Serine or Lysine; and $X_9FYX_{10}X_{11}REQX_{12}KEX_{13}$ [SEQ ID NO: 6] wherein X_9 is Alanine or Glutamate, X_{10} is Glutamine or Lysine, X_{11} is Arginine or Alanine, X_{12} is Asparagine or Alanine and X_{13} is Lysine or Serine.

X_5LKX_6IE [SEQ ID NO: 4] wherein X_5 is independently selected from Methionine or Phenylalanine and X_6 is independently selected from Histidine or Asparagine; and X_7VYSX_8K [SEQ ID NO: 5] wherein X_7 is Glutamate or Isoleucine and X_8 is one of Tryptophan, Serine or Lysine.

X_5LKX_6IE [SEQ ID NO: 4] wherein X_5 is independently selected from Methionine or Phenylalanine and X_6 is independently selected from Histidine or Asparagine; and

$X_9FYX_{10}X_{11}REQX_{12}KEX_{13}$ [SEQ ID NO: 6] wherein X_9 is Alanine or Glutamate, X_{10} is Glutamine or Lysine, X_{11} is Arginine or Alanine, X_{12} is Asparagine or Alanine and X_{13} is Lysine or Serine.

X_7VYSX_8K [SEQ ID NO: 5] wherein X_7 is Glutamate or Isoleucine and X_8 is one of Tryptophan, Serine or Lysine; and $X_9FYX_{10}X_{11}REQX_{12}KEX_{13}$ [SEQ ID NO: 6] wherein X_9 is Alanine or Glutamate, X_{10} is Glutamine or Lysine, X_{11} is Arginine or Alanine, X_{12} is Asparagine or Alanine and X_{13} is Lysine or Serine.

In another aspect, the present invention provides an isolated Cas protein or polypeptide fragment thereof having an amino acid sequence of SEQ ID NO: 1 or a sequence of at least 77% identity therewith, wherein the Cas protein or polypeptide is capable of DNA cleavage at a temperature in the range 50°C and 100°C inclusive.

Preferably, the Cas protein or polypeptide is capable of DNA cleavage at a temperature in the range 60°C to 80°C inclusive. For example, the DNA cleavage occurs at a temperature of 61°C, 62°C, 63°C, 64°C, 65°C, 66°C, 67°C, 68°C, 69°C, 70°C, 71°C, 72°C, 73°C, 74°C, 75°C, 76°C, 77°C, 78°C, 79°C or 80°C. More preferably the Cas protein or polypeptide is capable of DNA cleavage at a temperature in the range 60 to 65°C. In preferred aspects, a Cas protein or polypeptide fragment of the invention may comprises an amino acid sequence of at least 75% identity; preferably at least 85%; more preferably at least 90%; even more preferably at least 95% identity to SEQ ID NO: 1.

More particularly, a Cas protein or polypeptide of the invention may comprise an amino acid sequence with a percentage identity with SEQ ID NO:1 as follows: at least 60%, at least 61%, at least 62%, at least 63%, at least 64%, at least 65%, at least 66%, at least 67%, at least 68%, at least 69%, at least 70%, at least 71%, at least 72%, at least 73%, at least 74%, at least 75%, at least 76%, at least 77%, at least 78%, at least 79%, at least 80%, at least 81%, at least 82%, at least 83%, at least 84%, at least 85%, at least 86%, at least 87%, at least 88%, at least 89%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, at least 99.5% or at least 99.8%.

The percentage amino acid sequence identity with SEQ ID NO: 1 is determinable as a function of the number of identical positions shared by the sequences in a selected comparison window, taking into account the number of gaps, and the length of each gap, which need to be introduced for optimal alignment of the two sequences.

A Cas protein or polypeptide fragment of the invention may be characterised in terms of both the reference sequence SEQ ID NO: 1 and any aforementioned percentage variant thereof as defined by percentage sequence identity, alone or in combination with any of the aforementioned amino acid motifs (i.e. SEQ ID NOS 2 and/or 3 and/or 4 and/or 5 and/or 6) as essential features.

Also, the invention provides nucleic acids encoding any of the aforementioned proteins or polypeptides of the invention. The nucleic acids may be isolated or in the form of expression constructs.

In all aforementioned aspects of the present invention, amino acid residues may be substituted conservatively or non-conservatively. Conservative amino acid substitutions refer to those where amino acid residues are substituted for other amino acid residues with similar chemical properties (e.g., charge or hydrophobicity) and therefore do not alter the functional properties of the resulting polypeptide.

Similarly it will be appreciated by a person of average skill in the art that nucleic acid sequences may be substituted conservatively or non-conservatively without affecting the function of the polypeptide. Conservatively modified nucleic acids are those substituted for nucleic acids which encode identical or functionally identical variants of the amino acid sequences. It will be appreciated by the skilled reader that each codon in a nucleic acid (except AUG and UGG; typically the only codons for methionine or tryptophan, respectively) can be modified to yield a functionally identical molecule. Accordingly, each silent variation (i.e. synonymous codon) of a polynucleotide or polypeptide, which encodes a polypeptide of the present invention, is implicit in each described polypeptide sequence.

Cleavage Temperatures

The temperature range, including optimal temperature range of nuclease activity of the Cas proteins of the present invention is significantly higher than that of known Cas9 proteins. Also, the upper extent of the range in which it retains nuclease activity is much higher than that of known Cas9 proteins. A higher optimal temperature and functional range provides a significant advantage in genetic engineering at high temperatures and therefore, for example, in the editing the genomes of thermophilic organisms, many of which have utility in a range of industrial, agricultural and pharmaceutical processes conducted at elevated temperatures.

Advantageously, Cas proteins or polypeptides of the invention are capable of nucleic acid cleavage from 20°C to 100°C but are particularly useful at elevated temperatures, for example between 41°C and 122°C, preferably between 50°C and 100°C. Cas proteins and polypeptides of the invention are capable of cleaving DNA, RNA and synthetic nucleic acids. Cas proteins or polypeptides of the invention are may also provide operability for nuclease activity and gene editing applications at temperatures in the range 20 to 40 °C, for example.

Preferably, Cas proteins or polypeptides of the invention, when associated with suitable gRNA (guide RNA)which recognizes a target sequence in the polynucleotide molecule(s) to be cleaved, marked or modified, does so at temperatures in the range 50°C to 100°C, optionally in the range 55°C to 100°C, 60°C to 100°C, 65°C to 100°C, 70°C to 100°C, 75°C to 100°C, 80°C to 100°C, 85°C to 100°C, 90°C to 100°C, 95°C to 100°C. More preferably, Cas proteins of the invention cleave, mark or modify nucleic acids at temperatures in the range 51°C to 99°C, 52°C to 98°C, 53°C to 97°C, 54°C to 96°C, 55°C to 95°C, 56°C to 94°C, 57°C to 93°C, 58°C to 92°C, 59°C to 91°C, 60°C to 90°C, 61°C to 89°C, 62°C to 88°C, 63°C to 87°C, 64°C to 86°C, 65°C to 85°C, 66°C to 84°C, 67°C to 83°C, 68°C to 82°C, 69°C to 81°C, 70°C to 80°C, 71°C to 79°C, 72°C to 78°C, 73°C to 77°C, 74°C to 76°C, 75°C. Preferably, Cas proteins of the invention cleave, mark or modify nucleic acids at temperatures in the range 60°C to 80°C, 61°C to 79°C, 62°C to 78°C, 63°C to 77°C, 64°C to 76°C, 60°C to 75°C, 60°C to 70°C. Optimally Cas proteins of the invention cleave, mark or modify nucleic acids at temperatures in the range 60°C to 65°C, preferably at 65°C.

In all aspects of the invention, Cas proteins or polypeptides may be obtained or derived from bacteria, archaea or viruses; or alternatively may be synthesised *de novo*. In preferred embodiments, a Cas protein or polypeptide of the invention is derived from a thermophilic prokaryotic organism, which may be classified as an archaea or bacterium, but is preferably a bacterium. More preferably a Cas protein or polypeptide of the invention will be derived from a thermophilic bacterium. Herein, the term thermophilic is to be understood as meaning capable of survival and growth at relatively high temperatures, for example in the context of the invention, capable of nucleic acid cleavage between 41 and 122 °C (106 and 252 °F). Preferably a Cas protein or polypeptide of the invention may be isolated from one or more thermophilic bacteria and will function above 60°C. Preferably a Cas protein or polypeptide of the invention may be isolated from one or more thermophilic bacteria and will function in the range 60°C to 80°C and optimally between 60°C and 65°C. In preferred embodiments, a Cas protein or polypeptide of the invention is derived from *Geobacillus* sp. More preferably, a Cas protein of the invention is derived from *Geobacillus thermodenitrificans*. Even more preferably, a Cas protein of the invention is derived from *Geobacillus thermodenitrificans* T12. A Cas protein or polypeptide of the invention may be derived from a virus.

