The invention relates to chimeric endonucleases, comprising an endonuclease and a heterologous DNA binding domain, as well as methods of targeted integration, targeted deletion or targeted mutation of polynucleotides using chimeric endonucleases.
Chimeric Endonucleases and Uses Thereof

FIELD OF THE INVENTION

The invention relates to chimeric endonucleases, comprising a endonuclease and a heterologous DNA binding domain, as well as methods of targeted integration, targeted deletion or targeted mutation of polynucleotides using chimeric endonucleases.

BACKGROUND OF THE INVENTION

Genome engineering is a common term to summarize different techniques to insert, delete, substitute or otherwise manipulate specific genetic sequences within a genome and has numerous therapeutic and biotechnological applications. More or less all genome engineering techniques use recombinases, integrases or endonucleases to create DNA double strand breaks at predetermined sites in order to promote homologous recombination. In spite of the fact that numerous methods have been employed to create DNA double strand breaks, the development of effective means to create DNA double strand breaks at highly specific sites in a genome remains a major goal in gene therapy, agrotechnology, and synthetic biology.

One approach to achieve this goal is to use nucleases with specificity for a sequence that is sufficiently large to be present at only a single site within a genome. Nucleases recognizing such large DNA sequences of about 15 to 30 nucleotides are therefore called "meganucleases" or "homing endonucleases" and are frequently associated with parasitic or selfish DNA elements, such as group I self-splicing introns and inteins commonly found in the genomes of plants and fungi. Meganucleases are commonly grouped into four families: the LAGLIDADG family, the GIY-YIG family, the His-Cys box family and the HNH family. These families are characterized by structural motifs, which affect catalytic activity and the sequence of their DNA recognition sequences.

Natural meganucleases from the LAGLIDADG family have been used to effectively promote site-specific genome modifications in insect and mammalian cell cultures, as well as in many organisms, such as plants, yeast or mice, but this approach has been limited to the modification of either homologous genes that conserve the DNA recognition sequence or to preengineered genomes into which a recognition sequence has been introduced. In order to avoid these limitations and to promote the systematic implementation of DNA double strand break stimulated gene modification new types of nucleases have been created.

One type of new nucleases consists of artificial combinations of unspecific nucleases to a highly specific DNA binding domain. The effectiveness of this strategy has been demonstrated in a variety of organisms using chimeric fusions between an engineered zinc finger DNA-binding domain and the non-specific nuclease domain of the FokI restriction enzyme (e.g. WO03/089452) a variation of this approach is to use an inactive variant of a meganuclease as DNA binding domain fused to an unspecific nuclease like FokI as disclosed in Lippow et al., "Creation of a type IIS restriction endonuclease with a long recognition sequence", Nucleic Acid Research (2009), Vol.37, No.9, pages 3061 to 3073.

An alternative approach is to genetically engineer natural meganucleases in order to customize their DNA binding regions to bind existing sites in a genome, thereby creating engineered meganucleases having new specificities (e.g WO07093918, WO2008/093249, WO091 14321).
However, many meganucleases which have been engineered with respect to DNA cleavage specificity have decreased cleavage activity relative to the naturally occurring meganucleases from which they are derived (US2010/0071083). Most meganucleases do also act on sequences similar to their optimal binding site, which may lead to unintended or even detrimental off-target effects. Several approaches have already been taken to enhance the efficiency of meganuclease induced homologous recombination e.g. by fusing nuclease to the ligand binding domain of the rat Glucocorticoid Receptor in order to promote or even induce the transport of this modified nuclease to the cell nucleus and therefore its target sites by the addition of dexamethasone or similar compounds (WO2007/135022). Despite that fact, there is still a need in the art to develop meganucleases having high induction rates of homologous recombination and/or a high specificity for their binding site, thereby limiting the risk of off-target effects.

BRIEF SUMMARY OF THE INVENTION
The invention provides chimeric endonucleases comprising at least one endonuclease having DNA double strand break inducing activity and at least one heterologous DNA binding domain. Preferably at least one endonuclease of the chimeric endonuclease is a LAGLIDADG endonuclease. In one embodiment, at least one LAGLIDADG endonuclease is I-SceI, I-CreI, I-CeuI, I-Chui, I-Dmol, PI-SceI, I-MsoI, or I-Anil, or a LAGLIDADG endonuclease having at least 45% amino acid sequence identity to any one of these. In another embodiment of the invention, at least one LAGLIDADG endonuclease has at least 80% amino acid sequence identity to a polypeptide described by SEQ ID NO: 1, 2, 3 or 159. The LAGLIDADG endonuclease may be wild-type, engineered, optimized or optimized engineered LAGLIDADG endonuclease.

The heterologous DNA binding domain is preferably a transcription factor or an inactive nuclease, or a fragment comprising a DNA binding domain of a transcription factor or a nuclease.

In one embodiment at least one heterologous DNA binding domain is an inactive I-SceI, I-CreI, I-CeuI, I-Chui, I-Dmol, PI-SceI, I-MsoI, or I-Anil or an inactive homolog of these having at least 45% amino acid sequence identity. In one embodiment the heterologous DNA binding domain is an inactive version of a LAGLIDADG endonuclease having an amino acid sequence as described by at least one of SEQ ID NO: 1, 2, 3, 5, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 142 or 159, preferably having an amino acid sequence as described by any one of SEQ ID NO: 1, 2, 3, 5 or 159.

In another embodiment of the invention the heterologous DNA binding domain is a transcription factor or an DNA binding domain of a transcription factor. Preferably the transcription factor or the DNA binding domain of a transcription factor comprises a HTH domain. Even more preferred, the transcription factor or the DNA binding domain of a transcription factor comprises a HTH domain comprising an amino acid sequence of at least 80% sequence identity to at least one amino acid sequence described by SEQ ID NO: 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118 or 119, preferably described by 91, 92, 93, 94, 95, 112, 113, 114, 115, 116, 117, 118 or 119. In one embodiment of the invention, the heterologous DNA binding domain comprises a polypeptide having at least 80% amino acid sequence identity to a polypeptide described by SEQ ID NO: 6, 7 or 8. Preferably the chimeric endonuclease comprises a linker (or synonymous linker
polypeptide) to connect at least one endonuclease with at least one heterologous DNA binding domain. The chimeric endonuclease may comprise one or more NLS-sequences or one or more SecI or SecIV secretion signals or a combination of one or more NLS-sequences and one or more SecI or SecIV secretion signals or a combination of one or more SecI and SecIV secretion signals with one or more NLS-sequences. In one embodiment of the invention the DNA binding activity of the heterologous DNA binding domain is inducible. In another embodiment of the invention, the DNA double strand break inducing activity of the endonuclease is inducible by expression of the second monomer of a homo- or heterodimeric endonuclease, preferably a homo- or heterodimeric LAGLIDADG endonuclease. The chimeric endonucleases may comprise at least one NLS-sequence or at least one SecI or at least one SecIV secretion signal or a combination of one or more NLS-sequences, one or more SecI secretion signals or one or more SecIV secretion signals.

The invention does further provide isolated polynucleotides coding for a chimeric endonuclease. Preferably the isolated polynucleotide coding for a chimeric endonuclease is codon optimized, or has a low content of RNA instability motifs, or has a low content of cryptic splice sites, or has a low content of alternative start codons, or has a low content of restriction sites, or has a low content of RNA secondary structures, or has a combination of the features described above. A further embodiment of the invention is an expression cassette comprising an isolated polynucleotide coding for a chimeric endonuclease in functional combination with a promoter and an terminator sequence. An additional group of isolated polynucleotides provided by the invention are isolated polynucleotides comprising a chimeric recognition sequence having a length of about 15 to about 300 nucleotides and comprising a recognition sequence of an endonuclease and a recognition sequence of a heterologous DNA binding domain. Preferably the chimeric recognition sequence comprises a DNA recognition sequence of a LAGLIDADG endonuclease, even more preferred a DNA recognition sequence of a LAGLIDADG endonuclease having an amino acid sequence as described by at least one of SEQ ID NOs: 1, 2, 3, 5, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 142 or 159, preferably having an amino acid sequence as described by SEQ ID NO: 1, 2, 3, 5 or 159. In a further embodiment of the invention, the chimeric recognition site comprises a DNA recognition sequence of L-Scl, I-Crel, I-Dmol, I-Msol, I-Ceul, I-Chul, PI-Scel or I-Anil, and a recognition sequence of a heterologous DNA binding domain having at least 50% sequence amino acid sequence identity to scTet, scArc, LacR, MerR or MarA or to a DNA binding domain fragment of scTet, scArc, LacR, MerR or MarA. Preferred polynucleotides provided by the invention comprise a chimeric recognition sequence, comprising a DNA recognition sequence of L-Scl and a recognition sequence of scTet or scArc, wherein the DNA recognition sequence of L-Scl and the recognition sequence of scTet or scArc are directly connected, or are connected via a linker linker sequence of 1 to 10 nucleotides. In a preferred embodiment the isolated polynucleotide comprises a chimeric recognition sequence comprising a polynucleotide sequence as described by any one of SEQ ID NOs: 14, 15, 16, 17, 18, 19 or 20.

The invention does further provide a vector, host cell or non human organism comprising an isolated polynucleotide coding for a chimeric endonuclease, or an isolated polynucleotide as described above, or an expression cassette, or an isolated polynucleotide comprising a chimeric
recognition sequence or a chimeric endonuclease or comprising a combination of one or more of these. Preferably the non-human organism is a plant.

The invention provides methods of using the chimeric endonucleases and chimeric recognition sequences described herein to induce or facilitate homologous recombination or end joining events. Preferably methods for targeted integration or excision of sequences. Preferably the sequences being excised are marker genes.

One embodiment of the invention is a method for providing a chimeric endonuclease, comprising the steps of: a) providing at least one endonuclease coding region, b) providing at least one heterologous DNA binding domain coding region, c) providing a polynucleotide having a potential DNA recognition sequence or potential DNA recognition sequences of the endonuclease or endonucleases of step a) and having a potential recognition sequence or having potential recognition sequences of the heterologous DNA binding domain or heterologous DNA binding domains of step b), d) creating a translational fusion of the coding regions of all endonucleases of step b) and all heterologous DNA binding domains of step c), e) expressing a chimeric endonuclease from the translational fusion created in step d), f) testing the chimeric endonuclease expressed in step e) for cleavage of the polynucleotide of step c).

The invention does further provide a method for homologous recombination of polynucleotides comprising the following steps: a) providing a cell competent for homologous recombination, b) providing a polynucleotide comprising a chimeric recognition site flanked by a sequence A and a sequence B, c) providing a polynucleotide comprising sequences A' and B', which are sufficiently long and homologous to sequence A and sequence B, to allow for homologous recombination in said cell and d) providing a chimeric endonuclease as described herein or an expression cassette as described herein, e) combining b), c) and d) in said cell and f) detecting recombined polynucleotides of b) and c), or selecting for or growing cells comprising recombined polynucleotides of b) and c). Preferably the method for homologous recombination of polynucleotides leads to a homologous recombination, wherein a polynucleotide sequence comprised in the competent cell of step a) is deleted from the genome of the growing cells of step f). A further method of the invention is a method for targeted mutation comprising the following steps: a) providing a cell comprising a polynucleotide comprising a chimeric recognition site of an chimeric endonuclease, b) providing an chimeric endonuclease being able to cleave the chimeric recognition site of step a), c) combining a) and b) in said cell and d) detecting mutated polynucleotides, or selecting for growing cells comprising mutated polynucleotides. In another preferred embodiment of the invention, the methods described above comprise a step, wherein the chimeric endonuclease and the chimeric recognition site are combined in at least one cell via crossing of organisms, via transformation or via transport mediated via a Sec III or SecIV peptide fused to the optimized endonuclease.

BRIEF DESCRIPTION OF THE FIGURES

Figure 1 depicts a sequence alignment of different I-SceI homologs, wherein 1 is SEQ ID NO: 1, 2 is SEQ ID NO: 56, 3 is SEQ ID NO: 57, 4 is SEQ ID NO: 58, 5 is SEQ ID NO: 59.
Figure 2 depicts a sequence alignment of different l-Crel homologs, wherein 1 is SEQ ID NO: 60, 2 is SEQ ID NO: 61, 3 is SEQ ID NO: 62, 4 is SEQ ID NO: 63, 5 is SEQ ID NO: 64.

Figures 3a to 3c depicts a sequence alignment of different PI-Scel homologs, wherein 1 is SEQ ID NO: 79, 2 is SEQ ID NO: 80, 3 is SEQ ID NO: 81, 4 is SEQ ID NO: 82, 5 is SEQ ID NO: 83.

Figure 4 depicts a sequence alignment of different l-Ceul homologs, wherein 1 is SEQ ID NO: 65, 2 is SEQ ID NO: 66, 3 is SEQ ID NO: 67, 4 is SEQ ID NO: 68, 5 is SEQ ID NO: 69.

Figure 5 depicts a sequence alignment of different l-Chul homologs, wherein 1 is SEQ ID NO: 70, 2 is SEQ ID NO: 71, 3 is SEQ ID NO: 72, 4 is SEQ ID NO: 73, 5 is SEQ ID NO: 74.

Figure 6 depicts a sequence alignment of different l-Dmol homologs, wherein 1 is SEQ ID NO: 75, 2 is SEQ ID NO: 76, 3 is SEQ ID NO: 77, 4 is SEQ ID NO: 78.

Figure 7 depicts a sequence alignment of different l-Msol homologs, wherein 1 is SEQ ID NO: 84 and 2 is SEQ ID NO: 85.

Figure 8 depicts a sequence alignment of different TetR homologs, wherein 1 is SEQ ID NO: 86, 2 is SEQ ID NO: 87, 3 is SEQ ID NO: 88, 4 is SEQ ID NO: 89, 5 is SEQ ID NO: 90.

Figure 9a depicts a sequence alignment of HTH domains of different TetR homologs, wherein 1 is SEQ ID NO: 91, 2 is SEQ ID NO: 92, 3 is SEQ ID NO: 93, 4 is SEQ ID NO: 94, 5 is SEQ ID NO: 95.

Figure 9b depicts a sequence alignment of HTH domains of different ArcR homologs, wherein 1 is SEQ ID NO: 96, 2 is SEQ ID NO: 97, 3 is SEQ ID NO: 98, 4 is SEQ ID NO: 99, 5 is SEQ ID NO: 100.

Figure 10a depicts a sequence alignment of HTH domains of different LacR homologs, wherein 1 is SEQ ID NO: 101, 2 is SEQ ID NO: 102, 3 is SEQ ID NO: 103, 4 is SEQ ID NO: 104, 5 is SEQ ID NO: 105.

Figure 10b depicts a sequence alignment of HTH domains of different MerR homologs, wherein 1 is SEQ ID NO: 106, 2 is SEQ ID NO: 107, 3 is SEQ ID NO: 108, 4 is SEQ ID NO: 109, 5 is SEQ ID NO: 110, 6 is SEQ ID NO: 111.

Figure 11 depicts a sequence alignment of HTH domains of different MarA homologs, wherein 1 is SEQ ID NO: 112, 2 is SEQ ID NO: 113, 3 is SEQ ID NO: 114, 4 is SEQ ID NO: 115, 5 is SEQ ID NO: 116, 6 is SEQ ID NO: 117, 7 is SEQ ID NO: 118, 8 is SEQ ID NO: 119.

Figure 12 depicts a sequence alignment of different MarA homologs, wherein 1 is SEQ ID NO:
DESCRIPTION OF THE INVENTION

The invention provides chimeric endonucleases, which can be used as alternative DNA double strand break inducing enzymes. The invention also includes methods of using these chimeric endonucleases.

Chimeric endonucleases of the invention

The chimeric endonucleases of the invention comprise at least one endonuclease having DNA double strand break inducing activity and at least one heterologous DNA binding domain.

The endonuclease

Endonucleases suitable for the invention induce DNA double strand breaks in a DNA recognition sequence of at least 4, at least 6, at least 8, at least 10, at least 14, at least 16, at least 18 or at least 20 base pairs.

Preferred endonucleases induce double strand breaks in a DNA recognition sequence of at least 14 base pairs, more preferred of at least 16 base pairs, even more preferred of at least 18 base pairs.

The term "DNA recognition sequence" generally refers to those sequences which, under the conditions in a cell e.g., in a plant cell, enables recognition and cleavage by the endonuclease. Examples for DNA recognition sequences as well as endonucleases cutting those DNA recognition sequences can be found in Table 8 below.


Preferred homing endonucleases are GIY-YIG-, His-Cys box-, HNH- or LAGLIDADG -endonucleases. The GIY-YIG endonucleases have a GIY-YIG module of 70 to 100 amino acids length, which includes four or five conserved sequence motifs with four invariant residues (Van Roey et al (2002), Nature Struct. Biol. 9:806 to 811). His-Cys box endonucleases comprise a highly conserved sequence of histidines and cysteines over a region of several hundred amino acids.
acid residues. The HNH-endonucleases are defined by sequence motifs containing two pairs of conserved histidines surrounded by asparagine residues. Further information on His-Cys box- and HNH endonucleases is provided by Chevalier et al. (2001), Nucleic Acids Res. 29(18): 3757 to 3774).

Preferably, the homing endonuclease used in the chimeric endonucleases belongs to the group of LAGLIDADG endonucleases.

LAGLIDADG endonucleases can be found in the genomes of algae, fungi, yeasts, protozoan, chloroplasts, mitochondria, bacteria and archaea. LAGLIDADG endonucleases comprise at least one conserved LAGLIDADG motif. The name of the LAGLIDADG motif is based on a characteristic amino acid sequence appearing in all LAGLIDADG endonucleases. The term LAGLIDADG is an acronym of this amino acid sequence according to the one-letter-code as described in the STANDARD ST.25 i.e. the standard adopted by the PCIPI Executive Coordination Committee for the presentation of nucleotide and amino acid sequence listings in patent applications.

However, the LAGLIDADG motif is not fully conserved in all LAGLIDADG endonucleases, (see for example Chevalier et al. (2001), Nucleic Acids Res. 29(18): 3757 to 3774, or Dalgaard et al. (1997), Nucleic Acids Res. 25(22): 4626 to 4638), so that some LAGLIDADG endonucleases comprise some amino acid changes in their LAGLIDADG motif. LAGLIDADG endonucleases comprising only one LAGLIDADG motif usually act as homo- or heterodimers. LAGLIDADG endonucleases comprising two LAGLIDADG motifs act as monomers and comprise usually a pseudo-dimeric structure.

LAGLIDADG endonucleases can be isolated for example from polynucleotides of organisms mentioned for exemplary purposes in Table 1, 2, 3, 4, 5 and 6, or de novo synthesized by techniques known in the art, e.g. using sequence information available in public databases known to the person skilled in the art, for example Genbank Benson (2010), Nucleic Acids Res 38:D46-51 or Swissprot Boeckmann (2003), Nucleic Acids Res 31:365-70

A collection of LAGLIDADG endonucleases can be found in the PFAM-Database for protein families. The PFAM-Database accession number PF00961 describes the LAGLIDADG 1 protein family, which comprises about 800 protein sequences. PFAM-Database accession number PF03161 describes members of the LAGLIDADG 2 protein family, comprising about 150 protein sequences. An alternative collection of LAGLIDADG endonucleases can be found in the Inter-Pro data base, e.g. InterPro accession number IPR004860.

The term LAGLIDADG endonucleases shall also encompass artificial homo- und heterodimeric LAGLIDADG endonucleases, which can be created e.g. by modifying the protein-protein interaction regions of the monomers in order to promote homo- or heterodimer formation. Examples of artificial heterodimeric LAGLIDADG endonuclease comprising the LAGLIDADG endonuclease 1-Dmo I as one domain can be found in WO2009/074842 and WO2009/074873.
In addition to that, the term LAGLIDADG endonucleases shall also encompass artificial single chain endonucleases, which can be created by making translational fusions of monomers of homo- or heterodimeric LAGLIDADG endonucleases.

Accordingly in one embodiment of the invention, the chimeric endonucleases of the invention comprise at least one LAGLIDADG endonuclease.

In further embodiments the LAGLIDADG endonuclease comprised in the chimeric endonuclease can be a monomeric, homodimeric, artificial homo- or heterodimeric or artificial single chain LAGLIDADG endonuclease.

In one embodiment the LAGLIDADG endonuclease is a monomeric, homodimeric, heterodimeric, or artificial single chain LAGLIDADG endonuclease. Preferably the endonuclease is a monomeric or artificial single chain LAGLIDADG endonuclease.

Preferred LAGLIDADG endonucleases are: l-An/1, l-Sce I, l-Chu I, l-Dmo I, l-Cre I, l-Csm I, Pl-Sce I, Pl-Tli I, Pl-Mtu I, l-Ceu I, l-Sce II, l-Sce III, HO, Pl-Civ I, Pl Ctr I, Pl-Aae I, Pl-Bsu I, Pl-Dha I, Pl-Dra I, Pl-Mav I, Pl-Mch I, Pl-Mfu I, Pl-Mfl I, Pl-Mga I, Pl-Mgo I, Pl-Min I, Pl-Mka I, Pl-Mle I, Pl-Mma I, Pl-Msh I, Pl-Msm I, l-Mso I, Pl-Mth I, Pl-Mtu I, Pl-Mxe I, Pl-Npu I, Pl-Pfu I, Pl-Rma I, Pl-Spb I, Pl-Ssp I, Pl-Fac I, Pl-Mja I, Pl-Pho I, Pl-Tag I, Pl-Thy I, Pl-Tko I, and Pl-Tsp I and homologs of any one of these having at least 49%, 51%, 58%, 60%, 70%, 80%, 85%, 90%, 92%, 93%, 94%, 95%, 96%, 97%, 98% or 99% sequence identity on amino acid level; more preferred are: l-Sce I, l-Chu I, l-Dmo I, l-Cre I, l-Csm I, Pl-Pfu I, Pl-Sce I, Pl-Tli I, l-Mso I, Pl-Mtu I, l-Ceu I, l-Sce II, l-Sce III, and HO and homologs of any one of these having at least 49%, 51%, 58%, 60%, 70%, 80%, 85%, 90%, 92%, 93%, 94%, 95%, 96%, 97%, 98% or 99% sequence identity on amino acid level; even more preferred are: l-Sce I, l-Chu I, l-Dmo I, l-Cre I, l-Csm I, Pl-Sce I, Pl-Pfu I, Pl-Tli I, l-Mso I, Pl-Mtu I and l-Ceu I and homologs of any one of these having at least 49%, 51%, 58%, 60%, 70%, 80%, 85%, 90%, 92%, 93%, 94%, 95%, 96%, 97%, 98% or 99% sequence identity on amino acid level; still more preferred are l-Dmo I, l-Cre I, l-Sce I, l-Mso I and l-Chu I and homologs of any one of these having at least 49%, 51%, 58%, 60%, 70%, 80%, 85%, 90%, 92%, 93%, 94%, 95%, 96%, 97%, 98% or 99% sequence identity on amino acid level, most preferred is l-Sce I and homologs of l-Sce I having at least 49%, 51%, 58%, 60%, 70%, 80%, 85%, 90%, 92%, 93%, 94%, 95%, 96%, 97%, 98% or 99% sequence identity on amino acid level.