RNA Guides and Target Sequences

Cas proteins of the invention allow for sequence-specific cleavage, tagging, marking or modification of target nucleic acids at elevated temperatures. Target nucleic acids may be DNA (single-stranded or double-stranded), RNA or synthetic nucleic acids. A particularly useful application of the present invention is the sequence-specific targeting and modification of genomic DNA by one or more Cas proteins of the invention in complex with one or more guide RNAs (gRNAs) that complementarily bind to a targeted sequence of the genomic DNA. Consequently, the target nucleic acid is preferably double-stranded DNA. Such targeting may be performed *in vitro* or *in vivo*. Preferably such targeting is performed *in vivo*. In this way, Cas proteins of the invention may be used to target and modify specific DNA sequences located in the genomic DNA of a cell. It is envisaged that the Cas system may be used to modify genomes in a variety of cell types of and/or in different organisms.

Accordingly, the present invention provides a ribonucleoprotein complex comprising a Cas protein or a polypeptide of the invention as hereinbefore described, and further comprising at least one RNA molecule which has a targeting function in that it recognizes a particular nucleotide sequence in a target polynucleotide. Preferably, the RNA molecule is a single-stranded RNA molecule, e.g. a CRISPR RNA (crRNA) and is associated, e.g. by hybridization with a tracrRNA. The targeting RNA may be a chimera of a crRNA and tracrRNA. The aforementioned RNA molecules may have a ribonucleotide sequence of at least 90% identity, or complementarity to a target nucleotide sequence. Optionally, the RNA molecule has a ribonucleotide sequence of at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99% or 100% identity or complementarity to a target nucleotide sequence. The preferred target nucleotide sequence is a DNA.

In a preferred aspect, the present invention provides a ribonucleoprotein complex as hereinbefore described, wherein the at least one targeting RNA molecule is substantially complementary along its length to a target DNA sequence.

Alteration of the sequence of the RNA guide which associates with the Cas protein of the invention therefore allows the Cas protein to be programmed to mark or cut double-stranded DNA at sites complementary to the guide RNA.

Preferably, the length of the at least one targeting RNA molecule in a ribonucleoprotein complex of the invention is in the range 35 to 135 residues, optionally in the range 35 to 134 residues, 35 to 133 residues, 35 to 132 residues, 35 to 131 residues, 35 to 130 residues, 35 to 129 residues, 35 to 128 residues, 35 to 127 residues, 35 to 126 residues, 35 to 125 residues, 35 to 124 residues, 35 to 123 residues, 35 to 122 residues, 35 to 121 residues, 35 to 120 residues, 35 to 119 residues, 35 to 118 residues, 35 to 117 residues, 35 to 116 residues, 35 to 115 residues, 35 to 114 residues, 35 to 113 residues, 35 to 112 residues, 35 to 111 residues, 35 to 100 residues, 35 to 109 residues, 35 to 108 residues, 35 to 107 residues, 35 to 106 residues, 35 to 105 residues, 35 to 104 residues, 35 to 103 residues, 35 to 102 residues, 35 to 101 residues, 35 to 100 residues, 35 to 99

residues, 35 to 98 residues, 35 to 97 residues, 35 to 96 residues, 35 to 95 residues, 35 to 94 residues, 35 to 93 residues, 35 to 92 residues, 35 to 91 residues, 35 to 90 residues, 35 to 89 residues, 35 to 88 residues, 35 to 87 residues, 35 to 86 residues, 35 to 85 residues, 35 to 84 residues, 35 to 83 residues, 35 to 82 residues, 35 to 81 residues, 35 to 80 residues, 35 to 79 residues, 35 to 78 residues, 35 to 77 residues, 35 to 76 residues, 35 to 75 residues, 35 to 74 residues, 35 to 73 residues, 35 to 72 residues, 35 to 71 residues, 35 to 70 residues, 35 to 69 residues, 35 to 68 residues, 35 to 67 residues, 35 to 66 residues, 35 to 65 residues, 35 to 64 residues, 35 to 63 residues, 35 to 62 residues, 35 to 61 residues, 35 to 60 residues, 35 to 59 residues, 35 to 58 residues, 35 to 57 residues, 35 to 56 residues, 35 to 55 residues, 35 to 54 residues, 35 to 53 residues, 35 to 52 residues, 35 to 51 residues, 35 to 50 residues, 35 to 49 residues, 35 to 48 residues, 35 to 47 residues, 35 to 46 residues, 35 to 45 residues, 35 to 44 residues, 35 to 43 residues, 35 to 42 residues, 35 to 41 residues, 35 to 40 residues, 35 to 39 residues, 35 to 38 residues, 35 to 37 residues, 35 to 36 residues or 35 residues. Preferably, the length of the at least one RNA molecule is in the range 36 to 174 residues, 37 to 173 residues, 38 to 172 residues, 39 to 171 residues, 40 to 170 residues, 41 to 169 residues, 42 to 168 residues, 43 to 167 residues, 44 to 166 residues, 45 to 165 residues, 46 to 164 residues, 47 to 163 residues, 48 to 162 residues, 49 to 161 residues, 50 to 160 residues, 51 to 159 residues, 52 to 158 residues, 53 to 157 residues, 54 to 156 residues, 36 to 74 residues, 37 to 73 residues, 38 to 72 residues, 39 to 71 residues, 40 to 70 residues, 41 to 69 residues, 42 to 68 residues, 43 to 67 residues, 44 to 66 residues, 45 to 65 residues, 46 to 64 residues, 47 to 63 residues, 48 to 62 residues, 49 to 61 residues, 50 to 60 residues, 51 to 59 residues, 52 to 58 residues, 53 to 57 residues, 54 to 56 residues.

In preferred aspects, the present invention provides a ribonucleoprotein complex, wherein the complementary portion of the at least one RNA molecule is at least 30 residues long. Alternatively, the complementary portion of the at least one RNA molecule may be 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74 or 75 residues long.

The targeting RNA molecule will preferably require a high specificity and affinity for the target nucleic acid sequence. A dissociation constant (K_d) in the range 1 μM to 1 pM, preferably 1 nM to 1pM; more preferably 1 – 100 pM is desirable as may be determined by native gel electrophoresis, or alternatively isothermal titration calorimetry, surface plasmon resonance, or fluorescence based titration methods. Affinity may be determined using an electrophoretic mobility shift assay (EMSA), also called gel retardation assay (see Semenova E et al. (2011) PNAS 108: 10098–10103).

The targeting RNA molecule is preferably modeled on what are known from nature in prokaryotes as CRISPR RNA (crRNA) molecules. The structure of crRNA molecules is already established and explained in more detail in Jore *et al.*, 2011, Nature Structural & Molecular Biology 18: 529–537. In brief, a mature crRNA of type I-E is often 61 nucleotides long and consists of a 5' "handle" region of 8 nucleotides, the "spacer" sequence of 32 nucleotides, and a 3' sequence of 21 nucleotides which form a hairpin with a tetranucleotide loop (Fig 5). Type I systems differ from type II (Cas9) and details of different systems are described in Van der Oost 2014 Nat Rev Micr 12: 479–492. In type II (Cas9) systems there is a different processing mechanism, making use of a second RNA (tracrRNA) and two ribonucleases. Rather than a hairpin, the mature crRNA in type II remains attached to a fragment of the tracrRNA (Fig. 5). However, the RNA used in the invention does not have to be designed strictly to the design of naturally occurring crRNA, whether in length, regions or specific RNA sequences. What is clear though, is that RNA molecules for use in the invention may be designed based on gene sequence information in the public databases or newly discovered, and then made artificially, e.g. by chemical synthesis in whole or in part. The RNA molecules of the invention may also be designed and produced by way of expression in genetically modified cells or cell free expression systems and this option may include synthesis of some or all of the RNA sequence.

The structure and requirements of crRNA in type II (Cas9) has also been described in Jinek *et al.*, 2012 *ibid*. In type I, there is a so-called "SEED" portion forming the 5' end of the spacer sequence and which is flanked 5' thereto by the 5' handle of 8 nucleotides. Semenova *et al.* (2011, PNAS 108: 10098-10103), have found that all

residues of the SEED sequence should be complementary to the target sequence, although for the residue at position 6, a mismatch may be tolerated (Fig. 5). In type II, there is a SEED of 10-12 nucleotides that is located at the 3' end of the spacer (Fig. 5) (reviewed by Van der Oost 2014 *ibid.*). Similarly, when designing and making an RNA component of a ribonucleoprotein complex of the invention directed at a target locus (i.e. sequence), the necessary match and mismatch rules for the type II SEED sequence can be applied.