More preferred monomeric LAGLIDADG endonucleases are: l-Sce I, l-Chu I, l-Dmo I, l-Csm I,
Pl-Pfu I, Pl-Sce I, Pl-Tli I, Pl-Mtu I, l-Sce II, l-Sce III, and HO and homologs of any one of these having at least 49%, 51%, 58%, 60%, 70%, 80%, 85%, 90%, 92%, 93%, 94%, 95%, 96%, 97%, 98% or 99% sequence identity on amino acid level.

Even more preferred monomeric LAGLIDADG endonucleases are: l-Sce I, l-Chu I, l-Dmo I, l-Csm I, Pl-Sce I, Pl-Tli I, and Pl-Mtu I; homologs of any one these having at least 49%, 51%, 58%, 60%, 70%, 80%, 85%, 90%, 92%, 93%, 94%, 95%, 96%, 97%, 98% or 99% sequence identity on amino acid level.

Still more preferred monomeric LAGLIDADG endonucleases are: l-Dmo I, l-Sce I, and l-Chu I; homologs of any one these having at least 49%, 51%, 58%, 60%, 70%, 80%, 85%, 90%, 92%, 93%, 94%, 95%, 96%, 97%, 98% or 99% sequence identity on amino acid level.

One type of homolog LAGLIDADG endonucleases are artificial single chain LAGLIDADG endonucleases, which may comprise two sub-units of the same LAGLIDADG endonuclease, such as single-chain l-Cre, single-chain l-Ceu I or single-chain l-Ceu II as disclosed in WO03078619, or which may comprise two sub-units of different LAGLIDADG endonucleases. Artificial single chain LAGLIDADG endonucleases, which comprise two sub-units of different LAGLIDADG endonucleases are called hybrid meganucleases.


A particular preferred single chain LAGLIDADG endonuclease is single-chain l-Cre I.

Preferred dimeric LAGLIDADG endonucleases are: l-Cre I, l-Ceu I, l-Sce II, l-Mso I and l-Csm I and homologs of any one these having at least 49%, 51%, 58%, 60%, 70%, 80%, 85%, 90%, 92%, 93%, 94%, 95%, 96%, 97%, 98% or 99% sequence identity on amino acid level.

Preferred heterodimeric LAGLIDADG endonucleases are disclosed in WO 07/034262, WO
07/047859 and WO08093249.

Homologs of LAGLIDADG endonucleases can for example be cloned from other organisms or can be created by mutating LAGLIDADG endonucleases, e.g. by replacing, adding or deleting amino acids of the amino acid sequence of a given LAGLIDADG endonuclease, which preferably has no effect on its DNA-binding-affinity, its dimer formation affinity or will change its DNA recognition sequence.

As used herein, the term "DNA-binding affinity" means the tendency of a meganuclease or LAGLIDADG endonuclease to non-covalently associate with a reference DNA molecule (e.g. a DNA recognition sequence or an arbitrary sequence). Binding affinity is measured by a dissociation constant, \(K_D\) (e.g., the \(K_D\) of I-Crel for the WT DNA recognition sequence is approximately 0.1 nM). As used herein, a meganuclease has "altered" binding affinity if the \(K_D\) of the recombinant meganuclease for a reference DNA recognition sequence is increased or decreased by a statistically significant \((p < 0.05)\) amount relative to a reference meganuclease or LAGLIDADG endonuclease.

As used herein with respect to meganuclease monomers or LAGLIDADG endonuclease monomers, the term "affinity for dimer formation" means the tendency of a monomer to non-covalently associate with a reference meganuclease monomer or LAGLIDADG endonuclease monomer. The affinity for dimer formation can be measured with the same monomer (i.e., homodimer formation) or with a different monomer (i.e., heterodimer formation) such as a reference wild-type meganuclease or a reference LAGLIDADG endonuclease. Binding affinity is measured by a dissociation constant, \(K_D\). As used herein, a meganuclease has "altered" affinity for dimer formation, if the \(K_D\) of the recombinant meganuclease monomer or the recombinant LAGLIDADG endonuclease monomer for a reference meganuclease monomer or for a reference LAGLIDADG endonuclease is increased or decreased by a statistically significant \((p < 0.05)\) amount relative to a reference meganuclease monomer or the reference LAGLIDADG endonuclease monomer.

As used herein, the term "enzymatic activity" refers to the rate at which a meganuclease e.g. a LAGLIDADG endonuclease cleaves a particular DNA recognition sequence. Such activity is a measurable enzymatic reaction, involving the hydrolysis of phospho-diester-bonds of double-stranded DNA. The activity of a meganuclease acting on a particular DNA substrate is affected by the affinity or avidity of the meganuclease for that particular DNA substrate which is, in turn, affected by both sequence-specific and non-sequence-specific interactions with the DNA.

For example, it is possible to add nuclear localization signals to the amino acid sequence of a LAGLIDADG endonuclease and/or change one or more amino acids and/or delete parts of its sequence, e.g. parts of the N-terminus or parts of its C-terminus.

For example, it is possible to create a homolog LAGLIDADG endonuclease of l-Scel, by mutat-
ing amino acids of its amino acid sequence. Mutations which have little effect on the DNA binding affinity of l-Scel, or will change its DNA recognition sequence are: A36G, L40M, L40V, L43A, H91A and I123L.

In one embodiment of the invention, the homologs of LAGLIDADG endonucleases are being selected from the groups of artificial single chain LAGLIDADG endonucleases, including or not including hybrid meganucleases, homologs which can be cloned from other organisms, engineered endonucleases or optimized nuclease.

In one embodiment, the LAGLIDADG endonuclease is selected from the group comprising: l-See I, l-Cre I, l-Mso I, l-Ceu I, l-DMo I, l-Ani I, Pl-Sce I, l-Pfu I or homologs of any one these having at least 49%, 51%, 58%, 60%, 70%, 80%, 85%, 90%, 92%, 93%, 94%, 95%, 96%, 97%, 98% or 99% sequence identity on amino acid level.

In another embodiment the LAGLIDADG endonuclease is selected from the group comprising: l-Sce I, l-Chu I, l-Cre I, l-DMo I, l-Csm I, Pl-Sce I, Pl-Pfu I, Pl-77/1, Pl-Mtu I, and l-Ceu I and homologs of any one these having at least 49%, 51%, 58%, 60%, 70%, 80%, 85%, 90%, 92%, 93%, 94%, 95%, 96%, 97%, 98% or 99% sequence identity on amino acid level.

Table 1: Exemplary homologs of l-Scel, which can be cloned from other organisms.

<table>
<thead>
<tr>
<th>Uni-Prot Accession Nr.</th>
<th>Organism</th>
<th>SEQ ID NO:</th>
<th>Amino Acid Sequence Identity to l-Scel</th>
</tr>
</thead>
<tbody>
<tr>
<td>A7LCP1</td>
<td>S. cerevisiae</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>Q36760</td>
<td>S. cerevisiae</td>
<td>56</td>
<td>98</td>
</tr>
<tr>
<td>O63264</td>
<td>Z. bisporus</td>
<td>57</td>
<td>72</td>
</tr>
<tr>
<td>Q34839</td>
<td>K. thermotolerans</td>
<td>58</td>
<td>71</td>
</tr>
<tr>
<td>Q34807</td>
<td>P. canadensis</td>
<td>59</td>
<td>58</td>
</tr>
</tbody>
</table>

Table 2: Exemplary homologs of l-Crel, which can be cloned from other organisms.

<table>
<thead>
<tr>
<th>Uni-Prot Accession Nr.</th>
<th>Organism</th>
<th>SEQ ID NO:</th>
<th>Amino Acid Sequence Identity to l-Crel</th>
</tr>
</thead>
<tbody>
<tr>
<td>P05725</td>
<td>C. reinhardtii</td>
<td>60</td>
<td>100</td>
</tr>
<tr>
<td>Q8SMM1</td>
<td>C. lunzensis</td>
<td>61</td>
<td>56</td>
</tr>
<tr>
<td>Q8SML7</td>
<td>C. olivieri</td>
<td>62</td>
<td>58</td>
</tr>
<tr>
<td>Q1KVQ8</td>
<td>S. obliquus</td>
<td>63</td>
<td>49</td>
</tr>
</tbody>
</table>

Table 3: Exemplary homologs of Pl-Scel, which can be cloned from other organisms.

<table>
<thead>
<tr>
<th>Uni-Prot Accession Nr.</th>
<th>Organism</th>
<th>SEQ ID NO:</th>
<th>Amino Acid Sequence Identity to Pl-Scel</th>
</tr>
</thead>
<tbody>
<tr>
<td>P17255</td>
<td>S. cerevisiae</td>
<td>79</td>
<td>100</td>
</tr>
<tr>
<td>Q874G9</td>
<td>S. cerevisiae</td>
<td>80</td>
<td>99</td>
</tr>
<tr>
<td>Q874F9</td>
<td>S. pastorianus</td>
<td>81</td>
<td>97</td>
</tr>
</tbody>
</table>
Table 4: Exemplary homologs of l-Ceul, which can be cloned from other organisms.

<table>
<thead>
<tr>
<th>Uni-Prot Accession Nr.</th>
<th>Organism</th>
<th>SEQ ID NO:</th>
<th>Amino Acid Sequence Identity to l-Ceul</th>
</tr>
</thead>
<tbody>
<tr>
<td>P32761</td>
<td>C. moewusii</td>
<td>65</td>
<td>100%</td>
</tr>
<tr>
<td>Q8WKZ1</td>
<td>C. echinozygotum</td>
<td>66</td>
<td>63%</td>
</tr>
<tr>
<td>Q8WL12</td>
<td>C. elongatum</td>
<td>67</td>
<td>58%</td>
</tr>
<tr>
<td>Q8WL11</td>
<td>A. stipitatus</td>
<td>68</td>
<td>55%</td>
</tr>
<tr>
<td>Q8WKX7</td>
<td>C. monadina</td>
<td>69</td>
<td>51%</td>
</tr>
</tbody>
</table>

Table 5: Exemplary homologs of l-Chul, which can be cloned from other organisms.

<table>
<thead>
<tr>
<th>Uni-Prot Accession Nr.</th>
<th>Organism</th>
<th>SEQ ID NO:</th>
<th>Amino Acid Sequence Identity to l-Ceul</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q53X18</td>
<td>C. humicola</td>
<td>70</td>
<td>100%</td>
</tr>
<tr>
<td>Q8WL03</td>
<td>C. zebra</td>
<td>71</td>
<td>67%</td>
</tr>
<tr>
<td>Q8WKX6</td>
<td>C. monadina</td>
<td>72</td>
<td>62%</td>
</tr>
<tr>
<td>Q8WL10</td>
<td>A. stipitatus</td>
<td>73</td>
<td>58%</td>
</tr>
<tr>
<td>Q8SM16</td>
<td>N. aquatica</td>
<td>74</td>
<td>54%</td>
</tr>
</tbody>
</table>

Table 6: Exemplary homologs of l-Dmol, which can be cloned from other organisms.

<table>
<thead>
<tr>
<th>Uni-Prot Accession Nr.</th>
<th>Organism</th>
<th>SEQ ID NO:</th>
<th>Amino Acid Sequence Identity to l-Ceul</th>
</tr>
</thead>
<tbody>
<tr>
<td>P21505</td>
<td>D. mobilis</td>
<td>75</td>
<td>100%</td>
</tr>
<tr>
<td>Q6L6Z4</td>
<td>Thermoproteus sp.</td>
<td>76</td>
<td>51%</td>
</tr>
<tr>
<td>Q6L6Z5</td>
<td>Thermoproteus sp.</td>
<td>77</td>
<td>50%</td>
</tr>
<tr>
<td>A3MXB6</td>
<td>P. calidiformis</td>
<td>78</td>
<td>49%</td>
</tr>
</tbody>
</table>

Homologs of endonucleases, which are cloned from other organisms might have a different enzymatic activity, DNA-binding-affinity, dimer formation affinity or changes in its DNA recognition sequence, when compared to the reference endonucleases, like l-Scel for homologs described in Table 1, l-Crel for homologs described in Table 2, or PI-Scel for homologs described in Table 3, or l-Ceul for homologs described in Table 4, or l-Chul for homologs described in Table 5, or l-Dmol for homologs described in Table 6.

Preferred are LAGLIDADG endonucleases for which exact protein crystal structures have been determined, like I-Dmo I, H-Dre I, I-Sce I, I-Cre I, homologs of any one these having at least
Another way to create homologs of LAGLIDAG endonucleases is to mutate the amino acid sequence of an LAGLIDAG endonuclease in order to modify its DNA binding affinity, its dimer formation affinity or to change its DNA recognition sequence. The determination of protein structure as well as sequence alignments of homologs of LAGLIDAG endonucleases allows for rational choices concerning the amino acids, that can be changed to affect its enzymatic activity, its DNA-binding-affinity, its dimer formation affinity or to change its DNA recognition sequence.

Homologs of LAGLIDAG endonucleases, which have been mutated in order to modify their DNA binding affinity, its dimer formation affinity or to change its DNA recognition site are called engineered endonucleases.

One approach to create engineered endonucleases is to employ molecular evolution. Polynucleotides encoding a candidate endonuclease enzyme can, for example, be modulated with DNA shuffling protocols. DNA shuffling is a process of recursive recombination and mutation, performed by random fragmentation of a pool of related genes, followed by reassembly of the fragments by a polymerase chain reaction-like process. See, e.g., Stemmer (1994) Proc Natl Acad Sci USA 91:10747-10751; Stemmer (1994) Nature 370:389-391; and US 5,605,793, US 5,837,458, US 5,830,721 and US 5,811,238. Engineered endonucleases can also be created by using rational design, based on further knowledge of the crystal structure of a given endonuclease see for example Fajardo-Sanchez et al., "Computer design of obligate heterodimer meganucleases allows efficient cutting of custom DNA sequences", Nucleic Acids Research, 2008, Vol. 36, No. 7 2163-2173.


Engineered versions of l-Scel, l-Crel, l-Msol and l-Ceu1 having an increased or decreased DNA-binding affinity are for example disclosed in WO07/047859 and WO09/076292.

If not explicitly mentioned otherwise, all mutants will be named according to the amino acid numbers of the wildtype amino acid sequences of the respective endonuclease, e.g. the mutant L19 of l-Scel will have an amino acid exchange of leucine at position 19 of the wildtype l-Scel amino acid sequence, as described by SEQ ID NO: 1. The L19H mutant of l-Scel, will have a
replacement of the amino acid leucine at position 19 of the wildtype l-Scel amino acid sequence
with histidine.

For example, the DNA-binding affinity of l-Scel can be increased by at least one modification
 corresponding to a substitution selected from the group consisting of:
(a) substitution of D201, L19, L80, L92, Y151, Y188, 1191, Y199 or Y222 with H, N, Q, S, T, K or
 R; or
(b) substitution of N15, N17, S81, H84, N94, N120, T156, N157, S159, N163, Q165, S166, N194 or
 S202 with K or R.

DNA-binding affinity of l-Scel can be decreased by at least one mutation corresponding to a
 substitution selected from the group consisting of:
(a) substitution of K20, K23, K63, K122, K148, K153, K190, K193, K195 or K223 with H, N, Q,
 S, T, D or E; or
(b) substitution of L19, L80, L92, Y151, Y188, 1191, Y199, Y222, N15, N17, S81, H84, N94,
 N120, T156, N157, S159, N163, Q165, S166, N194 or S202 with D or E.

Engineered versions of l-Scel, l-Crel, l-Msol and l-CEul having a changed DNA recognition se-
 quence are disclosed in WO07/047859 and WO09/076292.

For example, an important DNA recognition site of l-Scel has the following sequence:

<table>
<thead>
<tr>
<th>Sense</th>
<th>5' - T T A C C C T G T T A T C C C T A G - 3'</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18</td>
</tr>
<tr>
<td>Antisense</td>
<td>3' - A A T G G G A C A A T A G G G A T C 5'</td>
</tr>
</tbody>
</table>

The following mutations of l-Scel will change the preference for C at position 4 to A: K50
The following mutations of l-Scel will keep the preference for C at position 4: K50, CE57
The following mutations of l-Scel will change the preference for C at position 4 to G: E50, R57,
 K57.

The following mutations of l-Scel will change the preference for C at position 4 to T: K57, M57,
 Q50.

The following mutations of l-Scel will change the preference for C at position 5 to A: K48, Q102.
The following mutations of l-Scel will keep the preference for C at position 5: R48, K48, E102,
 E59
The following mutations of l-Scel will change the preference for C at position 5 to G: E48, K102,
 R102.
The following mutations of l-Scel will change the preference for C at position 5 to T: Q48, C102,
 L102, V102.

The following mutations of l-Scel will change the preference for C at position 6 to A: K59.
The following mutations of l-Scel will keep the preference for C at position 6: R59, K59.
The following mutations of l-Scel will change the preference for C at position 6 to G: K84, E59.

The following mutations of l-Scel will change the preference for C at position 6 to T: Q59, Y46.

The following mutations of l-Scel will change the preference for T at position 7 to A: C46, L46, V46.

The following mutations of l-Scel will change the preference for T at position 7 to C: R46, K46, E86.

The following mutations of l-Scel will change the preference for T at position 7 to G: K86, R86, E46.

The following mutations of l-Scel will keep the preference for T at position 7: K68, C86, L86, Q46.

The following mutations of l-Scel will change the preference for G at position 8 to A: K61, S61, V61, A61, L61.

The following mutations of l-Scel will change the preference for G at position 8: E88, R61, H61.

The following mutations of l-Scel will keep the preference for G at position 8: E61, R88, K88.

The following mutations of l-Scel will change the preference for G at position 8 to T: K88, Q61, H61.

The following mutations of l-Scel will change the preference for T at position 9 to A: T98, C98, V98, L9B.

The following mutations of l-Scel will change the preference for T at position 9 to C: R98, K98.

The following mutations of l-Scel will change the preference for T at position 9 to G: E98, D98.

The following mutations of l-Scel will keep the preference for T at position 9: Q98.

The following mutations of l-Scel will change the preference for T at position 10 to A: V96, C96, A96.

The following mutations of l-Scel will change the preference for T at position 10 to C: K96, R96.

The following mutations of l-Scel will change the preference for T at position 10 to G: D96, E96.

The following mutations of l-Scel will keep the preference for T at position 10: Q96.

The following mutations of l-Scel will keep the preference for A at position 11: C90, L90.

The following mutations of l-Scel will change the preference for A at position 11 to C: K90, R90.

The following mutations of l-Scel will change the preference for A at position 11 to G: E90.

The following mutations of l-Scel will change the preference for A at position 11 to T: Q90.

The following mutations of l-Scel will change the preference for T at position 12 to A: Q193.

The following mutations of l-Scel will change the preference for T at position 12 to C: E165, E193, D193.

The following mutations of l-Scel will change the preference for T at position 12 to G: K165, R165.

The following mutations of l-Scel will keep the preference for T at position 12: C165, L165,

The following mutations of l-Scel will change the preference for C at position 13 to A: C193, L193.

The following mutations of l-Scel will change the preference for G at position 18 to T: H155, Y155.

The following mutations of l-Scel will keep the preference for C at position 13: K193, R193, D192.
The following mutations of l-Scel will change the preference for C at position 13 to G: E193, D193, K163, R192.
The following mutations of l-Scel will change the preference for C at position 13 to T: Q193, C163, L163.

The following mutations of l-Scel will change the preference for C at position 14 to A: L192, C192.
The following mutations of l-Scel will keep the preference for C at position 14: E161, R192, K192.
The following mutations of l-Scel will change the preference for C at position 14 to G: K147, K161, R161, R197, D192, E192.
The following mutations of l-Scel will change the preference for C at position 14 to T: K161, Q192.

The following mutations of l-Scel will change the preference for C at position 15 to A: none identified.
The following mutations of l-Scel will keep the preference for C at position 15: E151.
The following mutations of l-Scel will change the preference for C at position 15 to G: K151.
The following mutations of l-Scel will change the preference for C at position 15 to T: C151, L151, K151.

The following mutations of l-Scel will keep the preference for A at position 17: N152, S152, C150, L150, V150, T150.
The following mutations of l-Scel will change the preference for A at position 17 to C: K152, K150.
The following mutations of l-Scel will change the preference for A at position 17 to G: N152, S152, D152, D150, E150.
The following mutations of l-Scel will change the preference for A at position 17 to T: Q152, Q150.

The following mutations of l-Scel will change the preference for G at position 18 to A: K155, C155.
The following mutations of l-Scel will change the preference for G at position 18: R155, K155.
The following mutations of l-Scel will keep the preference for G at position 18: E155.
The following mutations of l-Scel will change the preference for G at position 18 to T: H155, Y155.
Combinations of several mutations may enhance the effect. One example is the triple mutant W149G, D150C and N152K, which will change the preference of l-Scel for A at position 17 to G.

In order to preserve the enzymatic activity of the LAGLIDAG endonucleases the following mutations should be avoided:


for l-Crel: Q47E,

for l-Ceul E66Q,

for l-Msol D22N,

for Pl-Scel mutations in D218, D229, D326 or T341.

Engineered endonuclease variants of l-Ani, having high enzymatic activity can be found in Takeuchi et al., Nucleic Acid Res. (2009), 73(3): 877 to 890. Preferred engineered endonuclease variants of l-Ani, as described by SEQ ID NO: 142, comprise the following mutations: F13Y and S111Y, or F13Y, S111Y and K222R, or F13Y, I55V, F91I, S92T and S111Y.

Mutations which alter the DNA-binding-affinity, the dimer formation affinity or change the DNA recognition sequence of a given endonuclease, e.g. a LAGLIDAG endonuclease, may be combined to create an engineered endonuclease, e.g. an engineered endonuclease based on l-Scel and having an altered DNA-binding-affinity and/or a changed DNA recognition sequence, when compared to l-Scel as described by SEQ ID NO: 1.

Optimized nucleases:
Nucleases can be optimized for example by inserting mutations to change their DNA binding specificity, e.g. to make their DNA recognition site more or less specific, or by adapting the polynucleotide sequence coding for the nuclease to the codon usage of the organism, in which the endonuclease is intended to be expressed, or by deleting alternative start codons, or by deleting cryptic polyadenylation signals from the polynucleotide sequence coding for the endonuclease.

Mutations and changes in order to create optimized nucleases may be combined with the mutations used to create engineered endonucleases, for example, a homologue of l-Scel may be an optimized nuclease as described herein, but may also comprise mutations used to alter its DNA-binding-affinity and/or change its DNA recognition sequence.

Further optimization of nucleases may enhance protein stability. Accordingly optimized nucleases do not comprise, or have a reduced number compared to the amino acid sequence of the non optimized nuclease of:

a) PEST-Sequences,
b) KEN-boxes
c) A-boxes,
d) D-boxes, or
e) comprise an optimized N-terminal end for stability according to the N-end rule,
f) comprise a glycine as the second N-terminal amino acid, or
g) any combination of a), b), c), d), e) and f).

PEST Sequences are required to contain at least one proline (P), one aspartate (D) or glutamate (E) and at least one serine (S) or threonine(T). Negatively charged amino acids are clustered within these motifs while positively charged amino acids, arginine (R), histidine (H) and lysine (K) are generally forbidden. PEST Sequences are for example described in Rechsteiner M, Rogers SW. “PEST sequences and regulation by proteolysis.” Trends Biochem. Sci. 1996; 21(7), pages 267 to271.