The invention therefore includes a method of detecting and/or locating a single base change in a target nucleic acid molecule comprising contacting a nucleic acid sample with a ribonucleoprotein complex of the invention as hereinbefore described, or with a Cas protein or polypeptide and separate targeting RNA component of the invention as hereinbefore described, and wherein the sequence of the targeting RNA (including when in the ribonucleoprotein complex) is such that it discriminates between a normal allele and a mutant allele by virtue of a single base change at, for example, position 6 of a contiguous sequence of 8 nucleotide residues.

Without wishing to be bound by a particular theory, a design rule which may be used in preparing a targeting RNA component of ribonucleoprotein complexes of the invention involves the so-called "PAM" (protospacer adjacent motif) sequence in the target polynucleotide strand. The PAM sequence may be a conserved triplet of nucleotide residues, in the type I-E system of *E.coli* 5'-CTT-3', 5'-CAT-3', 5'-CCT-3', 5'-CAC-3', 5'-TTT-3', 5'-ATT-3', and 5'-AWG-3', wherein W is A, T or U. In Type I, a PAM sequence located in the targeted strand is usually at a position corresponding to 5' of the SEED. In Type II, however, the PAM is located at the other end, on the displaced strand close to the 3' end of the crRNA spacer, at a position corresponding to 3' of the seed (Fig. 5) (Jinek *et al.*, 2012, *op. cit.*). For *Streptococcus pyogenes* Cas9, the PAM sequence has a conserved pair of nucleotide residues, 5'-NGG-3'. Recently, different Cas9 variants (Type IIA and Type IIC) (Ran *et al.*, 2015 Nature 520:186-191) – Fig 1A) have been characterized, and PAMs have been revealed (see Ran *et al.*, 2015, *ibid.* – Fig 1C). Currently established Cas9 PAMs include: Type IIA 5'-NGGNNNN-3' (*Streptococcus pyogenes*), 5'-NNGTNNN-3' (*Streptococcus pasteurianus*), 5'-NNGGAAN-3' (*Streptococcus thermophilus*), 5'-NNGGGNN-3' (*Staphylococcus aureus*), and Type IIC 5'-NGGNNNN-3'

(*Corynebacterium difteriae*), 5'-NNGGGTN-3' (*Campylobacter lari*), 5'-NNNCATN-3' (*Parvobaculum lavamentivorans*), 5'-NNNNGTA-3' (*Neiseria cinerea*). Cas9 of *Geobacillus thermodenitrificans* T12 (this invention) belongs to Type IIC (Ran *et al.*, 2015, *ibid.*). Therefore, although without wishing to be bound by a particular theory, the inventors expect that the PAM sequences more closely approximate those established for Type IIC Cas9 ribonucleoprotein complexes.

In embodiments of the invention, a targeting RNA molecule may have a length in the range of 35 – 75 residues. In preferred embodiments, the portion of the RNA which is complementary to and used for targeting a desired nucleic acid sequence is 31 or 32 residues long. In the context of a naturally-occurring crRNA, this would correspond to the spacer portion as shown for example in figure 1 of Semenova *et al.* (2011 *ibid.*).

A ribonucleoprotein complex of the invention may have a targeting component comprising 8 residues derived from the CRISPR repeat 5' to the RNA sequence which has substantial complementarity to the DNA target sequence. The RNA sequence having complementarity to the DNA target sequence would be understood to correspond in the context of a crRNA as being the spacer sequence. The 5' flanking sequence of the RNA would be considered to correspond to the 5' handle of a crRNA; as shown for example in figure 1 of Semenova *et al.* (2011 *ibid.*).

A ribonucleoprotein complex of the invention may have a hairpin and tetranucleotide loop forming sequence 3' to the targeting RNA sequence which has complementarity to a DNA target sequence, i.e. 3' to what would correspond to the 3' handle flanking the spacer sequence in a crRNA; for example as shown in figure 1 of Semenova *et al.* (2011 *ibid.*).

Without wishing to be bound by a particular theory, in a preferred ribonucleoprotein complex, the target sequence DNA strand which does not pair with the targeting RNA of the ribonucleoprotein complex may comprise a directly 3' adjacent triplet selected from 5'-NGGNNNN-3', 5'-NNGTNNN-3', 5'-NNGGAAN-3', 5'-NNGGGNN-3', 5'-NGGNNNN-3', 5'-NNGGGTN-3', 5'-NNNCATN-3', 5'-NNNNGTA-3' and wherein said triplet residues do not each pair with respective complementary DNA

strand residues. However, it will be appreciated that other triplets or combinations of nucleotides may be used depending on the desired application. These triplets correspond to what is termed "protospacer adjacent motif" or "PAM" in the context of naturally occurring crRNAs. In type IIC CRISPR/Cas systems these PAM triplets are required for stable interaction with the Cascade/crRNA complex with its dsDNA target, in order to ensure high degree of specificity of the crRNA – both in the natural system targets and therefore preferably also of the RNAs according to the present invention – for the target sequence.

Functional Moieties

Advantageously, the ability of Cas proteins, polypeptides and riboucleoprotein complexes of the invention to target any polynucleotide sequence in a sequence-specific manner may be exploited in order to modify the target nucleic acid in some way, for example by cleaving it and/or marking it and/or modifying it. It will therefore be appreciated that additional proteins may be provided along with the Cas protein or polypeptide to achieve this. Accordingly, Cas proteins, polypeptides or ribonucleoprotein complexes of the present invention may be provided as part of a protein complex comprising at least one further protein. In a preferred aspect the present invention provides a Cas protein, polypeptide or a ribonucleoprotein complex wherein the Cas protein or at least one further protein further comprises at least one functional moiety. The at least one functional moiety may be fused or linked to the Cas protein. Preferably, the at least one functional moiety may be translationally fused to the Cas protein through expression in natural or artificial protein expression systems. Alternatively, the at least one functional moiety may be covalently linked by a chemical synthesis step to the Cas protein. Preferably, the at least one functional moiety is fused or linked to the N-terminus and/or the C-terminus of the Cas protein; preferably the N-terminus.

Desirably, the at least one functional moiety will be a protein. It may be a heterologous protein or alternatively may be native to the bacterial species from which the Cas protein was derived. The at least one functional moiety may be a protein; optionally selected from a helicase, a nuclease, a helicase-nuclease, a DNA methylase, a histone methylase, an acetylase, a phosphatase, a kinase, a

transcription (co-)activator, a transcription repressor, a DNA binding protein, a DNA structuring protein, a marker protein, a reporter protein, a fluorescent protein, a ligand binding protein, a signal peptide, a subcellular localisation sequence, an antibody epitope or an affinity purification tag.

In a particularly preferred aspect, the present invention provides a Cas protein, polypeptide, or a ribonucleoprotein complex, wherein the at least one functional moiety is a marker protein, for example GFP.

Nuclease Activity

A Cas ribonucleoprotein of any aspect of the invention has nucleic acid cleavage activity between 50°C and 100°C. The ribonucleoproteins of the invention are capable of cleaving DNA, RNA or synthetic nucleic acids. In preferred aspects Cas ribonucleoproteins of the invention are capable of cleaving DNA in a sequence-specific manner, in particular double-stranded DNA.

In any aspect of the invention, Cas proteins, polypeptides or ribonucleoproteins of the invention may have more than one nuclease domain. Site-specific nucleases can permit the generation of double strand breaks (DSBs) at selected positions along a strand of DNA. In a target host cell, this enables DSBs to be made at specific pre-selected positions in the genome. The creation of such breaks by site-specific nucleases prompts the endogenous cellular repair machinery to be repurposed in order to insert, delete or modify DNA at desired positions in the genome of interest.

One or more nuclease activity sites of the protein or polypeptide molecule may be inactivated, e.g. so as to allow the activity of another functional moiety linked or fused to the protein or polypeptide, e.g. Fok1 nuclease.

Therefore notwithstanding the fact that the Cas proteins, polypeptides and ribonucleoproteins of the invention have endogenous nuclease activity, for certain applications it may be desirable to inactivate the native nuclease activity of the Cas protein and provide a Cas protein or a ribonucleoprotein complex wherein the native

Cas9 nuclease activity is inactivated and the Cas protein is linked to at least one functional moiety. Reducing the incidence of mis-targeting events by complementation of the native Cas9 nuclease activity is one such application. This may desirably be achieved by inactivation of the native Cas9 nuclease activity of the Cas protein or ribonucleoprotein complex and provision of a heterologous nuclease, preferably fused to the Cas protein. Accordingly, the present invention provides a Cas protein or a ribonucleoprotein complex, wherein the at least one functional moiety is a nuclease domain, preferably a FokI nuclease domain. In a particularly preferred aspect, the Cas protein or ribonucleoprotein complex of the invention fused to a FokI nuclease domain is provided as part of a protein complex, preferably comprising another Cas protein or ribonucleoprotein complex of the invention fused to a FokI nuclease domain and wherein the two complexes target opposite strands of the target genomic DNA.

For some applications it may be desirable to completely attenuate the nuclease activity of the Cas protein, polypeptide or ribonucleoprotein, for example in applications where the Cas protein or ribonucleoprotein complex is utilised to recognise and modify a specific target sequence in a nucleic acid, for instance to mark it as part of a diagnostic test. In such applications, the nuclease activity of the Cas protein may be inactivated and the functional moiety fused to the Cas protein may be a protein; optionally selected from a helicase, a nuclease, a helicase-nuclease, a DNA methylase, a histone methylase, an acetylase, a phosphatase, a kinase, a transcription (co-)activator, a transcription repressor, a DNA binding protein, a DNA structuring protein, a marker protein, a reporter protein, a fluorescent protein, a ligand binding protein, a signal peptide, a subcellular localisation sequence, an antibody epitope or an affinity purification tag.

In a particularly preferred aspect, the present invention provides a Cas protein or a ribonucleoprotein complex, wherein the nuclease activity of the Cas protein is inactivated and the at least one functional moiety is a marker protein, for example GFP. In this way it may be possible to specifically target a nucleic acid sequence of interest and to visualize it using a marker which generates an optical signal. Suitable markers may include for example, a fluorescent reporter protein, e.g. Green Fluorescent Protein (GFP), Yellow Fluorescent Protein (YFP), Red Fluorescent

Protein (RFP), Cyan Fluorescent Protein (CFP) or mCherry. Such a fluorescent reporter gene provides a suitable marker for visualisation of protein expression since its expression can be simply and directly assayed by fluorescence measurement. Alternatively, the reporter nucleic acid may encode a luminescent protein, such as a luciferase (e.g. firefly luciferase). Alternatively, the reporter gene may be a chromogenic enzyme which can be used to generate an optical signal, e.g. a chromogenic enzyme (such as beta-galactosidase (LacZ) or beta-glucuronidase (Gus)). Reporters used for measurement of expression may also be antigen peptide tags. Other reporters or markers are known in the art, and they may be used as appropriate.

Because the marker may be visualized, in certain embodiments where the target nucleic acid is RNA, specifically mRNA, it may be possible to quantify the transcriptional activity of a gene by detection and quantification of the optical signal provided by the marker, particularly where the optical signal generated by the marker is directly proportionate to the quantity of the expression product. Therefore in preferred embodiments of the invention, Cas proteins or ribonucleoproteins of the invention may be used to assay expression products of a gene of interest.

Throughout, the reference sequences of the Cas proteins of the invention may be defined as a nucleotide sequence encoding the amino acid sequence. For example the amino acid sequence of the motifs defined in SEQ ID's 2 to 6 also includes all nucleic acid sequences which encode that amino acid sequence.

Accordingly, the present invention also provides an isolated nucleic acid molecule encoding a Cas protein comprising;

a. the amino acid motif EKDGKYYC [SEQ ID NO: 2]; and/or
b. the amino acid motif $X_1X_2CTX_3X_4$ [SEQ ID NO: 3] wherein X_1 is independently selected from Isoleucine, Methionine or Proline, X_2 is independently selected from Valine, Serine, Asparagine or Isoleucine, X_3 is independently selected from Glutamate or Lysine and X_4 is one of Alanine, Glutamate or Arginine; and/or

c. the amino acid motif X_5LKX_6IE [SEQ ID NO: 4] wherein X_5 is independently selected from Methionine or Phenylalanine and X_6 is independently selected from Histidine or Asparagine; and/or
d. the amino acid motif X_7VYSX_8K [SEQ ID NO: 5] wherein X_7 is Glutamate or Isoleucine and X_8 is one of Tryptophan, Serine or Lysine; and/or
e. the amino acid motif $X_9FYX_{10}X_{11}REQX_{12}KEX_{13}$ [SEQ ID NO: 6] wherein X_9 is Alanine or Glutamate, X_{10} is Glutamine or Lysine, X_{11} is Arginine or Alanine, X_{12} is Asparagine or Alanine and X_{13} is Lysine or Serine;

wherein the Cas protein is capable of DNA cleavage between 50°C and 100°C when associated with at least one targeting RNA molecule, and a polynucleotide comprising a target nucleic acid sequence recognised by the targeting RNA molecule.

In another aspect the present invention also provides an isolated nucleic acid molecule encoding a clustered regularly interspaced short palindromic repeat (CRISPR)-associated (Cas) protein having an amino acid sequence of SEQ ID NO: 1 or a sequence of at least 77% identity therewith.

In another aspect the present invention also provides an isolated nucleic acid molecule, further comprising at least one nucleic acid sequence encoding a peptide which upon translation is fused to the Cas protein.

In another aspect the present invention also provides an isolated nucleic acid molecule, wherein the at least one nucleic acid sequence fused to the nucleic acid molecule encoding the Cas protein encodes a protein selected from a helicase, a nuclease, a helicase-nuclease, a DNA methylase, a histone methylase, an acetylase, a phosphatase, a kinase, a transcription (co-)-activator, a transcription repressor, a DNA binding protein, a DNA structuring protein, a marker protein, a reporter protein, a fluorescent protein, a ligand binding protein, a signal peptide, a subcellular localisation sequence, an antibody epitope or an affinity purification tag.

Expression Vectors

Nucleic acids of the present invention may be isolated. However, in order that expression of the nucleic acid sensing construct may be carried out in a chosen cell, the polynucleotide sequence encoding the Cas protein or ribonucleoprotein will preferably be provided in an expression construct. In some embodiments, the polynucleotide encoding the Cas protein or ribonucleoprotein will be provided as part of a suitable expression vector. In certain embodiments an expression vector of the present invention (with or without nucleotide sequence encoding amino acid residues which on expression will be fused to a Cas protein) may further comprise a nucleotide sequence encoding a targeting RNA molecule as hereinbefore defined. Consequently, such expression vectors can be used in an appropriate host to generate a ribonucleoprotein complex of the invention which can target a desired nucleotide sequence. Alternatively, nucleotide sequences encoding a targeting RNA molecule as hereinbefore defined may be provided in a separate expression vector or alternatively may be delivered to a target cell by other means.

Suitable expression vectors will vary according to the recipient cell and suitably may incorporate regulatory elements which enable expression in the target cell and preferably which facilitate high-levels of expression. Such regulatory sequences may be capable of influencing transcription or translation of a gene or gene product, for example in terms of initiation, accuracy, rate, stability, downstream processing and mobility.

Such elements may include, for example, strong and/or constitutive promoters, 5' and 3' UTR's, transcriptional and/or translational enhancers, transcription factor or protein binding sequences, start sites and termination sequences, ribosome binding sites, recombination sites, polyadenylation sequences, sense or antisense sequences, sequences ensuring correct initiation of transcription and optionally poly-A signals ensuring termination of transcription and transcript stabilisation in the host cell. The regulatory sequences may be plant-, animal-. bacteria-, fungal- or virus derived, and preferably may be derived from the same organism as the host cell. Clearly, appropriate regulatory elements will vary according to the host cell of interest. For example, regulatory elements which facilitate high-level expression in prokaryotic host cells such as in *E. coli* may include the pLac, T7, P(Bla), P(Cat), P(Kat), trp or tac promoters. Regulatory elements which facilitate high-level

expression in eukaryotic host cells might include the AOX1 or GAL1 promoter in yeast or the CMV- or SV40-promoters, CMV-enhancer, SV40-enhancer, *Herpes simplex virus* VIP16 transcriptional activator or inclusion of a globin intron in animal cells. In plants, constitutive high-level expression may be obtained using, for example, the *Zea mays* ubiquitin 1 promoter or 35S and 19S promoters of cauliflower mosaic virus.

Suitable regulatory elements may be constitutive, whereby they direct expression under most environmental conditions or developmental stages, developmental stage specific or inducible. Preferably, the promoter is inducible, to direct expression in response to environmental, chemical or developmental cues, such as temperature, light, chemicals, drought, and other stimuli. Suitably, promoters may be chosen which allow expression of the protein of interest at particular developmental stages or in response to extra- or intra-cellular conditions, signals or externally applied stimuli. For example, a range of promoters exist for use in *E. coli* which give high-level expression at particular stages of growth (e.g. osmY stationary phase promoter) or in response to particular stimuli (e.g. HtpG Heat Shock Promoter).