The amino acid consensus sequence of a KEN-box is: KENXXX(N/D)
The amino acid consensus sequence of a A-box is: AQRXLXXSXXXQRVL
The amino acid consensus sequence of a D-box is: RXXL

A further way to stabilize nucleases against degradation is to optimize the amino acid sequence of the N-terminus of the respective endonuclease according to the N-end rule. Nucleases which are optimized for the expression in eucaryotes comprise either methionine, valine, glycine, threonine, serine, alanine or cysteine after the start methionine of their amino acid sequence. Nucleases which are optimized for the expression in procaryotes comprise either methionine, valine, glycine, threonine, serine, alanine, cysteine, glutamic acid, glutamine, aspartic acid, asparagine, isoleucine or histidine after the start methionine of their amino acid sequence.

Nucleases may further be optimized by deleting 50, 40, 30, 20, 10, 9, 8, 7, 6, 5, 4, 3, 2, or 1 amino acids of its amino acid sequence, without destroying its endonuclease activity. For example, in case parts of the amino acid sequence of a LAGLIDADG endonuclease is deleted, it is important to retain the LAGLIDADG endonuclease motif described above.

It is preferred to delete PEST sequences or other destabilizing motifs like KEN-box, D-box and A-box. Those motifs can also be destroyed by introduction of single amino acid exchanges, e.g. introduction of a positively charged aminoacid (arginine, histidine and lysine) into the PEST sequence.

Another way to optimize nucleases is to add nuclear localization signals to the amino acid sequence of the nuclease. For example a nuclear localization signal as described by SEQ ID NO: 4.

Optimized nucleases may comprise a combination of the methods and features described above, e.g. they may comprise a nuclear localization signal, comprise a glycine as the second N-terminal amino acid or a deletion at the C-terminus or a combination of these features. Examples of optimized nucleases having a combination of the methods and features described above are for example described by SEQ ID NOs: 2, 3 and 5.
In one embodiment the optimized nuclease is an optimized l-Sce-l, which does not comprise an
amino acid sequence described by the sequence: HVCLLYDQWVLSPPH, LAYWFMDGGGK, KTI
PNVLVENYTPMSLASYWFMDGGK, KPIYIDSMSYLIFYNLK, KLPNTISSETFLK or TISSETFLK,
or which does not comprise an amino acid sequence described by the sequence: HVCLLY
DQWVLSPPH, LAYWFMDGGGK, KPIYIDSMSYLIFYNLK, KLPNTISSETFLK or TISSETFLK,
or which does not comprise an amino acid sequence described by the sequence: LAYWFMD
GGGK, KLPNTISSETFLK or TISSETFLK,
or which does not comprise an amino acid sequence described by the sequence: KLPNTIS
SETFLK or TISSETFLK,

In one embodiment the optimized nuclease is I-SceI, or its homologs having at least 49%, 51%,
58%, 60%, 70%, 80%, 85%, 90%, 92%, 93%, 94%, 95%, 96%, 97%, 98% or 99% sequence
identity on amino acid level in which the amino acid sequence TISSETFLK at the C-terminus of
wildtype I-SceI or its homologs having at least 49%, 51%, 58%, 60%, 70%, 80%, 85%, 90%,
92%, 93%, 94%, 95%, 96%, 97%, 98% or 99% sequence identity on amino acid level and hav
ing an amino acid sequence TISSETFLK at the C-terminus, is deleted or mutated.

The amino acid sequence TISSETFLK may be deleted or mutated, by deleting or mutating at
least 1, 2, 3, 4, 5, 6, 7, 8 or 9 amino acids of the C-terminus of wildtype I-SceI or its homologs
having at least 49%, 51%, 58%, 60%, 70%, 80%, 85%, 90%, 92%, 93%, 94%, 95%, 96%, 97%,
98% or 99% sequence identity on amino acid level and having an amino acid sequence TIS
SETFLK at the C-terminus.

Table 7: Different examples for deletions of the TISSETFLK amino acid sequence in wildtype I-
SceI

<table>
<thead>
<tr>
<th>Wildtype and optimized I-Sce</th>
<th>Amino Acid Sequence on C-terminus</th>
</tr>
</thead>
<tbody>
<tr>
<td>I-Sce I wildtype</td>
<td>TISSETFLK</td>
</tr>
<tr>
<td>I-Sce I - 1</td>
<td>TISSETFL</td>
</tr>
<tr>
<td>I-Sce I - 2</td>
<td>TISSETF</td>
</tr>
<tr>
<td>I-Sce I - 3</td>
<td>TISSET</td>
</tr>
<tr>
<td>I-Sce I - 4</td>
<td>TISSE</td>
</tr>
<tr>
<td>I-Sce I - 5</td>
<td>TISS</td>
</tr>
<tr>
<td>I-Sce I - 6</td>
<td>TIS</td>
</tr>
<tr>
<td>I-Sce I - 7</td>
<td>TI</td>
</tr>
<tr>
<td>I-Sce I - 8</td>
<td>T</td>
</tr>
<tr>
<td>I-Sce I - 9</td>
<td>all 9 amino acids on C-terminus of wt I-Sce deleted</td>
</tr>
</tbody>
</table>
Alternatively the amino acid sequence TISSETFLK may be mutated, e.g. to the amino acid sequence: TIKSETFLK (SEQ ID NO: 149), or AIANQAFLK (SEQ ID NO: 150).

Equally preferred, is to mutate serine at position 229 of the amino acid sequence of wildtype l-Scel as disclosed in SEQ ID No. 1 (being amino acid 230 if referenced to SEQ ID No. 2) to Lys, Ala, Pro, Gly, Glu, Gin, Asp, Asn, Cys, Tyr or Thr. Thereby creating the l-Scel mutants S229K, S229A, S229P, S229G, S229E, S229Q, S229D, S229N, S229C, S229Y, or S229T (amino acids are numbered according to SEQ ID No. 1).

In another embodiment of the invention, the amino acid methionine at position 203 of the amino acid sequence of wildtype l-Scel as disclosed in SEQ ID No. 1 (being amino acid 204 if referenced to SEQ ID No. 2), is mutated to Lys, His or Arg. Thereby creating the l-Scel mutant M203K, M203H and M203R.

Preferred optimized versions of l-Scel are the deletions l-Scel -1, l-Scel -2, l-Scel -3, l-Scel -4, l-Scel -5, l-Scel -6, l-Scel -7, l-Scel -8, l-Scel -9 and the mutants S229K and S229H, S229R even more preferred are the deletions l-Scel -1, l-Scel -2, l-Scel -3, l-Scel -4, l-Scel -5, l-Scel -6 and the mutant S229K.

It is also possible to combine the deletions and mutations described above, e.g. by combining the deletion l-Scel -1 with the mutant S229K, thereby creating the amino acid sequence TIKSETFLK at the C-terminus.

It is also possible to combine the deletions and mutations described above, e.g. by combining the deletion l-Scel -1 with the mutant S229A, thereby creating the amino acid sequence TIASETFLK at the C-terminus.

Further preferred optimized versions of l-Scel are the deletions l-Scel -1, l-Scel -2, l-Scel -3, l-Scel -4, l-Scel -5, l-Scel -6, l-Scel -7, l-Scel -8, l-Scel -9 or the mutants S229K and S229H, S229R, in combination with the mutation M203K, M203H, M203R.

Even more preferred are the deletions l-Scel -1, l-Scel -2, l-Scel -3, l-Scel -4, l-Scel -5, l-Scel -6 or the mutant S229K in combination with the mutation M203K.

In another embodiment of the invention, the amino acids glutamine at position 75, glutamic acid at position 130, or tyrosine at position 199 of the amino acid sequence of wildtype l-Scel as disclosed in SEQ ID No. 1 (being amino acids 76, 131 and 120 if referenced to SEQ ID No. 2), are mutated to Lys, His or Arg. Thereby creating the l-Scel mutants Q75K, Q75H, Q75R, E130K, E130H, E130R, Y199K, Y199H and Y199R.

The deletions and mutations described above will also be applicable to its homologs of l-Scel having at least 49%, 51%, 58%, 60%, 70%, 80%, 85%, 90%, 92%, 93%, 94%, 95%, 96%, 97%, 98% or 99% sequence identity on amino acid level and having an amino acid sequence TIS-
SETFLK at the C-terminus.

Accordingly, in one embodiment of the invention, the optimized endonuclease, is an optimized version of I-Scl or one of its homologs having at least 49%, 51%, 58%, 60%, 70%, 80%, 85%, 90%, 92%, 93%, 94%, 95%, 96%, 97%, 98% or 99% sequence identity on amino acid level, and having one or more of the mutations or deletions selected from the group of: I-Scl -1, I-Scl -2, I-Scl -3, I-Scl -4, I-Scl -5, I-Scl -6, I-Scl -7, I-Scl -8, I-Scl -9, S229K, S229A, S229P, S229G, S229E, S229Q, S229D, S229N, S229C, S229Y, S229T, M203K, M203H, M203R, Q77K, Q77H, Q77R, E130K, E130H, E130R, Y199K, Y199H and Y199R, wherein the amino acid numbers are referenced to the amino acid sequence as described by SEQ ID NO: 1.

In a further embodiment of the invention, the optimized endonuclease, is an optimized version of I-Scl or one of its homologs having at least 49%, 51%, 58%, 60%, 70%, 80%, 85%, 90%, 92%, 93%, 94%, 95%, 96%, 97%, 98% or 99% sequence identity on amino acid level, and having one or more of the mutations or deletions selected from the group of: I-Scl -1, I-Scl -2, I-Scl -3, I-Scl -4, I-Scl -5, I-Scl -6, S229K and M203K, wherein the amino acid numbers are referenced to the amino acid sequence as described by SEQ ID NO: 1.

A particular preferred optimized endonuclease is a wildtype or engineered version of I-Scl, as described by SEQ ID NO: 1 or one of its homologs having at least 49%, 51%, 58%, 60%, 70%, 80%, 85%, 90%, 92%, 93%, 94%, 95%, 96%, 97%, 98% or 99% sequence identity on amino acid level and having one or more mutations selected from the groups of:

- a) I-Scl -1, I-Scl -2, I-Scl -3, I-Scl -4, I-Scl -5, I-Scl -6, I-Scl -7, I-Scl -8 and I-Scl -9;
- c) a methionine, valine, glycine, threonine, serine, alanine, cysteine, glutamic acid, glutamine, aspartic acid, asparagine, isoleucine or histidine after the start methionine of their amino acid sequence; or
- d) a combination of one or more mutations selected from a) and b), a) and c), b) and c) or a) b) and c) above.

Heterologous DNA binding domains:
The chimeric endonuclease of the invention comprises at least one heterologous DNA binding domain.

Heterologous DNA binding domains are polypeptides binding to polynucleotides having a specific polynucleotide sequence (recognition sequence or operator sequence). Examples for heterologous DNA binding domains are eukaryotic, prokaryotic or viral transcription factors. In one embodiment of the invention, only the DNA binding domain of the eukaryotic, prokaryotic or viral transcription factor is used as heterologous DNA binding domain.

Preferably heterologous DNA binding domains are selected from eukaryotic, prokaryotic and
viral transcription factors or their respective DNA binding domains, which bind DNA as monomers or single chain variants, which bind their DNA recognition sequence with high affinity and specificity, and have an N- or C-Terminus on the surface of the protein.

Especially preferred are eukaryotic, prokaryotic and viral transcription factors or their respective DNA binding domains of which the three dimensional structure of at least a homolog of the respective eukaryotic, prokaryotic and viral transcription factors or their respective DNA binding domain has been determined.

The term heterologous DNA binding domain shall not comprise more than two repetitions of modular $C_2H_2$ zink finger domains, as disclosed for example in WO07/014275, WO08/076290, WO08/076290 or WO03/062455. $C_2H_2$ Zinc finger domains have conserved cysteine and histidine residues that tetrahedrally-coordinate the single zinc atom in each finger domain and are characterized by finger components having the general sequence: -Cys-(X$_{1,4}$)-Cys-(X$_{12}$)-His-(X$_{13,5}$)-His- in which X represents any amino acid. (the $C_2H_2$ ZFPs).

Numerous eukaryotic, prokaryotic and viral transcription factors as well as their respective recognition sequences or operator sequences have been described in the art. Information on eukaryotic, prokaryotic and viral transcription factors as well as their respective recognition sequences as well as numerous three dimensional structures can be found in public available databases and bioinformatic analysis tools, for example in:

JASPAR 2010 (Partales-Casamar et al. (2009), Nucl. Acids Res., 1 to 6),
PLACE (Higo et al. (1999), Nucl. Acids Res., 27 (1), 297 to 300).

RegulonDB (Gama-Castro, S., et al. (2008) Nucleic acids research 36, D120 to 124)
FlyReg (Bergman, CM., et al. (2005) Bioinformatics 21, 1747 to 1749)

Zhu, C., et al. (2009), Genome Res 19, 556 to 566


Maclsaac, K.D., et al. (2006) BMC bioinformatics 7, 113


Preferred heterologous DNA binding domains are proteins with known binding properties and recognition sequences; more preferable proteins which have been co-crystalized with their specific DNA target.

Eukaryotic, prokaryotic and viral transcription factors have been grouped in several protein families, having an individual PF-Number as identifier.

Heterologous DNA-binding domains can for example be found in the following protein families:
PF00126  Bacterial regulatory helix-turn-helix protein, lysR family
PF00486  Transcriptional regulatory protein, C terminal
PF04383  KilA-N domain
PF01381  Helix-turn-helix
PF02954  Bacterial regulatory protein, Fis family
PF00313  Cold-shock DNA-binding domain
PF00325  Bacterial regulatory proteins, crp family
PF01047  MarR family
PF00392  Bacterial regulatory proteins, gntR family
PF00165  Bacterial regulatory helix-turn-helix proteins, AraC family
PF05225  helix-turn-helix, Psq domain
PF00847  AP2 domain
PF04967  HTH DNA binding domain
PF00196  Bacterial regulatory proteins, luxR family
PF00010  Helix-loop-helix DNA-binding domain
PF00356  Bacterial regulatory proteins, lacl family
PF008279 HTH domain
PF01022  Bacterial regulatory protein, arsR family
PF00196  Bacterial regulatory proteins, luxR family
PF00010  Helix-loop-helix DNA-binding domain
PF00356  Bacterial regulatory proteins, lacl family
PF02082  Transcriptional regulator
PF00292  Paired box domain
PF04397  LytTr DNA-binding domain
PF03749  Sugar fermentation stimulation protein
PF04353  Regulator of RNA polymerase sigma70 subunit, Rsd/AlgQ

Preferably heterologous DNA binding domains are selected from members of the following protein families:

PF00126  Bacterial regulatory helix-turn-helix protein, lysR family
PF00165  Bacterial regulatory helix-turn-helix proteins, AraC family
PF01022  Bacterial regulatory protein, arsR family
PF00196  Bacterial regulatory proteins, luxR family
PF00010  Helix-loop-helix DNA-binding domain
PF00356  Bacterial regulatory proteins, lacl family

Even more preferred are members of the following protein families:

PF00126  Bacterial regulatory helix-turn-helix protein, lysR family
PF00165  Bacterial regulatory helix-turn-helix proteins, AraC family
PF00196  Bacterial regulatory proteins, luxR family
PF00356  Bacterial regulatory proteins, lacl family

A particular preferred group of heterologous DNA binding domains are proteins comprising a helix-turn-helix DNA binding domain (HTH domain). Such proteins are for example scTetR,
ArcR and proteins of the Lacl, AraC and MerR protein families.


Examples and common features of proteins belonging to the TetR protein family are given by SEQ ID NO: 86, 87, 88, 89 and 90 and the alignment shown in Figure 8, Examples and common features of the respective HTH domains are given by SEQ ID NO: 91, 92, 93, 94 and 95 and the alignment shown in Figure: 9a.


Examples and common features of the HTH domains of proteins belonging to the Lac Repressor protein family are given by SEQ ID NO: 101, 102, 103, 104 and 105 and the alignment shown in Figure: 10a.


Examples and common features of proteins belonging to the AraC protein family in particular homologs of MarA are given by SEQ ID NO: 120, 121, 122, 123, 124, 125, 126 and 127 and the alignment shown in Figure: 12. Examples and common features of the HTH domains of proteins belonging to the AraC protein protein family in particular homologs of MarA are given by SEQ ID NO: 112, 113, 114, 115, 116, 117, 118 and 119 and the alignment shown in Figure: 11.
Information about the MerR protein family and common features of their HTH domain can be found in: Brown N.L. et al. "The MerR family of transcriptional regulators" FEMS Microbiology Reviews (2003), Vol. 27, pages 145 to 163. Examples and common features of the HTH domains of proteins belonging to the MerR protein protein family are given by SEQ ID NO: 106, 107, 108, 109, 110 and 111 and the alignment shown in Figure: 10b.

Proteins similar to the scArcR protein as described by SEQ ID NO: 7 comprise a HTH domain for DNA binding, different examples and common features of these HTH domains are given by SEQ ID NO: 96, 97, 98, 99 and 100 and the alignment shown in Figure: 9b.

Members of the WRKY protein family and information about common features of these proteins are for example described in: Eulgem,T. et al. "The WRKY superfamily of plant transcription factors." (2000) Trends Plant Sci., 5, pages 199 to 206 and Ming-Rui Duan et al. "DNA binding mechanism revealed by high resolution crystal structure of Arabidopsis thaliana WRKY1 protein" (2007), Nucleic Acids Research, Vol. 35, No. 4 1145-1 154, which are included herein by reference in their entirety.

Other suitable heterologous DNA binding domains are inactive endonucleases. Such endonucleases may be inactive in the target organism because they act only under certain, usually more extreme conditions (for example, high temperature). Alternatively, one may use a mutated endonuclease, whereas said mutation renders the endonuclease inactive. Inactive endonucleases are for example, but not excluding others: l-Dmol or other termophyllic endonucleases employed at temperatures below 40°C, more preferable below 30°C, even more preferably below 25°C, and endonucleases having amino acid substitutions in their active center(s), for example l-Crel having the mutation of Q47 to E, l-Sce I having the mutation of D44 or D145 to N, l-CEul having the mutation of E66 to Q, or l-Msol having the mutation of D22 to N. A preferred inactive endonuclease is l-Sce I having the mutation of D44 to S (l-Scel D44S). For example the following amino acid residues of Pi-Scel: D218, D229, D326 and T341 Pingoud (2000) Biochemistry 39:15895-15900

In one embodiment at least one heterologous DNA binding domain is an inactive l-Scel, l-Crel, l-Ceul, l-Chul, l-Dmol, Pi-Scel, l-Msol, or l-Anil or an inactive homolog of these having at least 45%, 56%, 57%, 58%, 59%, 60%, 61%, 62%, 63%, 65%, 66%, 67%, 68%, 69%, 70%, 71%, 72%, 73%, 74%, 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99% amino acid sequence identity. In one embodiment the heterologous DNA binding domain is an inactive version of a LAGLIDADG endonucleases having an amino acid sequence as described by at least one of SEQ ID NO: 1, 2, 3, 5, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 142 or 159, preferably having an amino acid sequence as described by any one of SEQ ID NO: 1, 2, 3, 5 or 159.

In one preferred embodiment the chimeric endonuclease comprises l-Scel or an optimized version of l-Scel and an heterologous DNA binding domain comprising an inactive l-Scel or an in-
active version of an optimized version of l-Scel.

In one embodiment of the invention the term heterologous DNA binding domain does not comprise inactive endonucleases.

The heterologous DNA binding domain can comprise the full protein of a given transcription factor or a large fragment thereof or might only comprise a fragment more or less limited to the DNA binding domain of a transcription factor.

Examples for suitable transcription factors are for example, but not excluding others: scTet, scArcR, LacR, TraR, Gal, LambdaR, LuxR, WRKY and homologs of any one these having at least 50%, 60%, 70%, 80%, 85%, 90%, 92%, 93%, 94%, 95%, 96%, 97%, 98% or 99% of sequence identity on amino acid level.

In a preferred embodiment the DNA binding activity of the heterologous DNA binding domain is inducible or repressible via binding of an Inductor to at least one of the DNA binding domains. The Inductor can be a polypeptide or a small organic substance.

Examples for inducible or repressible or inducible and repressible heterologous DNA binding domains and their inducers or repressors are:

<table>
<thead>
<tr>
<th>Domain</th>
<th>Inducer or repressors</th>
</tr>
</thead>
<tbody>
<tr>
<td>scTet</td>
<td>Tetracyline and Anhydrotetracyline and other derivates</td>
</tr>
<tr>
<td>LacR</td>
<td>Lactose and IPTG</td>
</tr>
<tr>
<td>TraR</td>
<td>30C8HL (N-(3-oxo)-octanoly-L-homoserine lactone)</td>
</tr>
<tr>
<td>LuxR</td>
<td>LuxR family . acetylated homoserine lactones (AHL)</td>
</tr>
<tr>
<td>LuxR</td>
<td>30C6HL (N-(3-oxo)-hexa-L-homoserine lactone)</td>
</tr>
<tr>
<td>LasR</td>
<td>30C12HL (N-(3-oxo)-duodeca-L-homoserine lactone)</td>
</tr>
<tr>
<td>AraC</td>
<td>Arabinose</td>
</tr>
<tr>
<td>RhaR</td>
<td>Rhamnose</td>
</tr>
<tr>
<td>MerR</td>
<td>mercury ions</td>
</tr>
</tbody>
</table>

Preferably the heterologous DNA binding domain has a recognition sequence of at least 4, at least 6, at least 8, at least 10 or at least 12 base pairs.

Examples of recognition sequences of heterologous DNA binding domains are:

<table>
<thead>
<tr>
<th>Domain</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>scTet</td>
<td>5'-YTATCATTGATAG-3' (SEQ ID NO: 130)</td>
</tr>
<tr>
<td>TetR</td>
<td>5'-YTATC -3' (only one monomer)</td>
</tr>
<tr>
<td>scArcR</td>
<td>5'-AATGATAGAAGCACTCTACTAT-3' (SEQ ID NO: 7)</td>
</tr>
<tr>
<td>TraR</td>
<td>5'-ATGTG CAGATCTG CACAT-3' (SEQ ID NO: 131)</td>
</tr>
<tr>
<td>WRKY</td>
<td>5'-ATGTG CAGATCTG CACAT-3'</td>
</tr>
</tbody>
</table>
5'-YTGACY-3'
LacR (dimer or single chain variants)
5'-TTGTGAGC-3'
MarA (monomer)
5'-AYNGCACNNWNNRYYAAAYN-3'
(SEQ ID NO: 137)
5'-TTKACY-3',
MerR (monomer)
5'-TTKACYNNNNNNNNNNNNNTAAGGT-3'
(SEQ ID NO: 138)
wherein A stands for adenine, G for guanine, C for cytosine, T for thymine, R for guanine or adenine, Y for thymine or cytosine, K for guanine or thymine, W for adenine or thymine and n for adenine or guanine or cytosine or thymine.