Suitable expression vectors may comprise additional sequences encoding selectable markers which allow for the selection of said vector in a suitable host cell and/or under particular conditions.

The invention also includes a method of modifying a target nucleic acid in a cell, comprising transfecting, transforming or transducing the cell with any of the expression vectors as hereinbefore described. The methods of transfection, transformation or transduction are of the types well known to a person of skill in the art. Where there is one expression vector used to generate expression of a ribonucleoprotein complex of the invention and when the targeting RNA is added directly to the cell then the same or a different method of transfection, transformation or transduction may be used. Similarly, then there is one expression vector being used to generate expression of a ribonucleoprotein complex of the invention and when another expression vector is being used to generate the targeting RNA *in situ* via expression, then the same or a different method of transfection, transformation or transduction may be used.

In other embodiments, mRNA encoding the Cas protein or polypeptide is introduced into a cell so that the Cascade complex is expressed in the cell. The targeting RNA which guides the Cas protein complex to the desired target sequence is also introduced into the cell, whether simultaneously, separately or sequentially from the mRNA, such that the necessary ribonucleoprotein complex is formed in the cell.

Accordingly, the invention also provides a method of modifying, i.e. cleaving, tagging, marking or binding, a target nucleic acid comprising contacting the nucleic acid with a ribonucleoprotein complex as hereinbefore defined.

In addition, the invention also includes a method of modifying a target nucleic acid comprising contacting the nucleic acid with a Cas protein or polypeptide as hereinbefore defined, in addition to a targeting RNA molecule as hereinbefore defined.

In accordance with the above methods, modification of target nucleic acid may therefore be carried out *in vitro* and in a cell-free environment. In a cell-free environment, addition of each of the target nucleic acid, the Cas protein and the targeting RNA molecule may be simultaneous, sequential (in any order as desired), or separately. Thus it is possible for the target nucleic acid and targeting RNA to be added simultaneously to a reaction mix and then the Cas protein or polypeptide of the invention to be added separately at a later stage.

Equally, the modification of the target nucleic acid may be made *in vivo*, that is *in situ* in a cell, whether an isolated cell or as part of a multicellular tissue, organ or organism. In the context of whole tissue and organs, and in the context of an organism, the method may desirably be carried out *in vivo* or alternatively may be carried out by isolating a cell from the whole tissue, organ or organism, treating the cell ribonucleoprotein complex in accordance with the method and subsequently returning the cell treated with ribonucleoprotein complex to its former location, or a different location, whether within the same or a different organism.

In these embodiments, the ribonucleoprotein complex or the Cas protein or polypeptide requires an appropriate form of delivery into the cell. Such suitable delivery systems and methods are well known to persons skilled in the art, and include but are not limited to cytoplasmic or nuclear microinjection. In preferred modes of delivery, an Adeno-associated virus (AAV) is used; this delivery system is not disease causing in humans and has been approved for clinical use in Europe.

Accordingly the present invention provides a method of modifying a target nucleic acid comprising contacting the nucleic acid with:

a. a ribonucleoprotein complex as hereinbefore defined; or

b. a protein or protein complex as hereinbefore defined and an RNA molecule as as hereinbefore defined.

In a further aspect the present invention provides a method of modifying a target nucleic acid in a cell, comprising transforming, transfecting or transducing the cell with an expression vector comprising nucleotide sequences encoding a ribonucleoprotein complex as hereinbefore defined; or alternatively transforming, transfecting or transducing the cell with an expression vector comprising nucleotide sequences encoding a protein or protein complex as hereinbefore defined and a further expression vector comprising a nucleotide sequence encoding a targeting RNA molecule as hereinbefore defined.

In a further aspect, the present invention provides a method of modifying a target nucleic acid in a cell comprising transforming, transfecting or transducing the cell with an expression vector comprising nucleotide sequences encoding a protein or protein complex as hereinbefore defined, and then delivering a targeting RNA molecule as hereinbefore defined into the cell.

In embodiments where the guide (i.e. targeting) RNA (gRNA) molecule and the Cas protein or polypeptide are provided separately rather than as part of a ribonucleoprotein complex, the gRNA molecule requires an appropriate form of delivery into a cell, whether simultaneously, separately or sequentially with the Cas protein or protein complex. Such forms of introducing RNA into cells are well known to a person of skill in the art and may include in vitro or ex vivo delivery via

conventional transfection methods. Physical methods, such as microinjection and electroporation, as well as calcium co-precipitation, and commercially available cationic polymers and lipids, and cell-penetrating peptides, cell-penetrating (biolistic) particles may each be used. For example, viruses, particularly preferred is AAV, may be used as delivery vehicles, whether to the cytoplasm and/or nucleus, for example via the (reversible) fusion of Cas protein complex of the invention or a ribonucleoprotein complex of the invention to the viral particle.

In another aspect the present invention provides a method of modifying a target nucleic acid, wherein the at least one functional moiety is a marker protein or reporter protein and the marker protein or reporter protein associates with the target nucleic acid; preferably wherein the marker is a fluorescent protein, for example a green fluorescent protein (GFP).

In the aforementioned methods of modifying a target nucleic acid, the functional moiety may be a marker and the marker associates with the target nucleic acid; preferably wherein the marker is a protein; optionally a fluorescent protein, e.g. green fluorescent protein (GFP), yellow fluorescent protein (YFP), red fluorescent protein (RFP) or mCherry. Whether *in vitro*, *ex vivo* or *in vivo*, then methods of the invention can be used to directly visualise a target locus in a nucleic acid molecule, preferably in the form of a higher order structure such as a supercoiled plasmid or chromosome, or a single stranded target nucleic acid such as mRNA. Direct visualisation of a target locus may use electron micrography, or fluorescence microscopy. However, it will be appreciated that in the context of methods of the invention, other kinds of label may be used as the marker including organic dye molecules, radiolabels and spin labels which may be small molecules.

In methods of the invention for modifying a target nucleic acid wherein the target nucleic acid is dsDNA, the functional moiety may be a nuclease or a helicase-nuclease, and the modification is preferably a single stranded or a double stranded break at a desired locus. In this way unique sequence specific cutting of DNA can be engineered by using a suitable functional moiety fused to a ribonucleoprotein complex. The chosen sequence of the RNA component of the final

ribonucleoprotein complex provides the desired sequence specificity for the action of the functional moiety.

Therefore, the invention also provides a method of non-homologous end joining of a dsDNA molecule in a cell at a desired locus to remove at least a part of a nucleotide sequence from the dsDNA molecule; optionally to knockout the function of a gene or genes, wherein the method comprises making double stranded breaks using any of the methods of modifying a target nucleic acid as hereinbefore described.

The invention further provides a method of homologous recombination of a nucleic acid into a dsDNA molecule in a cell at a desired locus in order to modify an existing nucleotide sequence or insert a desired nucleotide sequence, wherein the method comprises making a double stranded break at the desired locus using any of the methods of modifying a target nucleic acid as hereinbefore described.

The invention therefore also provides a method of modifying gene expression in an organism comprising modifying a target nucleic acid sequence according to any of the methods hereinbefore described, wherein the nucleic acid is dsDNA and the functional moiety is selected from a DNA modifying enzyme (e.g. a methylase or acetylase), a transcription activator or a transcription repressor.

The invention additionally provides a method of modifying gene expression in an organism comprising modifying a target nucleic acid sequence according to any of the methods hereinbefore described, wherein the nucleic acid is an mRNA and the functional moiety is a ribonuclease; optionally selected from an endonuclease, a 3' exonuclease or a 5' exonuclease.

In any aspect of methods of the invention described herein, the target nucleic acid may be DNA, RNA or synthetic nucleic acid. Preferably the target nucleic acid is DNA; preferably dsDNA.

However, the target nucleic acid can be RNA; preferably mRNA. Alternatively therefore, the present invention also provides methods of modifying a target nucleic acid, wherein the target nucleic acid is RNA.

In another aspect the present invention provides a method of modifying a target nucleic acid, wherein the nucleic acid is dsDNA, the at least one functional moiety is a nuclease or a helicase-nuclease, and the modification is a single- stranded or a double- stranded break at a desired locus.

In another aspect the present invention provides a method of modifying a target nucleic acid in a cell, wherein modification results in a silencing of gene expression at a desired locus; and wherein the method includes the steps of;

a. making double-stranded breaks in the dsDNA molecule; and

b. repair of the dsDNA molecule in the cell by non-homologous end joining (NHEJ).

In another aspect the present invention provides a method of modifying a target nucleic acid in a cell; wherein the existing nucleotide sequence is modified or deleted and/or a desired nucleotide sequence is inserted at a desired location wherein the method includes the steps of;

a. making a double stranded break at the desired locus; and

b. repair of the dsDNA molecule in the cell by homologous recombination.

In another aspect the present invention provides a method of modifying gene expression in a cell comprising modifying a target nucleic acid sequence as hereinbefore described; wherein the nucleic acid is dsDNA and the functional moiety is selected from a DNA modifying enzyme (e.g. a methylase or acetylase), a transcription activator or a transcription repressor.

In another aspect the present invention provides a method of modifying gene expression in a cell comprising modifying a target nucleic acid sequence as hereinbefore described, wherein the nucleic acid is an mRNA and the functional moiety is a ribonuclease; optionally selected from an endonuclease, a 3' exonuclease or a 5' exonuclease.

In another aspect the present invention provides a method of modifying a target nucleic acid as hereinbefore described, wherein the method is carried out at a temperature between 50°C and 100°C. Preferably, the method is carried out at a

temperature at or above 60°C. More preferably, the method is carried out at a temperature between 60°C and 80°C. Optimally, the method is carried out at a temperature between60°C and 65°C.

In any of the methods of modifying a target nucleic acid hereinbefore described, the cell may be a prokaryotic cell or alternatively, may be a eukaryotic cell.

Host Cells

Advantageously, the present invention is of broad applicability and host cells of the present invention may be derived from any genetically tractable organism which can be cultured. Accordingly, the present invention provides a host cell transformed by a method as hereinbefore described.

Appropriate host cells may be prokaryotic or eukaryotic. In particular, commonly used host cells may be selected for use in accordance with the present invention including prokaryotic or eukaryotic cells which are genetically accessible and which can be cultured, for example prokaryotic cells, fungal cells, plant cells and animal cells including human cells (but not embryonic stem cells). Preferably, host cells will be selected from a prokaryotic cell, a fungal cell, a plant cell, a protist cell or an animal cell. Preferred host cells for use in accordance with the present invention are commonly derived from species which typically exhibit high growth rates, are easily cultured and/or transformed, display short generation times, species which have established genetic resources associated with them or species which have been selected, modified or synthesized for optimal expression of heterologous protein under specific conditions. In preferred embodiments of the invention where the protein of interest is eventually to be used in specific industrial, agricultural, chemical or therapeutic contexts, an appropriate host cell may be selected based on the desired specific conditions or cellular context in which the protein of interest is to be deployed. Preferably the host cell will be a prokaryotic cell. In preferred embodiments the host cell is a bacterial cell. The host cell may for instance be an *Escherichia coli* (*E. coli*) cell. Preferably the host cell will be a cell of a thermophilic bacterium.

BRIEF DESCRIPTION OF THE FIGURES

The invention will now be described in detail with reference to a specific embodiment and with reference to the accompanying drawings, in which:

Figure 1 shows a Neighbour-Joining tree of Cas9 protein sequences. All sequences having a sequence similarity above 40% with strain T12 based on pBLAST or PSI-BLAST were included, as well as currently well-characterized sequences (*S. pyogenes, S. thermophiles and A. naeslundii*), as well as all currently identified thermophilic sequences also when these were below 40% identity. For all thermophilic sequences, the percentage identity to T12 is indicated after the strain name. Gene identifier (gi) numbers are indicated before the species name. Legend: Closed circles: thermophilic (optimum above 60°C) Cas9 sequences, closed squares: thermotolerant (optimum <50°C) Cas9 sequences, open triangle: Cas9 sequence currently most used for genome editing purposes from mesophilic origin; no sign: mesophilic Cas9. Values at the nodes represent 1000-replicate bootstrap values; scale bar represents estimated amino acid substitutions per site.

Figure 2 shows a Neighbour-Joining tree of Cas9 gene sequences. Identity at the gene level was extremely poor; sequences from the same organisms as those used for the protein alignment were used for the gene alignment. Gene identifier (gi) numbers are indicated before the species name. Legend: Closed circles: thermophilic (optimum above 60°C) Cas9 sequences, closed squares: thermotolerant (optimum <50°C) Cas9 sequences, open triangle: Cas9 sequence currently most used for genome editing purposes from mesophilic origin; no sign: mesophilic Cas9. Values at the nodes represent 1000-replicate bootstrap values.

Figure 3 shows a protein sequence alignment for T12-Cas9 (SEQ ID NO: 1) (Type II-C) with well-characterized Type II-C (*A. naeslundii*'ana'; SEQ ID NO: 8) and Type II-A (*S. pyogenes*/'pyo'; SEQ ID NO: 9 and *S. thermophilus*) Cas9 sequences. Important active site residues are well conserved and indicated with black arrows. Protein domains as described for Ana-Cas9 and Pyo-Cas9 (Jinek, *et al.*, 2014, Science 343: 1247997) are indicated with shaded boxes and similarly coloured

letters. The PAM recognition domain has been determined for the *S. pyogenes* Type II-A system but not for any Type II-C system and is therefore only indicated in the *S. pyogenes* sequence.

Figure 4 shows protein architecture of *A. naeslundii* Cas9 (Cas9-Ana) (Jinek *et al.*, 2014). T12-Cas9 belongs to the same Type II-C CRISPR system and active site residues could be identified.

Figure 5 shows a comparison of crRNA-guided targeting of complementary dsDNA. Base pairing is indicated with dashed lines. RNA is depicted in black, DNA in grey. Base pairing between crRNA spacer and target protospacer is indicated with thick black dashed line, base pairing between DNA strands and between RNA strands is indicated with thick grey dashed lines. The 5' end of the crRNA is indicated. Note that PAM (small white box) in Type I resides downstream of target strand (protospacer), whereas in Type II it resides at the other end on the displaced strand. Likewise, the seed (the predicted sequence of the guide where base pairing with target DNA strand starts, and where no mismatches are allowed) is located close to the PAM, and as such differs in types I and II (Van der Oost, 2014 *ibid.*). Panel A shows a schematic of a Type I Cascade system of *E. coli*. crRNA has internal spacer (grey box, 31-32 nt that allows for target recognition), flanked bt a 8 nt 5' handle and a 29 nt 3' handle that consists of a stem-loop structure (hairpin) (Jore 2011 *ibid.*). Panel B shows a schematic of a Type II Cas9 system of *S. pyogenes*. crRNA basepairs with tracrRNA, that allows for processing by RNaseIII (opposite black triangles). Additionally, the 5' end of the crRNA is trimmed by an RNase (black triangle), typically resulting in a 20 nt spacer. Note that a synthetic loop may be introduced to link the crRNA and tracrRNA, resulting in a single guide RNA (sgRNA) (Jinek *et al.*, 2012 *ibid.*).

Below are polynucleotide and amino acid sequences of Cas proteins used in accordance with the invention.

[SEQ ID NO: 1] *Geobacillus thermodenitrificans* T12 Cas9 protein AA sequence
MKYKIGLDIGITSIGWAVINLDIPRIEDLGVRIFDRAENPKTGESLALPRRLARSARRR
LRRRKHRLERIRRLFVREGILTKEELNKLFEKKHEIDVWQLRVEALDRKLNNDELARI

LLHLAKRRGFRSNRKSERTNKENSTMLKHIEENQSILSSYRTVAEMVVKDPKFSLH
KRNKEDNYTNTVARDDLEREIKLIFAKQREYGNIVCTEAFEHEYISIWASQRPFASK
DDIEKKVGFCTFEPKEKRAPKATYTFQSFTVWEHINKLRLVSPGGIRALTDDERRLIY
KQAFHKNKITFHDVRTLLNLPDDTRFKGLLYDRNTTLKENEKVRFLELGAYHKIRKAI
DSVYGKGAAKSFRPIDFDTFGYALTMFKDDTDIRSYLRNEYEQNGKRMENLADKVY
DEELIEELLNLSFSKFGHLSLKALRNILPYMEQGEVYSTACERAGYTFTGPKKKQKT
VLLPNIPPIANPVVMRALTQARKVVNAIIKKYGSPVSIHIELARELSQSFDERRKMQK
EQEGNRKKNETAIRQLVEYGLTLNPTGLDIVKFKLWSEQNGKCAYSLQPIEIERLLE
PGYTEVDHVIPYSRSLDDSYTNKVLVLTKENREKGNRTPAEYLGLGSERWQQFETF
VLTNKQFSKKKRDRLLRLHYDENEENEFKNRNLNDTRYISRFLANFIREHLKFADSD
DKQKVYTVNGRITAHLRSRWNFNKNREESNLHHAVDAAIVACTTPSDIARVTAFYQ
RREQNKELSKKTDPQFPQPWPHFADELQARLSKNPKESIKALNLGNYDNEKLESL
QPVFVSRMPKRSITGAAHQETLRRYIGIDERSGKIQTVVKKKLSEIQLDKTGHFPMY
GKESDPRTYEAIRQRLLEHNNDPKKAFQEPLYKPKKNGELGPIIRTIKIIDTTNQVIPL
NDGKTVAYNSNIVRVDVFEKDGKYYCVPIYTIDMMKGILPNKAIEPNKPYSEWKEMT
EDYTFRFSLYPNDLIRIEFPREKTIKTAVGEEIKIKDLFAYYQTIDSSNGGLSLVSHDN
NFSLRSIGSRTLKRFEKYQVDVLGNIYKVRGEKRVGVASSSHSKAGETIRPL*