The person skilled in the art will acknowledge that most DNA binding domains will not be limited to bind only the exact recognition sequence, but also similar recognition sequences for example.
Examples for alternative recognition sequences of LacR dimmers are 5'-
TGTTTGATATCATATAAAACA-3'  (SEQ ID NO: 132) and
5'-GAATTGTGAGCGATAACATT-3'  (SEQ ID NO: 133) and
5'-GAATGTGAGCGAGTAACAACCG-3'  (SEQ ID NO: 134) and
5'-CGGCAGTGAGCGCAACGCAATT-3'  (SEQ ID NO: 135) and
5'-GAATGTGAGCGCTTACAATT-3'  (SEQ ID NO: 136)

Preferred heterologous DNA binding domains are monomeric DNA binding domains e.g. HTH domains of transcription factors or monomeric transcription factors.
Similar preferred are DNA binding domains having a high specificity for one or a small group of recognition sequences.
Equally preferred are DNA binding domains having a high affinity for one or a small group of recognition sequences.

In one embodiment the heterologous DNA-binding domain comprises at least one HTH domain of scTet, scArcR, TraR, LacR, LuxR, MarA, or MerR and homologs of any one these having at least 50%, 60%, 70%, 80%, 85%, 90%, 92%, at 93%, 94%, 95%, 96%, 97%, 98% or 99% of sequence identity on amino acid level.

In a further embodiment of the invention, the transcription factor or the DNA binding domain of a transcription factor comprises a HTH domain comprising an amino acid sequence of at least 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99% sequence identity to at least one amino acid sequence described by SEQ ID NO: 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118 or 119, preferably to at least one amino acid se-
sequence described by 91, 92, 93, 94, 95, 112, 113, 114, 115, 116, 117, 118 or 119.

In another embodiment of the invention, the heterologous DNA-binding domain comprises a HTH domain having a sequence identity of at least 50%, 60%, 70%, 80%, 85%, 90%, 92%, at 93%, 94%, 95%, 96%, 97%, 98% or 99% of sequence identity on amino acid level to any one of SEQ ID NO: 91, 92, 93, 94, 95, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118 or 119.

In one embodiment the heterologous DNA-binding domain is selected from the group consisting of: scTet, scArcR, TraR, LacR, LuxR, MarA, or MerR and homologs of any one these having at least 50%, 60%, 70%, 80%, 85%, 90%, 92%, at 93%, 94%, 95%, 96%, 97%, 98% or 99% of sequence identity on amino acid level or the DNA binding domain fragment of scTet, scArcR, TraR, LacR, LuxR, Gal4 and homologs of any one these having at least 50%, 60%, 70%, 80%, 85%, 90%, 92%, at 93%, 94%, 95%, 96%, 97%, 98% or 99% of sequence identity on amino acid level.

In another embodiment the heterologous DNA-binding domain is selected from the group consisting of: scTet, scArcR, TraR, LacR, LuxR and homologs of any one these having at least 50%, 60%, 70%, 80%, 85%, 90%, 92%, at 93%, 94%, 95%, 96%, 97%, 98% or 99% of sequence identity on amino acid level or the DNA binding domain fragment of scTet, scArcR, TraR, LacR, LuxR, Gal4 and homologs of any one these having at least 50%, 60%, 70%, 80%, 85%, 90%, 92%, at 93%, 94%, 95%, 96%, 97%, 98% or 99% of sequence identity on amino acid level.

In another embodiment the heterologous DNA-binding domain is scTet or scArcR and homologs of any one these having at least 50%, 60%, 70%, 80%, 85%, 90%, 92%, at 93%, 94%, 95%, 96%, 97%, 98% or 99% of sequence identity on amino acid level, or the DNA binding domain fragment of scTet or scArcR and homologs of any one these having at least 50%, 60%, 70%, 80%, 85%, 90%, 92%, at 93%, 94%, 95%, 96%, 97%, 98% or 99% of sequence identity on amino acid level.

In another embodiment the heterologous DNA-binding domain is scTet and homologs of scTet having at least 50%, 60%, 70%, 80%, 85%, 90%, 92%, at 93%, 94%, 95%, 96%, 97%, 98% or 99% of sequence identity on amino acid level, or the HTH domain of scTet and homologs of scTet having at least 50%, 60%, 70%, 80%, 85%, 90%, 92%, at 93%, 94%, 95%, 96%, 97%, 98% or 99% of sequence identity on amino acid level.

In another embodiment the heterologous DNA-binding domain is MarA and homologs of MarA having at least 50%, 60%, 70%, 80%, 85%, 90%, 92%, at 93%, 94%, 95%, 96%, 97%, 98% or 99% of sequence identity on amino acid level, or the HTH domain of MarA and homologs thereof having at least 50%, 60%, 70%, 80%, 85%, 90%, 92%, at 93%, 94%, 95%, 96%, 97%, 98% or 99% of sequence identity on amino acid level.
In another preferred embodiment, the heterologous DNA-dinding domain is a TAL effector protein or the DNA binding portion of a TAL effector. One may use native TAL effectors. Alternatively, TAL effectors can be designed to bind to certain recognition sequences (Moscou & Bogdanove, 2009, Science DOI: 10.126/science.1 178817; Boch et al. 2009, Science DOI: 10.126/science.1 1178811) and WO20 10/079430 and EP2206723. WO20 10/079430 and EP2206723 are included herein by reference. Examples for TAL effector proteins are AvBs3 (SEQ ID NO: 160), Hax2 (SEQ ID NO:161), Hax3 (SEQ ID NO: 162) and Hax4 (SEQ ID NO: 163).

The respective DNA binding site or the recognition sequence of AvBs3 is described by 5′-TCTNTAACCCTNNCCCTCT-3’ (SEQ ID NO:164), of Hax2 is described by 5′-TGTTATTCTCACACTCTTTAT-3’ (SEQ ID NO:165), of Hax3 is described by 5′-TACACCCNNNCAT-3’ (SEQ ID NO:166) and of Hax4 is described by 5′-TACCTNNACTANATAT-3’ (SEQ ID NO:167).

Accordingly, in another embodiment, at least one heterologous DNA dinding domain of the chimeric endonuclease is a TAL effector protein having an amino acid sequence identity of at least 80%, 81%, 82%, 83%, 84%, 85%,86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99% to an amino acid sequence described by SEQ ID NO: 160, 161 , 162 or 164, or a fragment of the DNA binding domain of a TAL effector protein having an amino acid sequence identity of at least 80%, 81%, 82%, 83%, 84%, 85%,86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99% to an amino acid sequence described by SEQ ID NO: 160, 161 , 162 or 164, comprising at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 or 20 repeat units derived from a transcription activator-like (TAL) effector, or a transcription activator-like (TAL) effector.

In another embodiment, at least one heterologous DNA dinding domain of the chimeric endonuclease is at least one repeat unit derived from a transcription activator-like (TAL) effector, or a transcription activator-like (TAL) effector. The term "repeat unit" is used to describe the modular portion of a repeat domain from a TAL effector, or an artificial version thereof, that contains one or two amino acids in positions 12 and 13 of the amino acid sequence of a repeat unit that determine recognition of a base pair in a target DNA sequence that such amino acids confer recognition of, as follows: HD for recognition of C/G; NI for recognition of A/T; NG for recognition of T/A; NS for recognition of C/G or A/T or T/A or G/C; NN for recognition of G/C or A/T; IG for recognition of T/A; N for recognition of C/G; HG for recognition of C/G or T/A; H for recognition of T/A; and NK for recognition of G/C. (the amino acids H, D, I, G, S, K are described in one-letter code, whereby A, T, C, G refer to the DNA base pairs recognized by the amino acids)

The number of repeat units to be used in a repeat domain can be ascertained by one skilled in the art by routine experimentation. Generally, at least 1.5 repeat units are considered as a minimum, although typically at least about 8 repeat units will be used. The repeat units do not
have to be complete repeat units, as repeat units of half the size can be used. A heterologous DNA binding domain of the invention can comprise, for example, 1.5, 2.5, 3, 3.5, 4, 4.5, 5, 5.5, 6, 6.5, 7, 7.5, 8, 8.5, 9, 9.5, 10, 10.5, 11, 11.5, 12, 12.5, 13, 13.5, 14, 14.5, 15, 15.5, 16, 16.5, 17, 17.5, 18, 18.5, 19, 19.5, 20, 20.5, 21, 21.5, 22, 22.5, 23, 23.5, 24, 24.5, 25, 25.5, 26, 26.5, 27, 27.5, 28, 28.5, 29, 29.5, 30, 30.5, 31, 31.5, 32, 32.5, 33, 33.5, 34, 34.5, 35, 35.5, 36, 36.5, 37, 37.5, 38, 38.5, 39, 39.5, 40, 40.5, 41, 41.5, 42, 42.5, 43, 43.5, 44, 44.5, 46, 46.5, 47, 47.5, 48, 48.5, 49, 49.5, 50, 50.5 or more repeat units.

A typical consensus sequence of a repeat with 34 amino acids (in one-letter code) is shown below:

\[
\text{LTPEQVVAIasnGgKQaletvrLLPvLCQAHG (SEQ ID NO: 128)}
\]

A further consensus sequence for a repeat unit with 35 amino acids (in one-letter code) is as follows:

\[
\text{LTPEQVVAIasnGgKQaletvrLLPvLCQAPHD (SEQ ID NO: 129)}
\]

The repeat units which can be used in one embodiment of the invention have an identity with the consensus sequences described above of at least 35%, 40%, 50%, 60%, 70%, 75%, 80%, 85%, 90% or 95%.

In one embodiment of the invention, the heterologous DNA binding domain is a transcription activator-like (TAL) effector of the group of transcription activator-like (TAL) effectors described by: AvrBs3, AvrBs3-repl6, AvrBs3-repl09, AvrHahl, AvrXa27, PthXol, PthXo6, PthXo7, or the members of the Hax sub-family Hax2, Hax3, Hax4 and Brgll, or homologs of these having at least 50%, 60%, 70%, 80%, 85%, 90%, 92%, at 93%, 94%, 95%, 96%, 97%, 98% or 99% of sequence identity on amino acid level.

In one embodiment of the invention, the heterologous DNA binding domain is not a TAL-Effector protein or a TAL-Effector repeat unit.

Preparation of Chimeric endonucleases:

Endonucleases and the heterologous DNA binding domains can be combined in many alternative ways.

For example, it is possible, to combine more than one endonuclease with one or more heterologous DNA binding domain or to combine more than one heterologous DNA binding domain with one endonuclease. It is also possible to combine more than one endonuclease with more than one heterologous DNA binding domain.

The heterologous DNA-binding domain or the heterologous DNA-binding-domains can be fused at the N-terminal or at the C-terminal end of the endonuclease. It is also possible, to fuse one or more heterologous DNA binding domains at the N-terminal end and one or more heterologous
DNA binding domains at the C-terminal end of the endonuclease. It is also possible to make alternating combinations of endonucleases and heterologous DNA binding domains. In case the chimeric endonuclease comprises more than one endonuclease or more than one heterologous DNA binding domain or more than one endonuclease and more than one heterologous DNA binding domain, it is possible to use several copies of the same heterologous DNA binding domain or endonuclease or to use different heterologous DNA binding domains or endonucleases.

It is also possible to apply the methods and features described for optimized nucleases above, to the full sequence of chimeric endonucleases, e.g. by adding a nuclear localization signal to a chimeric endonuclease or by reducing the number of:

- PEST-Sequences,
- KEN-boxes
- A-boxes,
- D-boxes, or
- comprise an optimized N-terminal end for stability according to the N-end rule,
- comprise a glycin as the second N-terminal amino acid, or
- any combination of a), b), c) d), e) and f). of the entire amino acid sequence of the chimeric endonuclease.

Chimeric endonucleases having a nuclear localization signal are for example described by the amino acid sequence described by SEQ ID NO: 11, or the polynucleotide sequence described by SEQ ID NO: 24, 25 or 26.

In one embodiment the chimeric endonucleases are combinations of:

- l-Scel and scTet, or l-Scel and scArc, or l-Crel and scTet, or l-Crel and scArcR orl-Msol and scTet, or l-Msol and scArcR, wherein scTet, or scArcR are fused N- or C-terminal to l-Scel, l-Crel or l-Msol and wherein l-Scel, l-Crel, l-Msol, scTet, scArcR, include their homologs having at least 50%, 49%, 51%, 58%, 60%, 70%, 80%, 85%, 90%, 92%, at 93%, 94%, 95%, 96%, 97%, 98% or 99% of sequence identity on amino acid level.

In another embodiment the chimeric endonucleases have the following structure:

- N-terminus-l-Scel-scTet- C-terminus, or
- N-terminus-l-Scel- scArcR - C-terminus, or
- N-terminus- l-Crel -scTet- C-terminus, or
- N-terminus- l-Crel - scArcR - C-terminus, or
- N-terminus- l-Msol -scTet- C-terminus, or
- N-terminus- l-Msol - scArcR - C-terminus,
- N-terminus- scTet-l-Scel- C-terminus, or
N-terminus- scArcR - l-Scel- C-terminus, or
N-terminus- scTet - l-Crel - C-terminus, or
N-terminus- scArcR - l-Crel - C-terminus, or

N-terminus- scTet- l-Msol - C-terminus, or
N-terminus- scArcR - l-Msol - C-terminus,

The chimeric endonuclease is preferably expressed as a fusion protein with a nuclear localization sequence (NLS). This NLS sequence enables facilitated transport into the nucleus and increases the efficacy of the recombination system. A variety of NLS sequences are known to the skilled worker and described, inter alia, by Jicks GR and Raikhel NV (1995) Annu. Rev. Cell Biol. 11:155-188. Preferred for plant organisms is, for example, the NLS sequence of the SV40 large antigen. Examples are provided in WO 03/060133 included herein by reference. The NLS may be heterologous to the endonuclease and/or the DNA binding domain or may be naturally comprised within the endonuclease and/or DNA binding domain.

In a preferred embodiment, the sequences encoding the chimeric endonucleases are modified by insertion of an intron sequence. This prevents expression of a functional enzyme in prokaryotic host organisms and thereby facilitates cloning and transformations procedures (e.g., based on E.coli or Agrobacterium). In eukaryotic organisms, for example plant organisms, expression of a functional enzyme is realized, since plants are able to recognize and “splice” out introns. Preferably, introns are inserted in the homing endonucleases mentioned as preferred above (e.g., into l-Scel or l-Crel).

In another preferred embodiment, the amino acid sequences of the endonuclease or the chimeric endonuclease can be modified by adding a Sec IV secretion signal to the N- or C-Terminus of the endonuclease or chimeric endonuclease.

In a preferred embodiment the SecIV secretion signal is a SecIV secretion signal comprised in Vir proteins of Agrobacterium. Examples of such Sec IV secretion signals as well as methods how to apply these are disclosed in WO 01/89283, in Vergunst et al, Positive charge is an important feature of the C-terminal transport signal of the VirB/D4-translocated proteins of Agrobacterium, PNAS 2005, 102, 03, pages 832 to 837 included herein by reference. A Sec IV secretion signal might also be added, by adding fragments of a Vir protein or even a complete Vir protein, for example a complete VirE2 protein to a endonuclease or chimeric endonuclease, in a similar way as described in the description of WO01/38504 included herein by reference, which describes a RecAA/irE2 fusion protein.

In another preferred embodiment the amino acid sequences of the endonuclease or the chimeric endonuclease can be modified by adding a Sec III secretion signal to the N- or C-Terminus of the endonuclease or chimeric endonuclease. Suitable SecIII secretion signals are for example disclosed in WO 00/02996, included herein by reference.
In case a Sec III secretion signal is added, it can be of advantage, to express this endonuclease or chimeric endonuclease in a cell, which does also comprise a recombinant construct encoding parts of, or a complete functional type III secretion system, in order to overexpress or complement parts or the complete functional type III secretion system in such cell.

Recombinant constructs encoding parts or a complete functional type III secretion system are for example disclosed inWO 00/02996 and WO05/085417 included herein by reference.

If a SecIV secretion signal is added to the chimeric endonuclease and the chimeric endonuclease is intended to be expressed for example in Agrobacterium rhizogenes or in Agrobacterium tumefaciens, it is of advantage to adapt the DNA sequence coding for the chimeric endonuclease to the codon usage of the expressing organism. Preferably the endonuclease or chimeric nuclease does not have or has only few DNA recognition sequences in the genome of the expressing organism. It is of even greater advantage, if the selected chimeric endonuclease does not have a DNA recognition sequence or less preferred DNA recognition sequence in the Agrobacterium genome. In case the nuclease or the chimeric endonuclease is intended to be expressed in a prokaryotic organism the nuclease or chimeric nuclease encoding sequence must not have an intron.

In one embodiment the endonuclease and the heterologous DNA binding domain are connected via a linker polypeptide.

Preferably the linker polypeptide consists of 1 to 30 amino acids, more preferred 1 to 20 and even more preferred 1 to 10 amino acids.

For example, the linker polypeptide can be composed of a plurality of residues selected from the group consisting of glycine, serine, threonine, cysteine, asparagine, glutamine, and proline.

Preferably the linker polypeptide is designed to lack secondary structures under physiological conditions and is preferably hydrophilic. Charged or non polar residues may be included, but they may interact to form secondary structures or may reduce solubility and are therefore less preferred.

In some embodiments the linker polypeptide consists essentially of a plurality of residues selected from glycine and serine. Examples of such linkers have the amino acid sequence (in one letter code): GS, or GGS, or GSGS, or GSGSGS, or GGSGG, or GGSGGGSGG, or GSGSGGSGG.

In case the linker consists of at least 3 amino acids, it is preferred that the amino acid sequence of the linker polypeptide comprises at least one third Glycines or Alanines or Glycines and Alanines.

In one preferred embodiment, the linker sequence has the amino acid sequence GSGS or GSGSGS.

Preferably the polypeptide linker is rationally designed using bioinformatic tools, capable of modeling both the DNA-binding site and the respective edonuclease, as well as the recognition site and the heterologous DNA-binding domain. Suitable bioinformatic tools are for example described in Desjarlais & Berg, (1994), PNAS, 90, 2256 to 2260 and in Desjarlais & Berg
DNA recognition sequences of chimeric endonucleases:

The chimeric endonucleases bind to DNA sequences being combinations of the DNA recognition sequence of the endonuclease and the recognition sequence of the heterologous DNA binding domain. In case the chimeric endonuclease comprises more than one endonuclease or more than one heterologous DNA binding domain the DNA the chimeric endonuclease will bind to DNA sequences being a combination of the DNA recognition sequence of the endonucleases used and the operator sequences of the heterologous DNA binding domains used. It is clear, that the sequence of the DNA, which is bound by the chimeric endonuclease will reflect the order, in which the endonuclease and the heterologous DNA binding domains are combined.

Endonucleases known in the art cut a huge variety of different polynucleotide sequences.

The terms DNA recognition sequence and DNA recognition site are used synonymously and refer to a polynucleotide of a particular sequence which can be bound and cut by a given endonuclease. A polynucleotide of a given sequence may therefore be a DNA recognition sequence or DNA recognition site for one endonuclease, but may or may not be a DNA recognition sequence or DNA recognition site for another endonuclease.

Examples of polynucleotide sequences which can be bound and cut by endonucleases, i.e. which represent a DNA recognition sequence or DNA recognition site for this endonuclease, are described in Table 8: the letter N represents any nucleotide, and can be replaced by A, T, G or C).

<table>
<thead>
<tr>
<th>Tabel 8</th>
<th>Organism of origin</th>
<th>DNA recognition sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>I-Crel Chlamydomonas reinhardtii</td>
<td></td>
<td>5’-CAAAACGTCTGAGACAGTTTC-3’ (SEQ ID NO: 138)</td>
</tr>
<tr>
<td>I-Ceul Chlamydomonas eugametos</td>
<td></td>
<td>5’-ATAACGGTCTAAGGTAGCGAA-3’ (SEQ ID NO: 139)</td>
</tr>
<tr>
<td>I-Dmol Desulfurococcus mobilis</td>
<td></td>
<td>5’-ATGCCCTTGCAGGATTTCCGGCAGCAT-3’ (SEQ ID NO: 140)</td>
</tr>
<tr>
<td>I-Msol Monomastix spec.</td>
<td></td>
<td>5’-CAGAACGTCTGAGACAGTTTC-3’ (SEQ ID NO: 153)</td>
</tr>
<tr>
<td>PI-PsiI S. cerevisia</td>
<td></td>
<td>5’-ATCTATGTCGGGTGCCAGAAAGAGGTAAT-3’ (SEQ ID NO: 154)</td>
</tr>
<tr>
<td>I-Anil Aspergillus nidulans</td>
<td></td>
<td>5’-CGCGCGTGAGAGGTTTCTGTAAAGCGCA-3’ (SEQ ID NO: 142)</td>
</tr>
</tbody>
</table>

Endonucleases do not have stringently-defined DNA recognition sequences, so that single base changes do not abolish cleavage but may reduce its efficiency to variable extents. A DNA rec-
ognition sequence listed herein for a given endonuclease represents only one site that is known to be recognized and cleaved.

Examples for deviations of a DNA recognition site are for example disclosed in Chevelier et al. (2003), J.Mol.Biol. 329, 253 to 269, in Marcaida et al. (2008), PNAS, 105 (44), 16888 to 16893 and in the Supporting Information to Marcaida et al. 10.1073/pnas.0804795105, in Doyon et al. (2006), J. AM. CHEM. SOC. 128, 2477 to 2484, in Argast et al. (1998), J.Mol.Biol. 280, 345 to 353, in Spiegel et al. (2006), Structure, 14, 869 to 880, in Posey et al. (2004), Nucl. Acids Res. 32 (13), 3947 to 3956, or in Chen et al. (2009), Protein Engineering, Design & Selection, 22 (4), 249 to 256.

It is therefore possible to identify a naturally occurring endonuclease having a predetermined polynucleotide sequence as a DNA recognition sequence. Methods to identify naturally occurring endonucleases, their genes and their DNA recognition sequences are disclosed for example in WO 2009/101625.

The cleavage specificity or respectively its degeneration of its DNA recognition sequence can be tested by testing its activity on different substrates. Suitable in vivo techniques are for example disclosed in WO09074873.

Alternatively, in vitro tests can be used, for example by employing labeled polynucleotides spotted on arrays, wherein different spots comprise essentially only polynucleotides of a particular sequence, which differs from the polynucleotides of different spots and which may or may not be DNA recognition sequences of the endonuclease to be tested for its activity. A similar technique is disclosed for example in US 2009/0197775.

However, it is possible to mutate the amino acid sequence of a given endonuclease, preferably a LAGLIDADG endonuclease, to bind and cut new polynucleotides, i.e. creating an engineered endonuclease having a changed DNA recognition site.


Therefore it is also possible to create an engineered endonuclease which will have a DNA recognition sequence identical to a particular predetermined polynucleotide sequence.

Preferably the DNA recognition sequence of the endonuclease and the operator sequence are separated by 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 or more base pairs. Preferably they are separated by 1 to 10, 1 to 8, 1 to 6, 1 to 4, 1 to 3, or 2 base pairs.

The amount of base pairs used to separate the DNA recognition sequence of the nuclease and the recognition sequence of the heterologous DNA binding domain depends on the distance of the DNA binding regions of the nuclease and the DNA binding region of the heterologous DNA binding domain in the chimeric endonuclease. A larger distance between the DNA binding re-
gions of the nuclease and the DNA binding region of the heterologous DNA binding domain will be reflected by a higher amount of base pairs separating the DNA recognition sequence of the nuclease and the recognition sequence of the heterologous DNA binding domain. The optimal amount of separating base pairs can be determined by using computer models or by testing the binding and cutting efficiency of a given chimeric endonuclease on several polynucleotides comprising a varying amount of base pairs between the DNA recognition sequence of the nuclease and the recognition sequence of the heterologous DNA binding domain.