[SEQ ID NO: 7] ***Geobacillus thermodenitrificans*** **T12 Cas9 DNA Sequence**

ATGAAGTATAAAATCGGTCTTGATATCGGCATTACGTCTATCGGTTGGGCTGTC
ATTAATTTGGACATTCCTCGCATCGAAGATTTAGGTGTCCGCATTTTTGACAGAG
CGGAAAACCCGAAAACCGGGGAGTCACTAGCTCTTCCACGTCGCCTCGCCCGC
TCCGCCCGACGTCGTCTGCGGCGTCGCAAACATCGACTGGAGCGCATTCGCC
GCCTGTTCGTCCGCGAAGGAATTTTAACGAAGGAAGAGCTGAACAAGCTGTTT
GAAAAAAAGCACGAAATCGACGTCTGGCAGCTTCGTGTTGAAGCACTGGATCG
AAAACTAAATAACGATGAATTAGCCCGCATCCTTCTTCATCTGGCTAAACGGCG
TGGATTTAGATCCAACCGCAAGAGTGAGCGCACCAACAAAGAAAACAGTACGAT
GCTCAAACATATTGAAGAAAACCAATCCATTCTTTCAAGTTACCGAACGGTTGCA
GAAATGGTTGTCAAGGATCCGAAATTTTCCCTGCACAAGCGTAATAAAGAGGAT
AATTACACCAACACTGTTGCCCGCGACGATCTTGAACGGGAAATCAAACTGATT
TTCGCCAAACAGCGCGAATATGGGAACATCGTTTGCACAGAAGCATTTGAACAC
GAGTATATTTCCATTTGGGCATCGCAACGCCCTTTTGCTTCTAAGGATGATATC
GAGAAAAAAGTCGGTTTCTGTACGTTTGAGCCTAAAGAAAAACGCGCGCCAAAA
GCAACATACACATTCCAGTCCTTCACCGTCTGGGAACATATTAACAAACTTCGT

CTTGTCTCCCCGGGAGGCATCCGGGCACTAACCGATGATGAACGTCGTCTTAT
ATACAAGCAAGCATTTCATAAAAATAAAATCACCTTCCATGATGTTCGAACATTG
CTTAACTTGCCTGACGACACCCGTTTTAAAGGTCTTTTATATGACCGAAACACCA
CGCTGAAGGAAAATGAGAAAGTTCGCTTCCTTGAACTCGGCGCCTATCATAAAA
TACGGAAAGCGATCGACAGCGTCTATGGCAAAGGAGCAGCAAAATCATTTCGT
CCGATTGATTTTGATACATTTGGCTACGCATTAACGATGTTTAAAGACGACACCG
ACATTCGCAGTTACTTGCGAAACGAATACGAACAAAATGGAAAACGAATGGAAA
ATCTAGCGGATAAAGTCTATGATGAAGAATTGATTGAAGAACTTTTAAACTTATC
GTTTTCTAAGTTTGGTCATCTATCCCTTAAAGCGCTTCGCAACATCCTTCCATAT
ATGGAACAAGGCGAAGTCTACTCAACCGCTTGTGAACGAGCAGGATATACATTT
ACAGGGCCAAAGAAAAAACAGAAAACGGTATTGCTGCCGAACATTCCGCCGAT
CGCCAATCCGGTCGTCATGCGCGCACTGACACAGGCACGCAAAGTGGTCAATG
CCATTATCAAAAAGTACGGCTCACCGGTCTCCATCCATATCGAACTGGCCCGG
GAACTATCACAATCCTTTGATGAACGACGTAAAATGCAGAAAGAACAGGAAGGA
AACCGAAAGAAAAACGAAACTGCCATTCGCCAACTTGTTGAATATGGGCTGACG
CTCAATCCAACTGGGCTTGACATTGTGAAATTCAAACTATGGAGCGAACAAAAC
GGAAAATGTGCCTATTCACTCCAACCGATCGAAATCGAGCGGTTGCTCGAACCA
GGCTATACAGAAGTCGACCATGTGATTCCATACAGCCGAAGCTTGGACGATAG
CTATACCAATAAAGTTCTTGTGTTGACAAAGGAGAACCGTGAAAAAGGAAACCG
CACCCCAGCTGAATATTTAGGATTAGGCTCAGAACGTTGGCAACAGTTCGAGAC
GTTTGTCTTGACAAATAAGCAGTTTTCGAAAAAGAAGCGGGATCGACTCCTTCG
GCTTCATTACGATGAAAACGAAGAAAATGAGTTTAAAAATCGTAATCTAAATGAT
ACCCGTTATATCTCACGCTTCTTGGCTAACTTTATTCGCGAACATCTCAAATTCG
CCGACAGCGATGACAAACAAAAAGTATACACGGTCAACGGCCGTATTACCGCC
CATTTACGCAGCCGTTGGAATTTTAACAAAAACCGGGAAGAATCGAATTTGCAT
CATGCCGTCGATGCTGCCATCGTCGCCTGCACAACGCCGAGCGATATCGCCCG
AGTCACCGCCTTCTATCAACGGCGCGAACAAAACAAAGAACTGTCCAAAAAGAC
GGATCCGCAGTTTCCGCAGCCTTGGCCGCACTTTGCTGATGAACTGCAGGCGC
GTTTATCAAAAAATCCAAAGGAGAGTATAAAAGCTCTCAATCTTGGAAATTATGA
TAACGAGAAACTCGAATCGTTGCAGCCGGTTTTTGTCTCCCGAATGCCGAAGC
GGAGCATAACAGGAGCGGCTCATCAAGAAACATTGCGGCGTTATATCGGCATC
GACGAACGGAGCGGAAAAATACAGACGGTCGTCAAAAAGAAACTATCCGAGAT
CCAACTGGATAAAACAGGTCATTTCCCAATGTACGGGAAAGAAAGCGATCCAAG
GACATATGAAGCCATTCGCCAACGGTTGCTTGAACATAACAATGACCCAAAAAA

GGCGTTTCAAGAGCCTCTGTATAAACCGAAGAAGAACGGAGAACTAGGTCCTAT
CATCCGAACAATCAAAATCATCGATACGACAAATCAAGTTATTCCGCTCAACGAT
GGCAAAACAGTCGCCTACAACAGCAACATCGTGCGGGTCGACGTCTTTGAGAA
AGATGGCAAATATTATTGTGTCCCTATCTATACAATAGATATGATGAAAGGGATC
TTGCCAAACAAGGCGATCGAGCCGAACAAACCGTACTCTGAGTGGAAGGAAAT
GACGGAGGACTATACATTCCGATTCAGTCTATACCCAAATGATCTTATCCGTATC
GAATTTCCCCGAGAAAAAACAATAAAGACTGCTGTGGGGGAAGAAATCAAAATT
AAGGATCTGTTCGCCTATTATCAAACCATCGACTCCTCCAATGGAGGGTTAAGT
TTGGTTAGCCATGATAACAACTTTTCGCTCCGCAGCATCGGTTCAAGAACCCTC
AAACGATTCGAGAAATACCAAGTAGATGTGCTAGGCAACATCTACAAAGTGAGA
GGGGAAAAGAGAGTTGGGGTGGCGTCATCTTCTCATTCGAAAGCCGGGGAAAC
TATCCGTCCGTTATAA

DETAILED DESCRIPTION

Example 1: Isolation of *G. thermodenitrificans*

G. thermodenitrificans was surprisingly discovered during a search of a library of ± 500 isolates for a thermophile capable of degrading lignocellulosic substrates under anaerobic conditions. At first a library of ± 500 isolates was established which, after several selection rounds by isolation on cellulose and xylan, was trimmed down to 110 isolates. This library of 110 isolates consisted solely of *Geobacillus* isolates with *G. thermodenitrificans* representing 79% of the library.

The isolated *G. thermodenitrificans* strain has been named “T12”.