Accordingly, in one embodiment of the invention, the chimeric recognition site comprises a DNA recognition sequence of a LAGLIDADG endonuclease, even more preferred a DNA recognition sequence of a LAGLIDADG endonuclease having an amino acid sequence as described by at least one of SEQ ID NOs: 1, 2, 3, 5, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99% to 1-

In a further embodiment of the invention, the chimeric recognition site comprises a DNA recognition sequence of l-Scel, l-Crel, l-Dmol, l-Msol, l-Ceul, l-Chul, Pi-Scel or l-Anil or a homolog of these having at least 56%, 57%, 58%, 59%, 60%, 61%, 62%, 63%, 65%, 66%, 67%, 68%, 69%, 70%, 71%, 72%, 73%, 74%, 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99% to l-Scel, l-Crel, l-Dmol, l-Msol, l-Ceul, l-Chul, Pi-Scel or l-Anil, and a recognition sequence of a heterologous DNA binding domain having at least 50% sequence amino acid sequence identity to scTet, scArc, LacR, MerR or MarA or to a DNA binding domain fragment of scTet, scArc, LacR, MerR or MarA.

In a further embodiment of the invention, the chimeric recognition site comprises two DNA recognition sequences of l-Scel, l-Crel, l-Dmol, l-Msol, l-Ceul, l-Chul, Pi-Scel or l-Anil or a homolog of these having at least 56%, 57%, 58%, 59%, 60%, 61%, 62%, 63%, 65%, 66%, 67%, 68%, 69%, 70%, 71%, 72%, 73%, 74%, 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99% to l-Scel, l-Crel, l-Dmol, l-Msol, l-Ceul, l-Chul, Pi-Scel or l-Anil.

Such chimeric recognition sites can be used with chimeric endonucleases comprising an active endonuclease and an inactive endonuclease as heterologous DNA binding domain.

One example for such types of combinations are a chimeric recognition site comprising two DNA recognition sequences of l-Scel, which can be used in combination with a chimeric endonuclease comprising an active version of l-Scel and an inactive version of l-Scel as heterologous DNA binding domain.

In a further embodiment of the invention, the chimeric recognition site comprises two DNA recognition sequences of l-Scel, l-Crel, l-Dmol, l-Msol, l-Ceul, l-Chul, Pi-Scel or l-Anil or a homolog of these having at least 56%, 57%, 58%, 59%, 60%, 61%, 62%, 63%, 65%, 66%, 67%, 68%, 69%, 70%, 71%, 72%, 73%, 74%, 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99% to l-
Seel, l-Crel, l-Dmol, l-Msol, l-Ceul, l-Chul, Pi-Scel or l-Anil and a DNA binding site of a TAL-effector protein, preferably comprising a polynucleotide sequence as described by SEQ ID NO: 164, 165, 166 or 167.

In another embodiment of the invention, the the chimeric recognition site comprises a two DNA recognition sequences of l-Scel, preferably described by SEQ ID NO: 13 and a DNA binding site of a TAL-effector protein, preferably comprising a polynucleotide sequence as described by SEQ ID NO: 164, 165, 166 or 167.

Examples for DNA recognition sequences of chimeric endonucleases (chimeric recognition site or target site of the respective chimeric endonuclease) are:

A chimeric endonuclease having the structure: l-Scel - scTet, preferably having an amino acid sequence described by SEQ ID NO: 8 or 9

I-Scel scTet target site 1 ctatacaatgatagcctaggataacaggttaat (SEQ ID NO: 14)
I-Scel scTet target site 2 ctatacaatgatagcctaggataacaggttaat (SEQ ID NO: 15)
I-Scel scTet target site 3 ctatacaatgatagcctaggataacaggttaat (SEQ ID NO: 16)

A chimeric endonuclease having the structure: l-Scel-scArcR, preferably having an amino acid sequence described by SEQ ID NO: 10 or 11

I-Scel scArc target site 1 tagggataacaggttaatactagtagtgct (SEQ ID NO: 17)
I-Scel scArc target site 2 tagggataacaggttaatactagtagtgct (SEQ ID NO: 18)
I-Scel scArc target site 3 tagggataacaggttaatactagtagtgct (SEQ ID NO: 19)
I-Scel scArc target site 4 tagggataacaggttaatactagtagtgct (SEQ ID NO: 20)

Poly nucleotides:
The invention does also comprise isolated polynucleotides coding for the chimeric endonucleases described above.

Examples of such isolated polynucleotides are isolated polynucleotides coding for amino acid sequences described by SEQ ID NO: 23, 24, 25 and 26 or amino acid sequences having at least 70%, 80%, 90% 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98% or 99% amino acid sequence similarity, preferably having at least 70%, 80%, 90% 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98% or 99% amino acid sequence identity to any one of the amino acid sequences described by SEQ ID NO: 23, 24, 25 and 26.

Preferably the isolated polynucleotide has a optimized codon usage for expression in a particular host organism, or has a low content of RNA instability motifs, or has a low content of codon repeats, or has a low content of cryptic splice sites, or has a low content of alternative start codons, or has a low content of restriction sites, or has a low content of RNA secondary structures or has any combination of these features.

The codon usage of the isolated polypeptide may be optimized e.g. for the expression in plants, preferably in a plant selected from the group comprising: rice, corn, wheat, rape seed, sugar
cane, sunflower, sugar beet, tobacco.

Preferably the isolated polynucleotide is combined with a promoter sequence and a terminator sequence suitable to form a functional expression cassette for expression of the chimeric endonuclease in a particular host organism.

Suitable promoters are for example constitutive, heat- or pathogen-inducible, or seed, pollen, flower or fruit specific promoters.

The person skilled in the art knows numerous promoters having those features.

For example several constitutive promoters in plants are known. Most of them are derived from viral or bacterial sources such as the nopaline synthase (nos) promoter (Shaw et al. (1984) Nucleic Acids Res. 12 (20) :7831-7846), the mannopine synthase (mas) promoter (Co-mai et al. (1990) Plant Mol Biol 15(3):373-381), or the octopine synthase (ocs) promoter (Leisner and Gelvin (1988) Proc Natl Acad Sci USA 85 (5) :2553-2557) from Agrobacterium tumefaciens or the CaMV35S promote from the Cauliflower Mosaic Vi-rus (US 5,352, 605). The latter was most frequently used in constitutive expression of transgenes in plants (Odell et al. (1985) Nature 313:810-812; Battraw and Hall (1990) Plant Mol Biol 15:527-538; Benfey et al. (1990) EMBO J 9(69):1 677-1 684; US 5,612,472). However, the CaMV 35S promoter demonstrates variability not only in different plant species but also in different plant tissues (Atanassova et al. (1998) Plant Mol Biol 37:275-85; Battraw and Hall (1990) Plant Mol Biol 15:527-538; Holtorf et al. (1995) Plant Mol Biol 29:637-646 ; Jefferson et al. (1987) EMBO J 6 :3901-3907). An additional disadvantage is an interference of the transcription regulating activity of the 35S promoter with wild-type CaMV virus (Al-Kaff et al. (2000) Nature Biotechnology 18 :995-99). Another viral promoter for constitutive expression is the Sugarcane bacilliform badnavirus (ScBV) promoter (Schenk et al. (1999) Plant Mol Biol 39 (6) :1221-1230).

Several plant constitutive promoters are described such as the ubiquitin promoter from Arabidopsis thaliana (Callis et al. (1990) J Biol Chem 265:12486- 12493; Holtorf S et al. (1995) Plant Mol Biol 29:637-747), which - however - is reported to be unable to regulate expression of selection markers (WO031 021 98), or two maize ubiquitin promoter (Ubi-1 and Ubi-2; US 5,510,474; US 6,020, 190; US 6,054574), which beside a constitutive expression profile demonstrate a heat-shock induction (Christensen et al. (1992) Plant. Mol. Biol. 18(4):675-689). A comparison of specificity and expression level of the CaMV 35S, the barley thionine promoter, and the Arabidopsis ubiquitin promoter based on stably transformed Arabidopsis plants demonstrates a high expression rate for the CaMV 35S promoter, while the thionine promoter was inactive in most lines and the ubil promoter from Arabisopsis resulted only in moderate expression activity (Holtorf et al. (1995) Plant Mol Biol 29 (4):637-646).

Chimeric recognition sequences:

The invention does also comprise isolated polynucleotides comprising a chimeric recognition
sequence, having a length of about 15 to about 300, or of about 20 to about 200 or of about 25 to about 100 nucleotides, comprising a DNA recognition sequence of an endonuclease and a recognition sequence of a heterologous DNA binding domain (also called binding site or operator).

Preferably isolated polynucleotides comprise a DNA recognition sequence of a homing endonuclease, preferably of a LAGLIDADG endonuclease.

In one embodiment the isolated polynucleotide comprises a DNA recognition sequence of l-Scel.

Preferably the recognition sequence of a heterologous DNA binding domain comprised in the isolated polynucleotide is a recognition sequence of a transcription factor. More preferably the recognition sequence is the recognition sequence of the transcription factors scTet or scArc.

In one embodiment the isolated polynucleotide comprises a DNA recognition sequence of l-Scel and a linker sequence of 0 to 10 polynucleotides and a recognition sequence of scTet or scArc.

Preferred chimeric recognition sequences comprise a combination of a DNA recognition sequence of l-Scel, l-Crel, l-Dmol, or l-Ceu, l-Msol, Pi-Scel or l-Anil in combination with a recognition site of scTet, TetR, scArcR, TraR, WRKY, LacR, MarA or MerR, wherein the DNA recognition sequence of l-Scel, l-Crel, l-Dmol, l-Msol, or l-Ceu may be fused in a distance of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 or 12 nucleotides up or downstream of a recognition site of scTet, TetR, scArcR, TraR, WRKY, LacR, MarA or MerR.

Preferred chimeric recognition sequences comprise a combination of a DNA recognition sequence of l-Scel, l-Crel, l-Dmol, or l-Ceu in combination with a recognition site of scTet, TetR, scArcR, TraR, MarA or MerR, wherein the DNA recognition sequence of l-Scel, l-Crel, l-Dmol, or l-Ceu may be fused in a distance of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 or 12 nucleotides up or downstream of a recognition site of scTet, TetR, scArcR, TraR, MarA or MerR.

Preferred chimeric recognition sequences comprise a combination of a DNA recognition sequence of l-Scel, l-Crel, l-Dmol or l-Msol in combination with a recognition site of scTet, TetR, scArcR, TraR, MarA or MerR, wherein the DNA recognition sequence of l-Scel, l-Crel, l-Dmol, or l-Ceu may be fused in a distance of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 or 12 nucleotides up or downstream of a recognition site of scTet, TetR, scArcR, TraR, MarA or MerR.

In one embodiment, the chimeric recognition sequence comprise a combination a DNA recognition sequence of l-Scel in combination with a recognition site of scTet, TetR, scArcR, TraR, MarA or MerR, wherein the DNA recognition sequence of l-Scel may be fused in a distance of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 or 12 nucleotides up or downstream of a recognition site of scTet, TetR, scArcR, TraR, MarA or MerR.
In one embodiment, the chimeric recognition sequence comprise a combination a DNA recognition sequence of I-SceI in combination with a recognition site of MarA wherein the DNA recognition sequence of I-SceI may be fused in a distance of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 or 12 nucleotides up or downstream of a recognition site of MarA. Preferably, the DNA recognition sequence of I-SceI is fused upstream of a recognition site of MarA.

In one embodiment the isolated polynucleotide comprise a sequence of a chimeric recognition site selected from the group comprising: SEQ ID NO: 30, 31, 32, 34, 35, 36 or 37.

The isolated polynucleotides may comprise a combination of a chimeric recognition site and a polynucleotide sequence coding for a chimeric nuclease.

In a preferred embodiment of the invention, a chimeric endonuclease having an amino acid sequence as described by SEQ ID NO: 8 or 9, is used in combination with a chimeric recognition sequence having a polynucleotide sequence selected from the group of sequences described by: SEQ ID NO: 14, 15 or 16.

In a preferred embodiment of the invention, a chimeric endonuclease having an amino acid sequence as described by SEQ ID NO: 10 or 11, is used in combination with a chimeric recognition sequence having a polynucleotide sequence selected from the group of sequences described by: SEQ ID NO: 17, 18, 19 or 20.

Vectors:
The polynucleotides described above may be comprised in a DNA vector suitable for transformation, transfection, cloning or overexpression.

In one example, the polynucleotides described above are comprised in a vector for transformation of non-human organisms or cells, preferably the non-human organisms are plants or plant cells.

The vectors of the invention usually comprise further functional elements, which may include but shall not be limited to:

i) Origins of replication which ensure replication of the expression cassettes or vectors according to the invention in, for example, E. coli. Examples which may be mentioned are ORI (origin of DNA replication), the pBR322 ori or the P15A ori (Sam-brook et al.: Molecular Cloning. A Laboratory Manual, 2nd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1989).

ii) Multiple cloning sites (MCS) to enable and facilitate the insertion of one or more nucleic acid sequences.

iii) Sequences which make possible homologous recombination or insertion into the genome of
a host organism.

iv) Elements, for example border sequences, which make possible the Agrobacterium-mediated
transfer in plant cells for the transfer and integration into the plant genome, such as, for exam-
ple, the right or left border of the T-DNA or the vir region.

The Marker Sequence

The term "marker sequence" is to be understood in the broad sense to include all nucleotide
sequences (and/or polypeptide sequences translated therefrom) which facilitate detection, iden-
tification, or selection of transformed cells, tissues or organism (e.g., plants). The terms "se-
quence allowing selection of a transformed plant material", "selection marker" or "selection
marker gene" or "selection marker protein" or "marker" have essentially the same meaning.

Markers may include (but are not limited to) selectable marker and screenable marker. A se-
lectable marker confers to the cell or organism a phenotype resulting in a growth or viability dif-
ference. The selectable marker may interact with a selection agent (such as a herbicide or anti-
biotic or pro-drug) to bring about this phenotype. A screenable marker confers to the cell or or-
ganism a readily detectable phenotype, preferably a visibly detectable phenotype such a color
or staining. The screenable marker may interact with a screening agent (such as a dye) to bring
about this phenotype.

Selectable marker (or selectable marker sequences) comprise but are not limited to
a) negative selection marker, which confers resistance against one or more toxic (in case of
plants phytotoxic) agents such as an antibiotica, herbicides or other biocides,
b) counter selection marker, which confer a sensitivity against certain chemical compounds
(e.g., by converting a non-toxic compound into a toxic compound), and
c) positive selection marker, which confer a growth advantage (e.g., by expression of key ele-
ments of the cytokinin or hormone biosynthesis leading to the production of a plant hormone e.
ge., auxins, gibberllins, cytokinins, abscisic acid and ethylene; Ebi-numa H et al. (2000) Proc Natl
Acad Sci USA 94:21 17-2121).

When using negative selection markers, only cells or plants are selected which comprise said
negative selection marker. When using counter selection marker, only cells or plants are se-
lected which lack said counter-selection marker. Counter-selection marker may be employed to
verify successful excision of a sequence (comprising said counter-selection marker) from a ge-
nome. Screenable marker sequences include but are not limited to reporter genes (e. g.
luciferase, glucuronidase, chloramphenicol acetyl transferase (CAT, etc.). Preferred marker se-
quences include but shall not be limited to:

i) Negative selection marker

As a rule, negative selection markers are useful for selecting cells which have success-fully un-
dergone transformation. The negative selection marker, which has been introduced with the
DNA construct of the invention, may confer resistance to a biocide or phytotoxic agent (for ex-
ample a herbicide such as phosphinothricin, glyphosate or bromoxynil, a metabolism inhibitor such as 2-deoxyglucose-6-phosphate (WO 98/45456) or an antibiotic such as, for example, tetracyclic ampicillin, kanamycin, G 418, neomycin, bleomycin or hygromycin to the cells which have successfully under-gone transformation. The negative selection marker permits the selection of the trans-formed cells from untransformed cells (McCormick et al. (1986) Plant Cell Reports 5:81-84). Negative selection marker in a vector of the invention may be employed to confer resistance in more than one organism. For example a vector of the invention may comprise a selection marker for amplification in bacteria (such as E.coli or Agrobacterium) and plants. Examples of selectable markers for E.coli include: genes specifying resistance to antibiotics, i.e., ampicillin, tetracycline, kanamycin, erythromycin, or genes conferring other types of selectable enzymatic activities such as galactosidase, or the lactose operon. Suitable selectable markers for use in mammalian cells include, for example, the dihydrofolate reductase gene (DHFR), the thymidine kinase gene (TK), or prokaryotic genes conferring drug resistance, gpt (xanthine-guanine phosphoribosyltransferase, which can be selected for with mycophenolic acid; neo (neomycin phosphotransferase), which can be selected for with G418, hygromycin, or puromycin; and DHFR (dihydrofolate reductase), which can be selected for with methotrexate (Mulligan & Berg (1981) Proc Natl Acad Sci USA 78:2072; Southern & Berg (1982) J Mol Appl Genet 1:327). Selection markers for plant cells often confer resistance to a biocide or an antibiotic, such as, for example, kanamycin, G 418, bleomycin, hygromycin, or chloramphenicol, or herbicide resistance, such as resistance to chlorsulfuron or Basta.

Especially preferred negative selection markers are those which confer resistance to herbicides. Examples of negative selection markers are:

- DNA sequences which encode phosphinothricin acetyltransferases (PAT), which acetylates the free amino group of the glutamine synthase inhibitor phosphinothricin (PPT) and thus brings about detoxification of PPT (de Block et al. (1987) EMBO J 6:2513-2518) (also referred to as Bialophos- resistance gene bar; EP 242236),
- 5-enolpyruvylshikimate-3-phosphate synthase genes (EPSP synthase genes), which confer resistance to Glyphosate- (N-(phosphonomethyl)glycine),
- the gox gene, which encodes the Glyphosate-degrading enzyme Glyphosate oxi-doreductase,
- the deh gene (encoding a dehalogenase which inactivates Dalapon-),
- acetolactate synthases which confer resistance to sulfonylurea and imidazolinone,
- bxn genes which encode Bromoxynil-degrading nitrilase enzymes,
- the kanamycin, or G418, resistance gene (NPTII). The NPTII gene encodes a neomycin phosphatransferase which reduces the inhibitory effect of kanamycin, neomycin, G418 and paromomycin owing to a phosphorylation reaction (Beck et al (1982) Gene 19: 327),
- the DOGR1 gene. The DOGR1 gene has been isolated from the yeast Saccharomyces cerevisiae (EP 0 807 836). It encodes a 2-deoxyglucose-6-phosphate phos-phatase which confers resistance to 2-DOG (Randez-Gil et al. (1995) Yeast 11:1233-1240),
- the hyg gene, which codes for the enzyme hygromycin phosphotransferase and confers resistance to the antibiotic hygromycin (Gritz and Davies (1983) Gene 25: 179);
- especially preferred are negative selection markers that confer resistance against the toxic
effects imposed by D-amino acids like e.g., D-alanine and D-serine (WO 03/060133; Erikson 2004). Especially preferred as negative selection marker in this contest are the daol gene (EC: 1.4. 3.3 : GenBank Acc.-No.: U60066) from the yeast Rhodotorula gracilis (Rhodosporidium torulooides) and the E. coli gene dsdA (D-serine dehydratase (D-serine deaminase) (EC: 4.3. 1.18; GenBank Acc.-No.: J01603).

ii) Positive selection marker
Positive selection marker comprise but are not limited to growth stimulating selection marker genes like isopentenyltransferase from Agrobacterium tumefaciens (strain: P022; Genbank Acc.-No.: AB025109) may - as a key enzyme of the cytokinin biosynthesis - facilitate regeneration of transformed plants (e.g., by selection on cyto-kinin-free medium). Corresponding selection methods are described (Ebinuma H et al. (2000) Proc Natl Acad Sci USA 94:21 17-2121; Ebinuma H et al. (2000) Selection of Marker-free transgenic plants using the oncogenes (ipt, rol A, B, C) of Agrobacterium as selectable markers, In Molecular Biology of Woody Plants. Kluwer Academic Publishers). Additional positive selection markers, which confer a growth advantage to a transformed plant in comparison with a non-transformed one, are described e.g., in EP-A 0 601 092. Growth stimulation selection markers may include (but shall not be limited to) beta-Glucuronidase (in combination with e.g., a cytokinin glucuronide), mannose-6-phosphate isomerase (in combination with mannose), UDP-galactose-4-epimerase (in combination with e.g., galactose), wherein mannose-6-phosphate isomerase in combination with mannose is especially preferred.

iii) Counter selection markers

In a preferred embodiment the excision cassette includes at least one of said counter-selection markers to distinguish plant cells or plants with successfully excised sequences from plant which still contain these. In a more preferred embodiment the excision cassette of the invention comprises a dual-function marker i.e. a marker with can be employed as both a negative and a counter selection marker depending on the substrate employed in the selection scheme. An example for a dual-function marker is the daol gene (EC: 1.4. 3.3 : GenBank Acc.-No.: U60066) from the yeast Rhodotorula gracilis, which can be employed as negative selection marker with D-amino acids such as D-alanine and D-serine, and as counter-selection marker with D-amino acids such as D-isoleucine and D-valine (see European Patent Appl. No.: 04006358.8 )
iv) Screenable marker (reporter genes)

Screenable marker (such as reporter genes) encode readily quantifiable or detectable proteins and which, via intrinsic color or enzyme activity, ensure the assessment of the transformation efficacy or of the location or timing of expression. Especially preferred are genes encoding reporter proteins (see also Schenborn E, Groskreutz D. (1999) Mol Biotechnol 13(1):29-44) such as


- Chloramphenicol transferase,
- -luciferase (Millar et al. (1992) Plant Mol Biol Rep 10:324-41 4; Ow et al. (1986) Science 234:856-859) permits selection by detection of bioluminescence,
- beta -galactosidase, encodes an enzyme for which a variety of chromogenic substrates are available,
- beta-glucuronidase (GUS) (Jefferson et al. (1987) EMBO J 6:3901-3907) or the uidA gene, which encodes an enzyme for a variety of chromogenic substrates,
- R locus gene product: protein which regulates the production of anthocyanin pig-ments (red coloration) in plant tissue and thus makes possible the direct analysis of the promoter activity without the addition of additional adjuvants or chromogenic substrates (Dellaporta et al. (1988) In: Chromosome Structure and Function: Impact of New Concepts, 18th Stadler Genetics Sym-
- posium, 11:263-282.),
- beta-lactamase (Sutcliffe (1978) Proc Natl Acad Sci USA 75:3737-3741), enzyme for a variety of chromogenic substrates (for example PADAC, a chromogenic cephalosporin),
- xylE gene product (Zukowsky et al. (1983) Proc Natl Acad Sci USA 80:1 101-1 105), catechol dioxygenase capable of converting chromogenic catechols,
- alpha-amylase (Ikuta et al. (1990) Bio/technol. 8:241-242),
- tyrosinase (Katz et al. (1983) J Gene Microbiol 129:2703-2714), enzyme which oxi-dizes tyro-
sine to give DOPA and dopaquinone which subsequently form melanine, which is readily de-
tectable,
-aequorin (Prasher et al. (1985) Biochem Biophys Res Commun 126(3):1259-1268), can be used in the calcium-sensitive bioluminescence detection.