Example 2: Defining the essential consensus sequences for Cas9 in *Geobacillus thermodenitrificans*

The following database searches and alignments were performed:

pBLAST and nBLAST were performed on the in-house BLAST server, in which either the protein or gene sequence of *G. thermodenitrificans* T12 was used as query sequence. This database was last updated May 2014 and therefore does not contain the most recently added *Geobacillus* genomes, but normal online BLAST

was not used to prevent publication of the T12 sequence. See Appendix 1 for the results of the in-house pBLAST in Excel format, the most relevant sequences of this (sequence identities greater than 40% are included in Figure 1).

To include more recent sequence data, the sequence of *Geobacillus* MAS1 (most closely related to T12 Cas9) was used to perform a PSI-BLAST on the NCBI website (Johnson *et al.*, 2008 Nucleic Acids Res. 36(Web Server issue): W5-9). Two consecutive rounds of PSI-BLAST were performed, in which only sequences that met the following criteria were used for the next round: minimum sequence coverage of 96% in the first round and 97% in the second and third round, minimum identity 40%, only one strain per species.

The sequences resulting from the PSI-BLAST, as well as the sequences with more than 40% identity to T12 from the internal server pBLAST that did not appear in the PSI-BLAST were aligned together with currently well-characterized mesophilic sequences and all currently identified thermophilic sequences also if these were more distantly related, from which a Neighbour-Joining tree was constructed (see Figure 1). Alignment was performed in Mega6 using ClustalW, after which a tree was constructed using the Neighbour-Joining method and bootstrap analysis was performed using 1000 replicates.

When BLASTn was performed using *Geobacillus sp.* MAS1 as the query sequence, only *Geobacillus sp.* JF8 Cas9 was identified with 88% identity, indicating very little homology at the gene level. Figure 2 is a Neighbour-Joining tree of Clustal-aligned Cas9 gene sequences.

Protein sequences of *G. thermodenitrificans* T12, *A. naeslundii* and *S. pyogenes* were further analyzed for protein domain homology (see Figure 3) by aligning them in CloneManager using BLOSUM62 with default settings.

Example 3: Identifying core amino acid motifs which are essential for the function of CAS9 and those which confer thermostability in thermophilic Cas9 nucleases

Percentages identity of the above described aligned protein sequences are provided in Figure 1. T12-Cas9 belongs to Type II-C. The best-studied and recently crystalized structure of a Type II-C system is from *Actinomyces naeslundii* (Jinek *et al.*, 2014, Science 343: 1247997). This protein sequence shows only 20% identity to T12-Cas9 but can be used to estimate highly conserved residues. Two well-characterized Type II-A systems (*S. pyogenes* and *S. thermophilus*) were also included in the analyses (Jinek *et al.*, 2014, Science 343: 1247997; Nishimasu *et al.*, 2014, Cell 156: 935-949). Alignments of these four protein sequences are shown in Figure 3; Figure 4 shows the protein architecture as determined for *A. naeslundii* ('Ana-Cas9') (Jinek *et al.*, 2014, Science 343: 1247997). The length of Cas9 from t12 and *Actinomyces naeslundii* is highly similar (*A. naeslundii* 1101 aa, T12 1082 aa) and T12 is expected to have similar protein architecture but this remains to be determined, as the overall sequence identity to cas9-Ana is only 20%. All active side residues described by Jinek *et al.* (Jinek *et al.*, 2014, Science 343: 1247997) in Cas9 from *A. naeslundii* and *S. pyogenes* could be identified in T12-Cas9 (see Figure 3). The PAM-binding domain has been determined for the *S. pyogenes* Type II-A system but not for any Type II-C system and is therefore only indicated in the *S. pyogenes* sequence. Moreover, the PAM-recognition site varies strongly, not only between CRISPR systems but also between species containing the same system. For more information on the PAM please see question 4 and future planning.

Example 4: Determination of the PAM sequence of *G. thermodenitrificans* T12 Cas9

It has been established that the prokaryotic CRISPR systems serve their hosts as adaptive immune systems (Jinek *et al.*, 2012, Science 337: 816-821) and can be used for quick and effective genetic engineering (Mali *et al.*, 2013, Nat Methods 10: 957-963.).

Cas9 proteins function as sequence-specific nucleases for the type II CRISPR systems (Makarova *et al.*, 2011, Nat Rev Micro 9: 467-477). Small crRNA molecules, which consist of a "spacer" (target) linked to a repetition region, are the transcription and processing products of a CRISPR loci. "Spacers" naturally originate from the genome of bacteriophages and mobile genetic elements, but they

can also be designed to target a specific nucleotide sequence during a genetic engineering process (Bikard *et al.*, 2013, Nucleic Acids Research 41: 7429-7437). The crRNA molecules are employed by the Cas9 as guides for the identification of their DNA targets. The spacer region is identical to the targeted for cleavage DNA region, the "protospacer" (Brouns *et al.*, 2012, Science 337: 808-809). A PAM (Protospacer Adjacent Motif), next to the protospacer, is required for the recognition of the target by the Cas9 (Jinek *et al.*, 2012, Science 337: 816-821).

Example 5: Target generation with randomized PAM

Two different spacers from the CRISPR II loci of the *G. thermodenitrificans* T12 strain were amplified by PCR using the *G. thermodenitrificans* T12 genomic DNA as template. Two pairs of degenerate primers were used for the amplification of each spacer:

Firstly, a pair that cause the introduction of six random nucleotides upstream of the "protospacer" fragment were used, leading to the production of a pool of protospacers with randomized PAM sequences.

Secondly, a pair that cause the introduction of six random nucleotides downstream of the "protospacer" fragment were used, leading to the production of a pool of protospacers with randomized PAM sequences.

The produced fragments were ligated to the pNW33n vector, producing 4 pools of "protospacer" constructs, with all the possible 4096 different combinations of 6-nucleotide long PAMs each. The assembled DNA was used for the transformation of *G. thermodenitrificans* T12 cells. The cells were plated on chloramphenicol selection and more than 2 x 10^6 cells from each protospacer pool will be pooled. The plasmid DNA was extracted from the pools, the target region will be PCR amplified and the products sent for deep sequencing. The PAMs with the fewest reads will be considered active and the process will be repeated only with pNW33n constructs that contain spacers with these PAMs. Reduced transformation efficiency of the *G. thermodenitrificans* T12 will confirm the activity of the PAMs.

Example 6. Functional temperature ranges for *Geobacillus thermodenitrificans* Cas9.

No temperature range experiments have been conducted to date on any publically available Cas9 protein. The Cas9 proteins used in research all have a mesophilic origin with a maximum growth temperature of 45°C for the host organism *Streptococcus pyogenes* (ATCC_700294).

Here the range of temperatures within which Cas9 from *Geobacillus thermodenitrificans* is stable and functional and further the temperature at which its activity is optimal was experimentally determined. Furthermore, the functional and preferred temperatures ranges for Cas9 nucleases from other *Geobacillus* species were also determined.

Temperature ranges of *Geobacillus thermodenitrificans* T12 Cas9 nucleases have been determined *in vitro* by using a single guide RNA (sgRNA). The sgRNA consists of the endogenous bacterial crRNA and tracrRNA into a single chimeric transcript and has been created by making use of the pT7 cloning vector containing the crRNA:tracrRNA backbone, preceded by a 20 basepair (bp) target sequence (Jao *et al.,* 2013, PNAS 110: 13904-13909). The target sequence starts with the 20 bp target and ends with the T12 PAM sequence (derived from Example 4). The desired single targeting sgRNA has been generated by *in vitro* transcription from a T7 promoter after linearizing the plasmid. The standard cloning vector pUC19 has been used as a target in our in vitro activity assays. A target site (20nt + PAM) in the plasmid has been constructed by two complementary oligos which have been annealed and ligated into the sgRNA cloning vector. The produced sgRNA was incubated together with the Cas9 nuclease of T12 at temperatures ranging from 20°C to 100°C. Cas9 cleavage activity has been assayed by agarose gel electrophoreses.

The organism isolated and characterised here (T12) has an optimum growth temperature of 65°C which also represents the optimal temperature for its Cas9

protein. The temperature range in which it is active would be from 20°C to 100°C, in particular 20°C to 80°C.

The optimal temperature range of Cas9 from *Geobacillus* species is much higher than that of Cas9 proteins which have been characterised to date. Similarly the upper extent of the range in which it retains nuclease activity is much higher than that of known Cas9 proteins. A higher optimal temperature and functional range provides a significant advantage in genetic engineering at high temperatures and therefore in editing the genomes of thermophilic organisms, which have utility in a range of industrial, agricultural and pharmaceutical processes conducted at elevated temperatures.