**Target Organisms**

Any organism suitable for transformation or delivery of chimeric endonuclease can be used as target organism. This includes prokaryotes, eukaryotes, and archaea, in particular non-human organisms, plants, fungi or yeasts, but also human or animal cells.

In one embodiment the target organism is a plant.

The term "plant" includes whole plants, shoot vegetative organs/structures (e. g. leaves, stems and tubers), roots, flowers and floral organs/structures (e. g. bracts, sepals, petals, stamens, carpels, anthers and ovules), seeds (including embryo, endosperm, and seed coat) and fruits (the mature ovary), plant tissues (e. g. vascular tissue, ground tissue, and the like) and cells (e. g.
g. guard cells, egg cells, trichomes and the like), and progeny of same. The class of plants that can be used in the method of the invention is generally as broad as the class of higher and lower plants amenable to transformation techniques, including angiosperms (monocotyledonous and dicotyledonous plants), gymnosperms, ferns, and multicellular algae. It includes plants of a variety of ploidy levels, including aneuploid, polyploid, diploid, haploid and hemizygous.

Included within the scope of the invention are all genera and species of higher and lower plants of the plant kingdom. Included are furthermore the mature plants, seed, shoots and seedlings, and parts, propagation material (for example seeds and fruit) and cultures, for example cell cultures, derived therefrom.

Preferred are plants and plant materials of the following plant families: Amaranthaceae, Brassicaceae, Carophyllaceae, Chenopodiaceae, Compositae, Cucurbitaceae, Labiatae, Leguminosae, Papilionoideae, Liliaceae, Linaceae, Malvaceae, Rosaceae, Saxi-fragaceae, Scrophulariaceae, Solanaceae, Tetragoniaceae.

Annual, perennial, monocotyledonous and dicotyledonous plants are preferred host organisms for the generation of transgenic plants. The use of the recombination system, or method according to the invention is furthermore advantageous in all ornamental plants, useful or ornamental trees, flowers, cut flowers, shrubs or turf. Said plant may include - but shall not be limited to - bryophytes such as, for example, Hepaticae (hepaticas) and Musci (mosses); pteridophytes such as ferns, horsetail and club-mosses; gymnosperms such as conifers, cycads, ginkgo and Gnetaeae; algae such as Chlorophyceae, Phaeophyceae, Rhodophyceae, Myxophyceae, Xanthophyceae, Bacillariophyceae (diatoms) and Euglenophyceae.

Plants for the purposes of the invention may comprise the families of the Rosaceae such as rose, Ericaceae such as rhododendrons and azaleas, Euphorbiaceae such as poinsettias and croton, Caryophyllaceae such as pinks, Solanaceae such as petunias, Gesneriaceae such as African violet, Balsaminaceae such as touch-me-not, Orchida-ceae such as orchids, Iridaceae such as gladioli, iris, freesia and crocus, Compositae such as marigold, Geraniaceae such as geraniums, Liliaceae such as dracaena, Moraceae such as ficus, Araceae such as philodendron and many others.

The transgenic plants according to the invention are furthermore selected in particular from among dicotyledonous crop plants such as, for example, from the families of the Leguminosae such as pea, alfalfa and soybean; Solanaceae such as tobacco and and many others; the family of the Umbelliferae, particularly the genus Daucus (very particularly the species carota (carrot)) and Apium (very particularly the species graveolens dulce (celery)) and many others; the family of the Solanaceae, particularly the genus Lycopersicon, very particularly the species esculentum (tomato) and the genus Solanum, very particularly the species tuberosum (potato) and melongena (au-bergine) and many others; and the genus Capsicum, very particularly the species an-num (pepper) and many others; the family of the Leguminosae, particularly the genus
Glycine, very particularly the species max (soybean) and many others; and the family of the
Cruciferae, particularly the genus Brassica, very particularly the species napus (oilseed rape),
campestris (beet), oleracea cv Tastie (cabbage), oleracea cv Snowball Y (cauliflower) and o-
eracea cv Emperor (broccoli); and the genus Arabidopsis, very particularly the species thaliana
and many others; the family of the Compositae, particularly the genus Lactuca, very particularly
the species sativa (lettuce) and many others.

The transgenic plants according to the invention are selected in particular among monocotyle-
donous crop plants, such as, for example, cereals such as wheat, barley, sorghum and millet,
rye, triticale, maize, rice or oats, and sugar cane.

Especially preferred are Arabidopsis thaliana, Nicotiana tabacum, oilseed rape, soybean, corn
(maize), wheat, linseed, potato and tagetes.

Plant organisms are furthermore, for the purposes of the invention, other organisms which are
capable of photosynthetic activity, such as, for example, algae or cyanobacteria, and also
mosses. Preferred algae are green algae, such as, for example, algae of the genus Haemato-
coccus, Phaedactylum tricornatum, Volvox or Dunaliella.

Genetically modified plants according to the invention which can be consumed by humans or
animals can also be used as food or feedstuffs, for example directly or following processing
known in the art.

Construction of Polynucleotide Constructs.

Typically, polynucleotide constructs (e.g., for an expression cassette) to be introduced into non-
human organism or cells, e.g. plants or plant cells are prepared using transgene expression
techniques. Recombinant expression techniques involve the construction of recombinant nu-
cleic acids and the expression of genes in transfected cells. Molecular cloning techniques to
achieve these ends are known in the art. A wide variety of cloning and in vitro amplification
methods suitable for the construction of recombinant nucleic acids are well-known to persons of
skill in the art. Examples of these techniques and instructions sufficient to direct persons of skill
in the art through many cloning exercises are found in Berger and Kimmel, Guide to Molecular
(Berger); Current Protocols in Molecular Biology, F. M. Ausubel et al., eds., Current Protocols,
Supplement), T. Maniatis, E.F. Fritsch and J. Sambrook, Molecular Cloning: A Laboratory Man-
ual, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY (1989), in T.J. Silhavy, M.L. Ber-
man and L.W. Enquist, Experiments with Gene Fusions, Cold Spring Harbor Laboratory, Cold
Spring Harbor, NY (1984). Preferably, the DNA constructs employed in the invention are gener-
ated by joining the abovementioned essential constituents of the DNA construct together in the
abovementioned sequence using the recombination and cloning techniques with which the
skilled worker is familiar.
The construction of polynucleotide constructs generally requires the use of vectors able to replicate in bacteria. A plethora of kits are commercially available for the purification of plasmids from bacteria. The isolated and purified plasmids can then be further manipulated to produce other plasmids, used to transiently or stably transform cells or incorporated into Agrobacterium tumefaciens or Agrobacterium rhizogenes to infect and transform plants. Where Agrobacterium is the means of transformation, shuttle vectors are constructed.

Methods for Introducing Constructs into Target Cells
A DNA construct employed in the invention may advantageously be introduced into cells using vectors into which said DNA construct is inserted. Examples of vectors may be plasmids, cosmids, phages, viruses, retroviruses or agrobacteria. In an advantageous embodiment, the expression cassette is introduced by means of plasmid vectors. Preferred vectors are those which enable the stable integration of the expression cassette into the host genome.

A DNA construct can be introduced into the target plant cells and/or organisms by any of the several means known to those of skill in the art, a procedure which is termed transformation (see also Keown et al. (1990) Meth Enzymol 185:527-537). For instance, the DNA constructs can be introduced into cells, either in culture or in the organs of a plant by a variety of conventional techniques. For example, the DNA constructs can be introduced directly to plant cells using ballistic methods, such as DNA particle bombardment, or the DNA construct can be introduced using techniques such as electroporation and microinjection of cells. Particle-mediated transformation techniques (also known as "biolistics") are described in, e.g., Klein et al. (1987) Nature 327:70-73; Vasil V et al. (1993) BiolTechnol 11:1553-1 558; and Becker D et al. (1994) Plant J 5:299-307. These methods involve penetration of cells by small particles with the use of a DNA-coated gold or tungsten microcarriers toward target cells. The process is applicable to a wide range of tissues and cells from organisms, including plants. Other transformation methods are also known to those of skill in the art.

Microinjection techniques are known in the art and are well described in the scientific and patent literature. Also, the cell can be permeabilized chemically, for example using polyethylene glycol, so that the DNA can enter the cell by diffusion. The DNA can also be introduced by protoplast fusion with other DNA-containing units such as minicells, cells, lysosomes or liposomes. The introduction of DNA constructs using polyethylene glycol (PEG) precipitation is described in Paszkowski et al. (1984) EMBO J 3:2717. Liposome-based gene delivery is e.g., described in WO 93/24640; Mannino and Gould-Fogerite (1988) BioTechniques 6(7):682-691 ; US 5,279,833; WO 91/06309; and Feigner et al. (1987) Proc Natl Acad Sci USA 84:7413-7414).

Another suitable method of introducing DNA is electroporation, where the cells are permeabilized reversibly by an electrical pulse. Electroporation techniques are described in Fromm et al. (1985) Proc Natl Acad Sci USA 82:5824. PEG-mediated transformation and electroporation of
general methods which may be mentioned are the calcium-phosphate-mediated transfection,
the DEAE-dextran-mediated transfection, the cationic lipid-mediated transfection, electropora-
tion, transduction and infection. Such methods are known to the skilled worker and described,
for example, in Davis et al., Basic Methods In Molecular Biology (1986). For a review of gene
transfer methods for plant and cell cultures, see, Fisk et al. (1993) Scientia Horticul-turae 55:5-

Methods are known for introduction and expression of heterologous genes in both monocot and
particular can use various techniques including electroporation (e.g., Shimamoto et al. (1992)
Nature 338:274-276; biolistics (e.g., EP-A1 270,356); and Agrobacterium (e.g., Bytebier et al.

In plants, methods for transforming and regenerating plants from plant tissues or plant cells with
which the skilled worker is familiar are exploited for transient or stable transformation. Suitable
methods are especially protoplast transformation by means of poly-ethylene-glycol-induced
DNA uptake, biolistic methods such as the gene gun ("particle bombardment" method), electro-
poration, the incubation of dry embryos in DNA-containing solution, sonication and microinj-
etion, and the transformation of intact cells or tissues by micro- or macroinjection into tissues or
embryos, tissue electroporation, or vacuum infiltration of seeds. In the case of injection or elec-
 troporation of DNA into plant cells, the plasmid used does not need to meet any particular re-
quirement. Simple plasmids such as those of the pUC series may be used. If intact plants are to
be regenerated from the transformed cells, the presence of an additional selectable marker
gene on the plasmid is useful.

In addition to these "direct" transformation techniques, transformation can also be carried out by
bacterial infection by means of Agrobacterium tumefaciens or Agrobacterium rhizogenes. These
strains contain a plasmid (Ti or Ri plasmid). Part of this plasmid, termed T-DNA (transferred
DNA), is transferred to the plant following Agrobacterium infection and integrated into the ge-
nome of the plant cell.

For Agrobacterium-mediated transformation of plants, a DNA construct of the invention may be
combined with suitable T-DNA flanking regions and introduced into a conventional Agrobacte-
rion tumefaciens host vector. The virulence functions of the A. tumefaciens host will direct the
insertion of a transgene and adjacent marker gene(s) (if present) into the plant cell DNA when
the cell is infected by the bacteria. Agrobacterium tumefaciens-mediated transformation tech-
niques are well described in the scientific literature. See, for example, Horsch et al. (1984) Sci-
L'Academie Des Sciences Serie Ill-Sciences De La Vie-Life Sciences 316:1 194-1 199, Valve-

A DNA construct of the invention is preferably integrated into specific plasmids, either into a shuttle, or intermediate, vector or into a binary vector). If, for example, a Ti or Ri plasmid is to be used for the transformation, at least the right border, but in most cases the right and the left border, of the Ti or Ri plasmid T-DNA is linked with the expression cassette to be introduced as a flanking region. Binary vectors are preferably used. Bi-nary vectors are capable of replication both in E. coli and in Agrobactenum. As a rule, they contain a selection marker gene and a linker or polylinker flanked by the right or left T-DNA flanking sequence. They can be transformed directly into Agrobacterium (Holsters et al. (1978) Mol Gen Genet 163:181-187). The selection marker gene permits the selection of transformed agrobacteria and is, for example, the nptll gene, which imparts resistance to kanamycin. The Agrobacterium, which acts as host organism in this case, should already contain a plasmid with the vir region. The latter is required for transferring the T-DNA to the plant cell. An Agrobacterium thus transformed can be used for transforming plant cells.

Many strains of Agrobacterium tumefaciens are capable of transferring genetic material - for example a DNA constructs according to the invention -, such as, for example, the strains EHA101 (pEHA101) (Hood EE et al. (1996) J Bacteriol 168(3):1291-1301), EHA105(pEHA105) (Hood et al. 1993, Transgenic Research 2, 208-218), LBA4404(pAL4404) (Hoekema et al. (1983) Nature 303:1 79-181), C58C1 (pMP90) (Koncz and Schell (1986) Mol Gen Genet 204,383-396) and C58C1(pGV2260) (De-blaere et al. (1985) Nucl Acids Res. 13, 4777-4788).

The agrobacterial strain employed for the transformation comprises, in addition to its disarmed Ti plasmid, a binary plasmid with the T-DNA to be transferred, which, as a rule, comprises a gene for the selection of the transformed cells and the gene to be transferred. Both genes must be equipped with transcriptional and translational initiation and termination signals. The binary plasmid can be transferred into the agrobacterial strain for example by electroporation or other transformation methods (Mozo & Hooykaas (1991) Plant Mol Biol 16:91 7-918). Coculture of the plant explants with the agrobacterial strain is usually performed for two to three days.

A variety of vectors could, or can, be used. In principle, one differentiates between those vectors which can be employed for the Agrobacterium-mediated transformation or agroinfection, i.e. which comprise a DNA construct of the invention within a T-DNA, which indeed permits stable integration of the T-DNA into the plant genome. Moreover, border-sequence-free vectors may be employed, which can be transformed into the plant cells for example by particle bombardment, where they can lead both to transient and to stable expression.

(1985) EMBO J 4:277-287). Various binary vectors are known, some of which are commercially available such as, for example, pBIN19 (Clontech Laboratories, Inc. USA).

To transfer the DNA to the plant cell, plant explants are cocultured with Agrobacterium tumefaciens or Agrobacterium rhizogenes. Starting from infected plant material (for example leaf, root or stalk sections, but also protoplasts or suspensions of plant cells), intact plants can be regenerated using a suitable medium which may contain, for example, antibiotics or biocides for selecting transformed cells. The plants obtained can then be screened for the presence of the DNA introduced, in this case a DNA construct according to the invention. As soon as the DNA has integrated into the host genome, the genotype in question is, as a rule, stable and the insertion in question is also found in the subsequent generations. As a rule, the expression cassette integrated contains a selection marker which confers a resistance to a biocide (for example a herbicide) or an antibiotic such as kanamycin, G 418, bleomycin, hygromycin or phosphinotricin and the like to the transformed plant. The selection marker permits the selection of transformed cells (McCormick et al., Plant Cell Reports 5 (1986), 81-84). The plants obtained can be cultured and hybridized in the customary fashion. Two or more generations should be grown in order to ensure that the genomic integration is stable and hereditary.


The DNA construct of the invention can be used to confer desired traits on essentially any plant. One of skill will recognize that after DNA construct is stably incorporated in transgenic plants and confirmed to be operable, it can be introduced into other plants by sexual crossing. Any of a number of standard breeding techniques can be used, depending upon the species to be crossed.

The nuclease s or chimeric endonuclease may alternatively be expressed transiently. The chimeric endonuclease may be transiently expressed as a DNA or RNA delivered into the target cell and/or may be delivered as a protein. Delivery as a protein may be achieved with the help of cell penetrating peptides or by fusion with SEciV signal peptides fused to the nuclease s or chimeric endonucleases, which mediate the secretion from a delivery organism into a cell of a target organism e.g. from Agrobacterium rhizogenes or Agrobacterium tumefaciens to a plant cell.

Regeneration of Transgenic Plants
Transformed cells, i.e. those which comprise the DNA integrated into the DNA of the host cell, can be selected from untransformed cells if a selectable marker is part of the DNA introduced. A
marker can be, for example, any gene which is capable of conferring a resistance to antibiotics or herbicides (for examples see above). Transformed cells which express such a marker gene are capable of surviving in the presence of concentrations of a suitable antibiotic or herbicide which kill an untransformed wild type. As soon as a transformed plant cell has been generated, an intact plant can be obtained using methods known to the skilled worker. For example, callus cultures are used as starting material. The formation of shoot and root can be induced in this as yet undifferentiated cell biomass in the known fashion. The shoots obtained can be planted and cultured.

Transformed plant cells, derived by any of the above transformation techniques, can be cultured to regenerate a whole plant which possesses the transformed genotype and thus the desired phenotype. Such regeneration techniques rely on manipulation of certain phytohormones in a tissue culture growth medium, typically relying on a biocide and/or herbicide marker that has been introduced together with the desired nucleotide sequences. Plant regeneration from cultured protoplasts is described in Evans et al., Protoplasts Isolation and Culture, Handbook of Plant Cell Culture, pp. 124176, Macmillan Publishing Company, New York (1983); and in Binding, Regeneration of Plants, Plant Protoplasts, pp. 21-73, CRC Press, Boca Raton, (1985). Regeneration can also be obtained from plant callus, explants, somatic embryos (Dandekar et al. (1989) J Tissue Cult Meth 12:145; McGranahan et al. (1990) Plant Cell Rep 8:512), organs, or parts thereof. Such regeneration techniques are described generally in Klee et al. (1987) Ann Rev Plant Physiol 38:467-486.

Combination with other recombination enhancing techniques
In a further preferred embodiment, the efficacy of the recombination system is increased by combination with systems which promote homologous recombination. Such systems are described and encompass, for example, the expression of proteins such as RecA or the treatment with PARP inhibitors. It has been demonstrated that the intrachromosomal homologous recombination in tobacco plants can be increased by using PARP inhibitors (Puchta H et al. (1995) Plant J. 7:203-210). Using these inhibitors, the homologous recombination rate in the recombination cassette after induction of the sequence-specific DNA double-strand break, and thus the efficacy of the deletion of the transgene sequences, can be increased further. Various PARP inhibitors may be employed for this purpose. Preferably encompassed are inhibitors such as 3-amino-nobenza mede, 8-hydroxy-2-methylquinazoloin-4-one (NU1025), 1,1b-dihydro-(2H)benzopyrano(4,3,2-de)isoquinolin-3-one (GPI 6150), 5-aminoisoquino-linone, 3,4-dihydro-5-(4-(1-piperidiny)ibutoxy)-1 (2H)-isoquinolinone, or the compounds described in WO 00/26192, WO 00/29384, WO 00/32579, WO 00/64878, WO 00/68206, WO 00/67734, WO 01/23386 and WO 01/23390.

In addition, it was possible to increase the frequency of various homologous recombination reactions in plants by expressing the E. coli RecA gene (Reiss B et al. (1996) Proc Natl Acad Sci USA 93(7):3094-3098). Also, the presence of the protein shifts the ratio between homologous and illegitimate DSB repair in favor of homologous repair (Reiss B et al. (2000) Proc Natl Acad
Sci USA 97(7):3358-3363). Reference may also be made to the methods described in WO 97/08331 for increasing the homologous recombination in plants. A further increase in the efficacy of the recombination system might be achieved by the simultaneous expression of the RecA gene or other genes which increase the homologous recombination efficacy (Shalev G et al. (1999) Proc Natl Acad Sci USA 96(13):7398-402). The above-stated systems for promoting homologous recombination can also be advantageously employed in cases where the recombination construct is to be introduced in a site-directed fashion into the genome of a eukaryotic organism by means of homologous recombination.

Methods of providing chimeric endonucleases:

The current invention provides a method of providing a chimeric endonuclease as described above.

The method comprises the steps of:

a. providing at least one endonuclease coding region
b. providing at least one heterologous DNA binding domain coding region,
c. providing a polynucleotide having a potential DNA recognition sequence or potential DNA recognition sequences of the endonuclease or endonucleases of step a) and having a potential recognition sequence or having potential recognition sequences of the heterologous DNA binding domain or heterologous DNA binding domains of step b),
d. creating a translational fusion of all endonuclease coding regions of step b) and all heterologous DNA binding domains of step c),
e. expressing a chimeric endonuclease from the translational fusion created in step d),
f. testing the chimeric endonuclease expressed in step e) for cleavage of the polynucleotide of step c).

Depending on the intended purpose, the method steps a), b), c) and d) can be used in varying order. For example, the method can be used to provide a particular combination of at least one endonuclease and at least one heterologous DNA binding domain and providing thereafter a polynucleotide comprising potential DNA recognition sites and potential recognition sites reflecting the order in which the at least one nuclease and the at least one heterologous DNA binding site were arranged in the translational fusion, and testing the chimeric endonuclease for cleaving activity on a polynucleotide having potential DNA recognition sites and potential recognition sites for the nucleases and heterologous DNA binding domains comprised by the chimeric endonuclease and selecting at least one polynucleotide that is cut by the chimeric endonuclease.

The method can also be used to design a chimeric endonuclease for cleaving activity on a pre-selected polynucleotide, by first providing a polynucleotide having a specific sequence, thereafter selecting at least one endonuclease and at least one heterologous DNA binding domain having non-overlapping potential DNA recognition sites and potential recognition sites in the nucleotide sequence of the polynucleotide, creating a translational fusion of the at least one endonuclease and the at least one heterologous DNA binding domain, expressing the chimeric
endonuclease encoded by said translational fusion and testing the chimeric endonuclease of cleavage activity on the preselected polynucleotide sequence, and selecting a chimeric endonuclease having such cleavage activity.

This method can be used to design a chimeric endonuclease having an enhanced cleavage activity on a specific polynucleotide, for example, if a polynucleotide comprises a DNA recognition site of a nuclease it will be possible to identify a potential recognition site of a heterologous DNA binding domain, which can be used to create a chimeric endonuclease comprising the nuclease and the heterologous DNA binding domain.

Alternatively, this method can also be used to create a chimeric endonuclease having cleavage activity on a specific polynucleotide comprising a recognition site of a heterologous DNA binding domain. For example, in case the specific polynucleotide is known to be bound by a heterologous DNA binding domain, e.g. a particular transcription factor or a virulence factor of a pathogen having a specific DNA binding activity, like Tal-Type Effector proteins or there repeat units in particular Tal-Type III Effector proteins of Xanthomonas species, it is possible to identify a endonuclease having a potential DNA recognition site close to but not overlapping with the recognition site of the identified heterologous DNA binding domain. By creating a translational fusion and expressing the chimeric endonuclease comprising the identified endonuclease and the heterologous DNA binding domain, it will be possible to test the chimeric endonuclease for cleavage activity on said preselected polynucleotide.

Suitable endonucleases and heterologous DNA binding domains can be identified by searching databases comprising DNA recognition sites of endonucleases and recognition sites of DNA binding proteins like transcription factors or virulence factors.

Further, it is possible to mutate the amino acid sequence of endonucleases, like l-Scel, l-Crel, l-Dmol or l-Msol to create new binding and DNA cleavage activity. Similar techniques are available to create new binding activities of zink-finger comprising proteins or virulence factors of the Tal-Type III Effector proteins of Xanthomonas species, which can be used as heterologous DNA binding domains. By creating chimeric endonucleases comprising endonucleases like l-Scel, l-Crel, l-Dmol or l-Msol and heterologous DNA binding domains derived from or comprising zink-finger proteins or Tal-Type III Effector proteins of Xanthomonas species in combination with mutational techniques to adapt their DNA binding activity to the sequence of preselected polypeptides, it is possible to create chimeric endonucleases which will bind and cleave such preselected polypeptides.

Accordingly one embodiment of the invention comprises chimeric endonucleases comprising

a) at least one endonuclease selected from the group of l-Scel, l-Crel, l-Dmol or l-Msol or homologs of l-Scel, l-Crel, l-Dmol or l-Msol having at least 80%, 85%, 90% 95%, 96%, 97%, 98% or 99% sequence identity, and

b) a heterologous DNA binding domain comprising either at least one zink finger protein or comprising at least one Tal-Type III Effector protein of Xanthomonas species or comprising at least one zink finger protein and comprising at least one Tal-Type III Effector pro-
tein of Xanthomonas species or comprising at least one homolog of zink finger proteins or Tal-Type III Effector proteins of Xanthomonas species having at least 80%, 85%, 90% 95%, 96%, 97%, 98% or 99% sequence identity.

The cleavage activity of endonucleases and chimeric endonucleases as well as the DNA binding activity of endonucleases, heterologous DNA binding domains and chimeric endonucleases can be tested by in vitro and in vivo techniques known in the art. For example by techniques as disclosed in the examples herein.

Methods for homologous recombination and targeted mutation using chimeric endonucleases.

The current invention provides a method for homologous recombination of polynucleotides comprising:

a. providing a cell competent for homologous recombination,

b. providing a polynucleotide comprising a recombinant polynucleotide flanked by a sequence A and a sequence B,

c. providing a polynucleotide comprising sequences A and B', which are sufficiently long and homologous to sequence A and sequence B, to allow for homologous recombination in said cell and

d. providing a chimeric endonuclease or an expression cassette coding for a chimeric endonuclease,

e. combining b), c) and d) in said cell and

f. detecting recombined polynucleotides of b) and c), or selecting for or growing cells comprising recombined polynucleotides of b) and c).

In one embodiment of the invention, the polynucleotide provided in step b) comprises at least one chimeric recognition site, preferably a chimeric recognition site selected from the group of sequences described by SEQ ID NO: 14, 15, 16, 17, 18, 19 or 20.

In one embodiment of the invention, the polynucleotide provided in step c) comprises at least one chimeric recognition site, preferably selected from the group of sequences described by SEQ ID NO: 14, 15, 16, 17, 18, 19 or 20.

In one embodiment of the invention, the polynucleotide provided in step b) and the polynucleotide provided in step c) comprise at least one chimeric recognition site, preferably selected from the group of sequences described by SEQ ID NO: 14, 15, 16, 17, 18, 19 or 20.

In one embodiment of the invention, step e) leads to deletion of a polynucleotide comprised in the polynucleotide provided in step c).

In one embodiment of the invention the deleted polynucleotide comprised in the polynucleotide provided in step c) codes for a marker gene or parts of a marker gene.

In one embodiment of the invention, the polynucleotide provided in step b) comprises at least one expression cassette.
one expression cassette, leading to expression of a selection marker gene or a reporter gene. In one embodiment of the invention, the polynucleotide provided in step b) comprises at least one expression cassette, leading to expression of a selection marker gene or a reporter gene and comprises at least one DNA recognition site or at least one chimeric recognition site.

A further embodiment of the invention provides a method for targeted mutation of polynucleotides comprising:

a. providing a cell comprising a polynucleotide comprising a chimeric recognition site,
b. providing a chimeric endonuclease, e.g. an chimeric endonuclease comprising an endonuclease having a sequence selected from the group of sequences described by SEQ ID NO: 2, 3, or 5, and being able to cleave the chimeric recognition site of step a),
c. combining a) and b) in said cell and

d. detecting mutated polynucleotides, or selecting for growing cells comprising mutated polynucleotides.

The invention provides in another embodiment a method for homologous recombination as described above or a method for targeted mutation of polynucleotides as described above, comprising:

combining the chimeric endonuclease and the chimeric recognition site via crossing of organisms, via transformation of cells or via a SecIV peptide fused to the chimeric endonuclease and contacting the cell comprising the chimeric recognition site with an organism expressing the chimeric endonuclease and expressing a SecIV transport complex able to recognize the SecIV peptide fused to the chimeric endonuclease.

Examples

General methods:

The chemical synthesis of oligonucleotides can be effected for example in the known manner using the phosphoamidite method (Voet, Voet, 2nd edition, Wiley Press New York, pages 896-897). The cloning steps carried out for the purposes of the present invention, such as, for example, restriction cleavages, agarose gel electrophoresis, purification of DNA fragments, the transfer of nucleic acids to nitrocellulose and nylon membranes, the linkage of DNA fragments, the transformation of E.coli cells, bacterial cultures, the propagation of phages and the sequence analysis of recombinant DNA are carried out as described by Sambrook et al. (1989) Cold Spring Harbor Laboratory Press; ISBN 0-87969-309-6. Recombinant DNA molecules were sequenced using an ALF Express laser fluorescence DNA sequencer (Pharmacia, Upsala [sic], Sweden) following the method of Sanger (Sanger et al., Proc. Natl. Acad. Sci. USA 74 (1977), 5463-5467).

Example 1: Constructs harboring sequence specific DNA-endonuclease expression cassettes for expression in *E.coli*
Example 1a: Basic construct
In this example we present the general outline of a vector, named "Construct I" suitable for transformation in E. coli. This general outline of the vector comprises an ampicillin resistance gene for selection, a replication origin for E. coli and the gene araC, which encodes an Arabinose inducible transcription regulator. Different genes, encoding the different versions of the sequence specific DNA-endonuclease, can be expressed from the Arabinose inducible pBAD promoter (Guzman et al., J Bacteriol 177: 4121-4130(1995)). The sequences of the genes encoding the different nuclease versions are given in the following examples.
The control construct, in which encodes the sequence of l-Scel (SEQ ID NO: 22), was called VC-SAH40-4.

Example 1b: scTet - l-Scel fusion constructs
In JOURNAL OF BACTERIOLOGY 150(2), 633-642 (1982) Beck et al. described the TetR protein. TetR acts as a dimer, but single chain variants (scTetR) are well described in NUCLEIC ACIDS RESEARCH 31(12), 3050-3056 (2003) by Krueger et al. The scTetR encoding sequence was fused to l-Scel, with a single lysine as a short. The linker was designed in a way that the resulting fusion protein recognizes a cognate binding site, which represents a combination of the binding sites of l-Scel and TetR. TetR is a transcriptional repressor, which binds to the DNA in absence of the inducer. It is displaced from the recognition sequence in the presence of tetracycline. This could provide the potential to regulate the activity or DNA binding affinity of the fusion protein in the same manner. The resulting plasmid was called VC-SAH54-4.
The sequence of the construct is identical to the sequence of construct I, whereas the nuclease encoding gene was replaced by the sequence described by SEQ ID NO: 23.
A similar construct was generated, which in addition to the latter contains a NLS sequence. The resulting plasmid was called VC-SAH53-10. The sequence of the construct is identical to the sequence of construct I, whereas the nuclease encoding gene was replaced by the sequence described by SEQ ID NO: 24.

Example 1c: scArc - l-Scel fusion constructs
In J Mol Biol 185 (2), 445-6 (1985) Jordan et al. described the crystallization of the Arc Repressor of Salmonella phage P22 Arc. It is active as a dimer, but single chain variants (scArc) are described in Biochemistry 35 (1), 109-16 (1996) by Robinsons et al.. The coding sequence for this single chain variant was fused to l-Scel, with a linker that encompasses a NLS. The linker having the amino acid sequence: RSGGGSGGGTGGGSGGGAPKKKRKVLE (SEQ ID NO: 151) was designed in a way that the resulting fusion protein recognizes a cognate binding site, which represents a combination of the binding sites of l-Scel and Arc. The resulting plasmid was called VC-SAH28-5. The sequence of the construct is identical to the sequence of construct I, whereas the encoded gene is described by SEQ ID NO: 25. Also a fusion with a shorter linker the linker having the amino acid sequence: RSAPKKKRKVLE (SEQ ID NO: 152) between scArc and l-Scel was generated, which still encompasses a NLS. The resulting plasmid was called VC-SAH46-4. The sequence of the construct is identical to the sequence of Construct I, whereas the encoded gene is described by SEQ ID NO: 26.
Example 2: Constructs harboring nuclease recognition sequences/target sites to monitor l-Scel activity in *E. coli*

Example 2a: Basic construct

In this example we present the general outline of a vector, named "Construct II" suitable for transformation in *E. coli*. This general outline of the vector comprises a Kanamycin resistance gene for selection, a replication origin for *E. coli*, which is compatible with the ori of Construct I. SEQ ID NO: 27 shows a sequence stretch of "NNNNNNNNNN". This is meant to be a placeholder for different recognition/target sites for the diverse versions and protein fusions of the sequence specific DNA-endonucleases. The control construct, in which the placeholder is replaced by a sequence stretch encompassing the native target sequence of l-Scel (SEQ ID NO: 28), was called VC-SAH6-1. A control plasmid without a target site was called VC-SAH7-1 (SEQ ID NO 29)

The different combined target sites are given in the following examples.

Example 2b: target sites combined of l-Scel recognition sequence and scTet binding sequence

Combined target sites were generated, that consist of the target site of the nuclease l-Scel and TetR. Different combined target sites with varying distances of the single sites were generated. The goal was to identify the one that is best recognized by the cognate l-Scel fusion protein. The resulting plasmids were called VC-SAH60-5, VC-SAH61-1, VC-SAH62-1. The sequence of the constructs is identical to the sequence of Construct II, whereas the sequence "NNNNNNNNNN" was replaced by the sequences described by SEQ ID NO: 30, NO: 31, NO: 32, respectively.

Example 2c: target sites combined of l-Scel recognition sequence and scArc binding sequence

In *PNAS* 96, 811-817 ^999) Schildbach et al. described the Arc Protein in contact with its cognate recognition sequence. Combined target sites were generated, that consist of the target site of the nuclease l-Scel and Arc, with varying distances. The goal is to identify the one that is best recognized by the cognate l-Scel fusion protein. The resulting plasmids are called VC-SAH132-1, VC-SAH133-8, VC-SAH134-1 and VC-SAH135-1. The sequences of these plasmids is identical to the sequence of Construct III (SEQ ID NO: 33), where the sequence "NNNNNNNNNN" is replaced by the sequences consisting of different versions of the combined target sites, described by SEQ ID NO: 34, NO: 35, NO: 36, NO: 37 respectively.

Example 3: Cotransformation of DNA endonuclease encoding constructs and constructs harboring nuclease recognition sequences

Two plasmids with different selection markers and identical concentrations were transformed in chemical competent *E. coli* Top10 cells, according to the manufacturer description. The cells were plated on LB with the respective antibiotics for selection, and grown over night at 37°C. With this method constructs harboring sequence specific DNA-endonuclease expression cas-
settes and cognate constructs harboring nuclease recognition sequences/target sites were combined in the same transformant to allow monitoring of the nuclease activity.

Example 4: Demonstration of the endonuclease activity in *E. coli*

Cotransformants which carry the combination of two plasmids, one encoding a nuclease or a nuclease-fusion (Construct I) and the other one harboring a compatible target site (Construct II) were grown over night in LB with Ampicillin and Kanamycin. The cultures were diluted 1:100 and grown until they reached \( \text{OD}_{600} = 0.5 \). The expression of the fusion protein from Construct I was induced by addition of Arabinose for 3 to 4 hours. The pBAD promoter is described to be dose dependent (Guzman 1995), therefore the culture was divided in different aliquots and protein expression was induced with Arabinose concentrations varying from 0.2% to 0.0002%. 5 \( \mu \)l of each aliquot were plated on LB solid media, supplemented with Ampicillin and Kanamycin. The plates were incubated over night at 37 °C and cell growth was analyzed semi quantitatively. Active nuclease fusions did cut the constructs, which harbor the target site. This led to the loss of Construct II or Construct III, which confer Kanamycin resistance. Therefore, activity of the fusion protein was observed due to the lost ability of the cotransformants to grow on Kanamycin containing medium.

RESULTS:

The result are simplified and summarized in Table 9. ++ and + represent very strong and strong growth, which indicates no or little activity of the expressed nuclease towards the respective target site. - and — represent reduced or no growth, which indicates high or very high activity of the nuclease towards the respective target site.

Table 9: l-Scel - scTet fusions: *E. coli* growth assay indicates endonuclease activity (enzymatic activity) against the respective target sites.

<table>
<thead>
<tr>
<th>VC-SAH740-4</th>
<th>VC-SA545-4</th>
<th>VC-SA535-10</th>
</tr>
</thead>
<tbody>
<tr>
<td>VC-SA740-1</td>
<td>++</td>
<td>++</td>
</tr>
<tr>
<td>VC-SA640-1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>VC-SA60-5</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>VC-SA61-1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>VC-SA62-1</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Example 5: Transformation of *Arabidopsis thaliana*

*A. thaliana* plants were grown in soil until they flowered. *Agrobacterium tumefaciens* (strain C58C1 [pMP90]) transformed with the construct of interest was grown in 500 mL in liquid YEB medium (5 g/L Beef extract, 1 g/L Yeast Extract (Duchefa), 5 g/L Peptone (Duchefa), 5 g/L sucrose (Duchefa), 0.49 g/L MgSO₄ (Merck)) until the culture reached an \( \text{OD}_{600} \) 0.8-1.0. The bacterial cells were harvested by centrifugation (15 minutes, 5,000 rpm) and resuspended in 500 mL infiltration solution (5% sucrose, 0.05% SILWET L-77 [distributed by Lehle seeds, Cat. No.
Flowering plants were dipped for 10-20 seconds into the Agrobacterium solution. Afterwards the plants were kept in the dark for one day and then in the greenhouse until seeds could be harvested. Transgenic seeds were selected by plating surface sterilized seeds on growth medium A (4.4g/L MS salts [Sigma-Aldrich], 0.5g/L MES [Duchefa]; 8g/L Plant Agar [Duchefa]) supplemented with 50 mg/L kanamycin for plants carrying the nptII resistance marker gene, and 10 mg/L Phosphinotricin for plants carrying the pat gene, respectively. Surviving plants were transferred to soil and grown in the greenhouse.

Example 6: Constructs harbouring sequence specific DNA-endonuclease expression cassettes for A. thaliana

Example 6a: Basic construct
In this example we present the general outline of a binary vector, named "Construct IV" suitable for plant transformation. This general outline of the binary vector comprises a T-DNA with a p-Mas1del100::cBAR::t-Ocs1 cassette, which enables selection on Phosphinotricin, when integrated into the plant genome. SEQ ID NO: 38 shows a sequence stretch of "NNNNNNNNNN". This is meant to be a placeholder for genes encoding the different versions of the sequence specific DNA-endonuclease. The sequence of the latter is given in the following examples.

Example 6b: scTet - l-Scel fusion constructs
The sequence stretch of "NNNNNNNNNN" of construct IV is separately replaced by genes encoding the different versions of l-Scel -scTet fusions. The scTetR encoding sequence was fused to l-Scel, with a short linker, as described in Example 1 c). The resulting plasmid is called VC-SAHI40. The sequence of the construct is identical to the sequence of construct IV, whereas the sequence "NNNNNNNNNN" is replaced by the sequence described in Example 1.

Example 6c: scArc - l-Scel fusion constructs
The sequence stretch of "NNNNNNNNNN" of construct IV was separately replaced by genes encoding the different versions of l-Scel -scArc fusions. The scArc encoding sequence was fused to l-Scel, as described in Example 1 d). The resulting plasmid was called VC-SAHI89.10.

Example 7: Constructs harboring nuclease recognition sequences/target sites to monitor nucle-
ase activity in *A. thaliana*

Example 7a: Basic construct

In this example we present the general outline of a binary vector, named "Construct V", suitable for transformation in *A. thaliana*. This general outline of the vector comprises a T-DNA with a nos-promoter::nptII::nos-terminator cassette, which confers kanamycin resistance when integrated into the plant genome.

The T-DNA also comprises a partial *uidA* (GUS) gene (called "GU") and another partial *uidA* gene (called "US"). Between GU and US a stretch of "NNNNNNNNNN" is shown in SEQ ID NO: 39. This is meant to be a placeholder for different recognition/target sites for the diverse versions and protein fusions of the sequence specific DNA-endonucleases. The sequences of the different target sites are given in the following examples.

If the recognition sequence is cut by the respective nuclease, the partially overlapping and non-functional halves of the GUS gene (GU and US) will be restored as a result of intrachromosomal homologous recombination (ICHR). This can be monitored by histochemical GUS staining (Jefferson 1985).

Example 7b: Target sites combined of nuclease recognition sequence and scTet binding sequence

Combined target sites are generated, that consist of the target site of the nuclease l-Scel and TetR. Different combined target sites with varying distances of the single sites are generated. The goal is to identify the one that is best recognized by the cognate l-Scel fusion protein. The resulting plasmids are called VC-SAH1 13, VC-SAH1 14, VC-SAH1 15. The sequence of the constructs is identical to the sequence of Construct II, whereas the sequence "NNNNNNNNNN" is replaced by the sequences described by SEQ ID NO: 40, NO: 41, NO: 42, respectively.

Example 7c: Target sites combined of nuclease recognition sequence and scArc binding sequence

Combined target sites were generated, that consist of the target site of the nuclease l-Scel and Arc. Different combined target sites with varying distances of the single sites were generated. The goal was to identify the one that is best recognized by the cognate l-Scel fusion protein. The resulting plasmids were called VC-SAH16-4, VC-SAH17-8, VC-SAH18-7, VC-SAH19-15. The sequence of the constructs is identical to the sequence of Construct V, whereas the sequence "NNNNNNNNNN" was replaced by the sequences described by SEQ ID NO: 43, NO: 44, NO: 45, NO: 46 respectively.

Example 8: Transformation of sequence-specific DNA endonuclease encoding constructs into *A. thaliana*

Plasmids VC-SAH87-4 VC-SAH140, VC-SAH139-20, VC-SAH89-10, VC-SAH90 were/ are transformed into *A. thaliana* according to the protocol described in Example 5. Selected transgenic lines (T1 generation) are grown in the greenhouse and some flowers will be used for crossings (see below).
Example 9: Transformation of constructs harboring combined target sites to monitor recombination into *A. thaliana*

Plasmids VC-SAH1 11, VC-SAH1 12, VC-SAH1 13, VC-SAH1 14, VC-SAH1 15, VC-SAH16-4, VC-SAH17-8, VC-SAH18-7 and VC-SAH 19-1 5 were/are transformed into *A. thaliana* according to the protocol described in Example 5. Selected transgenic lines (T1 generation) are grown in the greenhouse and some flowers are used for crossings (see Example 10).

Example 10: Monitoring activity of the nuclease fusions in *A. thaliana*

Transgenic lines of Arabidopsis harboring a T-DNA encoding a sequence-specific DNA endonuclease are crossed with lines of Arabidopsis harboring the T-DNA carrying a GU-US reporter construct with a corresponding combined target site. As a result of l-Scel activity on the target site a functional GUS gene will be restored by homologous intrachromosomal recombination (ICHR). This can be monitored by histochemical GUS staining (Jefferson et al. (1987) EMBO J 6:3901-3907).

To visualize l-Scel activity of the scTet fusions, transgenic lines of *Arabidopsis* harboring the T-DNA of the nuclease encoding constructs VC-SAH139-20 and VC-SAH 140 are crossed with lines of *Arabidopsis* harboring the T-DNA of constructs VC-SAH 113, VC-SAH 114, VC-SAH 115, harboring the target sites.

To visualize l-Scel activity of the scArc fusions, transgenic lines of *Arabidopsis* harboring the T-DNA of the nuclease encoding constructs VC-SAH89-1 0, VC-SAH90 are crossed with lines of *A. thaliana* harboring the T-DNA of constructs VC-SAH16-4, VC-SAH17-8, VC-SAH18-7, VC-SAH19-1 5, harboring the target sites.

F1 seeds of the crosses are harvested. The seeds are surface sterilized and grown on medium A supplemented with the respective antibiotics and/or herbicides. Leafs are harvested and used for histochemical GUS staining. The percentage of plants showing blue staining is an indicator of the frequency of ICHR and therefore for l-Scel activity.

Activity of the different fusion proteins is determined by comparison of the number ICHR events of these crossings. An increase in specificity of the l-Scel fusions with respect to the native nuclease will be observed by comparing these results with control crosses. For these all transgenic lines of *Arabidopsis* harboring the T-DNA of constructs encoding the different fusions of l-Scel are crossed with lines of *Arabidopsis* harboring the T-DNA of the construct carrying the native l-Scel target site (VC-SAH743-4).

The next generation of these plants is analyzed for fully blue seedlings.
Claims

1. A chimeric endonuclease comprising at least one endonuclease having DNA double strand break inducing activity and at least one heterologous DNA binding domain.

2. A chimeric endonuclease as claimed in claim 1, wherein at least l-Scel, l-Crel, l-Ceul, l-Chul, l-Dmol, Pi-Scel, l-Msol, or l-Anil, or a LAGLIDADG endonuclease having at least 45% amino acid sequence identity to any one of these.

3. A chimeric endonuclease as claimed in claim 1 or 2, wherein the LAGLIDADG endonuclease has at least 80% amino acid sequence identity to a polypeptide described by SEQ ID NO: 1, 2, 3 or 159.

4. A chimeric endonuclease as claimed in any one of claims 1 to 3, comprising a heterologous DNA binding domain derived from a transcription factor or an inactive nuclease, or a fragment comprising a DNA binding domain of a transcription factor or a nuclease.

5. A chimeric endonuclease as claimed in any one of claims 1 to 4, wherein at least one heterologous DNA binding domain is an inactive l-Scel, l-Crel, l-Ceul, l-Chul, l-Dmol, Pi-Scel, l-Msol, or l-Anil or an inactive homolog of these having at least 45% amino acid sequence identity to l-Scel, l-Crel, l-Ceul, l-Chul, l-Dmol, Pi-Scel, l-Msol, or l-Anil.

6. A chimeric endonuclease as claimed in any one of claims 1 to 5, wherein the chimeric endonuclease comprises an engineered endonuclease or an optimized endonuclease or an engineered optimized endonuclease.

7. A chimeric endonuclease as claimed in any one of claims 1 to 6, wherein at least one heterologous DNA binding domain is a transcription factor or a DNA binding domain of a transcription factor comprising a HTH domain.

8. A chimeric endonuclease as claimed in any one of claims 1 to 6, wherein, at least one transcription factor or the DNA binding domain of a transcription factor comprises a HTH domain comprising an amino acid sequence of at least 80% sequence identity to at least one amino acid sequence described by SEQ ID NO: 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118 or 119, preferably described by 91, 92, 93, 94, 95, 112, 113, 114, 115, 116, 117, 118 or 119.

9. A chimeric endonuclease as claimed in any one of claims 1 to 7, wherein the heterologous DNA binding domain comprises a polypeptide having at least 80% amino acid sequence identity to a polypeptide described by SEQ ID NO: 6, or 7.
10. A chimeric endonuclease as claimed in any one of claims 1 to 8, wherein the endonuclease having DNA double strand break inducing activity and the heterologous DNA binding domain are connected via a linker polypeptide.

11. A chimeric endonuclease as claimed in any one of claims 1 to 9, wherein the DNA binding activity of the heterologous DNA binding domain is inducible.

12. A chimeric endonuclease as claimed in any of claims 1 to 19, wherein the DNA binding activity of the heterologous DNA binding domain is inducible by at least one mechanism selected from the group of,
   a. binding of an inducer molecule,
   b. binding of the second monomer of the DNA binding domain
   c. phosphorylation or dephosphorylation,
   d. a rising of temperature or a lowering of temperature.

13. A chimeric endonuclease as claimed in any one of claims 1 to 11, wherein the DNA double strand break inducing activity of the endonuclease is inducible by expression of the second monomer of a homo- or heterodimeric endonuclease.

14. A chimeric endonuclease as claimed in any one of claims 1 to 12, comprising at least one NLS-sequence or one or more SecIII or SecIV secretion signals or a combination of one or more NLS-sequences and one or more SecIII or SecIV secretion signals or a combination one or more SecIII and SecIV secretion signals with one or more NLS-sequences.

15. An isolated polynucleotide comprising a nucleotide sequence, which codes for a chimeric endonuclease as claimed in any one of claims 1 to 13.

16. An isolated polynucleotide comprising a nucleotide sequence, as claimed in claim 14, wherein the sequence of the isolated polynucleotide
   a. is codon optimized,
   b. has a low content RNA instability motives
   c. has a low content of codon repeats,
   d. has a low content of cryptic splice sites,
   e. has a low content of alternative start codons,
   f. has a low content of restriction sites,
   g. has a low content of RNA secondary structures
   h. has any combination of a), b), c), d), e), f) or g).

17. An expression cassette comprising an isolated polynucleotide as claimed in claim 14 or 15 in functional combination with a promoter and an terminator sequence.

18. An isolated polynucleotide comprising a chimeric recognition sequence having a length of about 15 to about 300 nucleotides and comprising
   a. a recognition sequence of an endonuclease and
b. recognition sequence of a heterologous DNA binding domain.

19. An isolated polynucleotide comprising a chimeric recognition sequence as claimed in claim 17, wherein the recognition sequence of the endonuclease is a recognition sequence of a LAGLIDADG endonuclease.

20. An isolated polynucleotide comprising a chimeric recognition sequence as claimed in any one of claims 18 or 19, comprising,
   a. a DNA recognition sequence of l-SceI,
   b. a recognition sequence of scTet or scArc and
   c. a linker sequence of 0 to 10 nucleotides connecting the DNA recognition sequence of l-SceI and the recognition sequence of scTet or scArc.

21. An isolated polynucleotide comprising a chimeric recognition sequence as claimed in any one of claims 18 to 20, comprising a polynucleotide sequence as described by any one of SEQ ID NO: 14, 15, 16, 17, 18, 19 or 20.

22. A vector, host cell or non-human organism comprising
   a. a polynucleotide coding for a chimeric endonuclease as claimed in any of claims 1 to 14, or
   b. an expression cassette as claimed in claim 16, or
   c. an expression cassette as claimed in claim 17, or
   d. an isolated polynucleotide comprising a chimeric recognition sequence as claimed in any one of claims 18 to 21, or
   e. any combination of a), b), c) and d).

23. A non-human organism as claimed in claim 21, wherein the non-human organism is a plant.

24. A method for providing a chimeric endonuclease, comprising the steps of:
   a. providing at least one endonuclease coding region
   b. providing at least one heterologous DNA binding domain coding region,
   c. providing a polynucleotide having a potential DNA recognition sequence or potential DNA recognition sequences of the endonuclease or endonucleases of step a) and having a potential recognition sequence or having potential recognition sequences of the heterologous DNA binding domain or heterologous DNA binding domains of step b),
   d. creating a translational fusion of the coding regions of all endonucleases of step b) and all heterologous DNA binding domains of step c),
   e. expressing a chimeric endonuclease from the translational fusion created in step d),
   f. testing the chimeric endonuclease expressed in step e) for cleavage of the polynucleotide of step c).
25. A method for homologous recombination of polynucleotides comprising:
   a. providing a cell competent for homologous recombination,
   b. providing a polynucleotide comprising an isolated polynucleotide as claimed in any one
      of claims 18 to 21 flanked by a sequence A and a sequence B,
   c. providing a polynucleotide comprising sequences A’ and B’, which are sufficiently
      long and homologous to sequence A and sequence B, to allow for homologous re-
      combination in said cell and
   d. providing a chimeric endonuclease as claimed in any one of claims 1 to 14 or an ex-
      pression cassette as claimed in claim 17,
   e. combining the polynucleotides of b), c) and the chimeric endonuclease of d) in said
      cell and
   f. detecting recombined polynucleotides of b) and c), or selecting for and/or growing
      cells comprising recombined polynucleotides of b) and c).

26. A method for homologous recombination of polynucleotides as claimed in claim 25 wherein
upon homologous recombination a polynucleotide sequence comprised in the competent
   cell of step a) is deleted from the genome of the growing cells of step f)

27. A method for targeted mutation of polynucleotides comprising:
   a. providing a cell comprising a polynucleotide comprising a chimeric recognition site or
      a DNA recognition site,
   b. providing a chimeric endonuclease as claimed in any one of claims 1 to 14 or an ex-
      pression cassette as claimed in claim 17 and being able to cleave the chimeric rec-
      ognition site or the DNA recognition site of step a),
   c. combining the polynucleotide of a) and the chimeric endonuclease of b) in said cell
      and
   d. detecting mutated polynucleotides, or selecting for or growing cells comprising mu-
      tated polynucleotides.

28. A method for homologous recombination or targeted mutation as claimed in any one of
   claims 25 to 27, wherein the chimeric endonuclease and the chimeric recognition site are
   combined in at least one cell via crossing of organisms, via transformation or via transport
   mediated via a Seel ffor SecIV peptide fused to the chimeric endonuclease.
Figure 1: Sequence alignment of different I-Scel homologs

Consensus

(MK IKKNQIMNLGPNSKLLKeyKSQLT LT EQ EAGIGLILGDAYI RSRDEGKTYCMQFEWKNKAYIDHVCLLYDEWVLSPPPKK)

Consensus

(KVNLGNCVLFWQAQtFNFQAFNLANLFVNNKIKIPELVYELPMSLAYWMDGGKWDYNKMSINSKSLVINQSPFBEYVLL)

Consensus

(KVNLGNCVLFWQAQtFNFQAFNLANLFVNNKIKIPELVYELPMSLAYWMDGGKWDYNKMSINSKSLVINQSPFBEYVLL)

Consensus

(RWNLGNVIIIWGAQtFNFQAFNLANLFVNNKIKIPELVYELPMSLAYWMDGGKWDYNKMSINSKSLVINQSPFBEYVLL)

Consensus

(KGLNKFLOLNCVYKIDKKNKPIYIDMSYLVIFYNLIFKFYLYLEQMYKLPENTSSLSEETFLK---)

Consensus

(KGLNKFLOLNCVYKIDKKNKPIYIDMSYLVIFYNLIFKFYLYLEQMYKLPENTSSLSEETFLK---)

Consensus

(GLNKFNLNCMKFNKNKPIYIPS SY IFYNLI PIYPEMKYKLP I S)
Figure 2: Sequence alignment of different I-Crel homologs

<table>
<thead>
<tr>
<th>Consensus</th>
<th>(1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>M L D LIYLAGFDGDGSITIAQIV R DFKKFQ1RLTVQTQLT3K3 FLE I DEIG G VRDR TVSDYVLT</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Consensus</th>
<th>(90)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(90)</td>
<td>K VY LLTQLQPFLRLK KQANLVLKIEQLPSSK S D FLELC LVDQVA LND SK RK TADEV A L L</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Consensus</th>
<th>(179)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(179)</td>
<td>193</td>
</tr>
<tr>
<td>(1)</td>
<td>164</td>
</tr>
<tr>
<td>(164)</td>
<td>NSGI</td>
</tr>
<tr>
<td>(166)</td>
<td>ETDLIEIEEDPSSI</td>
</tr>
<tr>
<td>(166)</td>
<td>168</td>
</tr>
<tr>
<td>(167)</td>
<td>167</td>
</tr>
<tr>
<td>(167)</td>
<td>167</td>
</tr>
</tbody>
</table>
Figure 3a: Sequence alignment of different PI-Scel homologs

(1) MAGAIENARKEIKRISLEHAESEYGAITYSVSGPVIAENMIGCAMYELVKVGHDLNLVGEWIRIDGDKATI0QVYETAGLTVGDPLVR
(2) -------------GAIYSVSVPVIAENMIGCAMYELVKVGHDLNLVGEWIRIDGDKATI0QVYETAGLTVGDPLVR
(3) -------------------------------------------------------------
(4) --------------IYSSLEHAESEYGAITYSVSGPVIAENMIGCAMYELVKVGHDLNLVGEWIRIDGDKATI0QVYETAGLTVGDPLVR
(5) -------------------------------------------------------------

Consensus

(1) GAIYSVSVPVIAENMIGCAMYELVKVGHDLNLVGEWIRIDGDKATI0QVYETAGLTVGDPLVR

(89) 89
1 (89) TGKFLSVELGPGMELIYDGIUQLKAIKESQSIYIPFRGIDTPALDRTIKWQTFTGKFQVGDHISGDYIGSVFENSLISSHLKILP
2 (64) TGKFLSVELGPGMELIYDGIUQLKAIKESQSIYIPFRGIDTPALDRTIKWQTFTGKFQVGDHISGDYIGSVFENSLISSHLKILP
3 (75) TGKFLSVELGPGMELIYDGIUQLKAIKESQSIYIPFRGIDTPALDRTIKWQTFTGKFQVGDHISGDYIGSVFENSLISSHLKILP
4 (1) --------------
5 (1) --------------

Consensus

(89) 177
1 (177) FRSRGITITWIAPAGEYTLDEKILEVEFQEGKSDFPLFHTWFPVYRPREVPTEKLSADYPLLTGQRLDALDFPCVQGCTTCIFPGAFGCGLT
2 (175) FRSRGITITWIAPAGEYTLDEKILEVEFQEGKSDFPLFHTWFPVYRPREVPTEKLSADYPLLTGQRLDALDFPCVQGCTTCIFPGAFGCGLT
3 (163) FRSRGITITWIAPAGEYTLDEKILEVEFQEGKSDFPLFHTWFPVYRPREVPTEKLSADYPLLTGQRLDALDFPCVQGCTTCIFPGAFGCGLT
4 (1) --------------
5 (1) --------------

Consensus

(177) FRSRGITITWIAPAGEYTLDEKILEVEFQEGKSDFPLFHTWFPVYRPREVPTEKLSADYPLLTGQRLDALDFPCVQGCTTCIFPGAFGCGLT

(265) 285
1 (265) VISQSLKYSNDAIYYVGCFAKTNVLADGSLIECICNIEVQNEVQGKGDTHERPWKPRHETMSVSYYQSQVHKSADDSEVPE
2 (243) VISQSLKYSNDAIYYVGCFAKTNVLADGSLIECICNIEVQNEVQGKGDTHERPWKPRHETMSVSYYQSQVHKSADDSEVPE
3 (251) VISQSLKYSNDAIYYVGCFAKTNVLADGSLIECICNIEVQNEVQGKGDTHERPWKPRHETMSVSYYQSQVHKSADDSEVPE
4 (1) --------------
5 (1) --------------

Consensus

(265) V1S4SLKYSNDAIYYVGCFAKTNVLADGSLIECICNIEVQNEVQGKGDTHERPWKPRHETMSVSYYQSQVHKSADDSEVPE
**Figure 3b: Sequence alignment of different PI-Scel homologs**

<table>
<thead>
<tr>
<th></th>
<th>353</th>
<th>360</th>
<th>370</th>
<th>380</th>
<th>390</th>
<th>400</th>
<th>410</th>
<th>420</th>
<th>430</th>
<th>440</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>LFKTTCHHELVVRTPR5VRKLSRTIKGVEPYEVITFEMGQKAPDGKRVELVEKSVSYPISGPANLVEYKAKSNKAF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>LFKTTCHHELVVRTPR5VRKLSRTIKGVEPYEVITFEMGQKAPDGKRVELVEKSVSYPISGPANLVEYKAKSNKAF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>LFKTTCHHELVVRTPR5VRKLSRTIKGVEPYEVITFEMGQKAPDGKRVELVEKSVSYPISGPANLVEYKAKSNKAF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>LFKTTCHHELVVRTPR5VRKLSRTIKGVEPYEVITFEMGQKAPDGKRVELVEKSVSYPISGPANLVEYKAKSNKAF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>ALFMTCHHELVVRTPR5VRKLSRTIKGVEPYEVITFEMGQKAPDGKRVELVEKSVSYPISGPANLVEYKAKSNKAF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Consensus:** LFKTTCHHELVVRTPR5VRKLSRTIKGVEPYEVITFEMGQKAPDGKRVELVEKSVSYPISGPANLVEYKAKSNKAF

<table>
<thead>
<tr>
<th></th>
<th>441</th>
<th>450</th>
<th>460</th>
<th>470</th>
<th>480</th>
<th>490</th>
<th>500</th>
<th>510</th>
<th>528</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ETITEARLSLGLGVVATKKYTSAPATHY1ENDHFFDYMQLKSKHLYTECPRVLATYLGWLWGGGSLDSFATSSVDRSRTLMEYTA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>ETITEARLSLGLGVVATKKYTSAPATHY1ENDHFFDYMQLKSKHLYTECPRVLATYLGWLWGGGSLDSFATSSVDRSRTLMEYTA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>ETITEARLSLGLGVVATKKYTSAPATHY1ENDHFFDYMQLKSKHLYTECPRVLATYLGWLWGGGSLDSFATSSVDRSRTLMEYTA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>ETITEARLSLGLGVVATKKYTSAPATHY1ENDHFFDYMQLKSKHLYTECPRVLATYLGWLWGGGSLDSFATSSVDRSRTLMEYTA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>ETITEARLSLGLGVVATKKYTSAPATHY1ENDHFFDYMQLKSKHLYTECPRVLATYLGWLWGGGSLDSFATSSVDRSRTLMEYTA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Consensus:** EITTEARDLSLGLGVVATKKYTSAPATHY1ENDHFFDYMQLKSKHLYTECPRVLATYLGWLWGGGSLDSFATSSVDRSRTLMEYTA

<table>
<thead>
<tr>
<th></th>
<th>529</th>
<th>540</th>
<th>550</th>
<th>560</th>
<th>570</th>
<th>580</th>
<th>590</th>
<th>600</th>
<th>616</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>EKLNLCAEKYDKREPQQAKTVNLGGSGYMLKRRPDGSKNIPFPNSPFDALTIFSGLDSGYVTD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>EKLNLCAEKYDKREPQQAKTVNLGGSGYMLKRRPDGSKNIPFPNSPFDALTIFSGLDSGYVTD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>EKLNLCAEKYDKREPQQAKTVNLGGSGYMLKRRPDGSKNIPFPNSPFDALTIFSGLDSGYVTD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>EKLNLCAEKYDKREPQQAKTVNLGGSGYMLKRRPDGSKNIPFPNSPFDALTIFSGLDSGYVTD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>EKLNLCAEKYDKREPQQAKTVNLGGSGYMLKRRPDGSKNIPFPNSPFDALTIFSGLDSGYVTD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Consensus:** EKLNLCAEKYDKREPQQAKTVNLGGSGYMLKRRPDGSKNIPFPNSPFDALTIFSGLDSGYVTD

<table>
<thead>
<tr>
<th></th>
<th>617</th>
<th>630</th>
<th>640</th>
<th>650</th>
<th>660</th>
<th>670</th>
<th>680</th>
<th>690</th>
<th>704</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>EGHGAKTIKTHTSVGRDLSLGRVSLGSVNGAPARDINNGISYAYMSGGDVLNLNSLGCAGSKKFRAPFAPAARAARECRGF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>EGHGAKTIKTHTSVGRDLSLGRVSLGSVNGAPARDINNGISYAYMSGGDVLNLNSLGCAGSKKFRAPFAPAARAARECRGF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>EGHGAKTIKTHTSVGRDLSLGRVSLGSVNGAPARDINNGISYAYMSGGDVLNLNSLGCAGSKKFRAPFAPAARAARECRGF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>EGHGAKTIKTHTSVGRDLSLGRVSLGSVNGAPARDINNGISYAYMSGGDVLNLNSLGCAGSKKFRAPFAPAARAARECRGF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>EGHGAKTIKTHTSVGRDLSLGRVSLGSVNGAPARDINNGISYAYMSGGDVLNLNSLGCAGSKKFRAPFAPAARAARECRGF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Consensus:** EGHGAKTIKTHTSVGRDLSLGRVSLGSVNGAPARDINNGISYAYMSGGDVLNLNSLGCAGSKKFRAPFAPAARAARECRGF
Figure 3c: Sequence alignment of different PI-Scel homologs

<table>
<thead>
<tr>
<th></th>
<th>(705)</th>
<th>710</th>
<th>720</th>
<th>730</th>
<th>740</th>
<th>750</th>
<th>760</th>
<th>770</th>
<th>780</th>
<th>792</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>YELELKEDDLDDYGVITLSDDDHQLPLLANQQVHNCGERGNEAEVMHMEFPETYEMTSKEPMKRTTLVANTSMNPVAAREASIYT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>YELELKEDDLDDYGVITLSDDDHQLPLLANQQVHNCGERGNEAEVMHMEFPETYEMTSKEPMKRTTLVANTSMNPVAAREASIYT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>YELELKEDDLDDYGVITLSDDDHQLPLLANQQVHNCGERGNEAEVMHMEFPETYEMTSKEPMKRTTLVANTSMNPVAAREASIYT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>YELELKEDDLDDYGVITLSDDDHQLPLLANQQVHNCGERGNEAEVMHMEFPETYEMTSKEPMKRTTLVANTSMNPVAAREASIYT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>YELELKEDDLDDYGVITLSDDDHQLPLLANQQVHNCGERGNEAEVMHMEFPETYEMTSKEPMKRTTLVANTSMNPVAAREASIYT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consensus</td>
<td>YELELKEDDLDDYGVITLSDDDHQLPLLANQQVHNCGERGNEAEVMHMEFPETYEMTSKEPMKRTTLVANTSMNPVAAREASIYT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>(793)</th>
<th>800</th>
<th>810</th>
<th>820</th>
<th>830</th>
<th>840</th>
<th>850</th>
<th>860</th>
<th>870</th>
<th>880</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GTLAELFRDQGKVMNIMADSSRKWEALRISGRLGEMPAQDFGPAYLGAFLSAEYERAGKAVALGSPDRTGVSIVAAVSPAGDF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>GTLAELFRDQGKVMNIMADSSRKWEALRISGRLGEMPAQDFGPAYLGAFLSAEYERAGKAVALGSPDRTGVSIVAAVSPAGDF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>GTLAELFRDQGKVMNIMADSSRKWEALRISGRLGEMPAQDFGPAYLGAFLSAEYERAGKAVALGSPDRTGVSIVAAVSPAGDF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>GTLAELFRDQGKVMNIMADSSRKWEALRISGRLGEMPAQDFGPAYLGAFLSAEYERAGKAVALGSPDRTGVSIVAAVSPAGDF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>GTLAELFRDQGKVMNIMADSSRKWEALRISGRLGEMPAQDFGPAYLGAFLSAEYERAGKAVALGSPDRTGVSIVAAVSPAGDF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consensus</td>
<td>GTLAELFRDQGKVMNIMADSSRKWEALRISGRLGEMPAQDFGPAYLGAFLSAEYERAGKAVALGSPDRTGVSIVAAVSPAGDF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>(881)</th>
<th>890</th>
<th>890</th>
<th>890</th>
<th>890</th>
<th>900</th>
<th>910</th>
<th>920</th>
<th>930</th>
<th>940</th>
<th>950</th>
<th>968</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SEDYSTATLGTQFGLKLDKLQARQKEFESSNHYSYQYVQLVXAGSKLSADSD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>SEDYSTATLGTQFGLKLDKLQARQKEFESSNHYSYQYVQLVXAGSKLSADSD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>SEDYSTATLGTQFGLKLDKLQARQKEFESSNHYSYQYVQLVXAGSKLSADSD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>SEDYSTATLGTQFGLKLDKLQARQKEFESSNHYSYQYVQLVXAGSKLSADSD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>SEDYSTATLGTQFGLKLDKLQARQKEFESSNHYSYQYVQLVXAGSKLSADSD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consensus</td>
<td>SEDYSTATLGTQFGLKLDKLQARQKEFESSNHYSYQYVQLVXAGSKLSADSD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>(969)</th>
<th>980</th>
<th>990</th>
<th>1000</th>
<th>1010</th>
<th>1020</th>
<th>1030</th>
<th>1040</th>
<th>1050</th>
<th>1056</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>KITLDEVATLRTEDFLQQGNSYSTYDAFCPIMKTFDMRAFIYSYKDEAQRACAVANGANWKLADSTGDVHKSSEFEPSRGEKHYEG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>KITLDEVATLRTEDFLQQGNSYSTYDAFCPIMKTFDMRAFIYSYKDEAQRACAVANGANWKLADSTGDVHKSSEFEPSRGEKHYEG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>KITLDEVATLRTEDFLQQGNSYSTYDAFCPIMKTFDMRAFIYSYKDEAQRACAVANGANWKLADSTGDVHKSSEFEPSRGEKHYEG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>KITLDEVATLRTEDFLQQGNSYSTYDAFCPIMKTFDMRAFIYSYKDEAQRACAVANGANWKLADSTGDVHKSSEFEPSRGEKHYEG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>KITLDEVATLRTEDFLQQGNSYSTYDAFCPIMKTFDMRAFIYSYKDEAQRACAVANGANWKLADSTGDVHKSSEFEPSRGEKHYEG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consensus</td>
<td>KITLDEVATLRTEDFLQQGNSYSTYDAFCPIMKTFDMRAFIYSYKDEAQRACAVANGANWKLADSTGDVHKSSEFEPSRGEKHYEG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>(1057)</th>
<th>1057</th>
<th>1073</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>FKLHSTMQRFAESTD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>-------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>FKLHLS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>FKLHLS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>FKLHLS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consensus</td>
<td>FKLHSTMQRFAESTD</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 4: Sequence alignment of different I-CeuI homologs
Figure 5: Sequence alignment of different I-Chul homologs

Consensus
Figure 6: Sequence alignment of different I-Dmol homologs

Consensus: MSVAyllGLVGDGGYLYALRYKGRTEYRVVTQKDEAVVEKAVY MLEALLRELGKSKVQVRGRSRETEVRVSSKALWQ
Figure 7: Sequence alignment of different I-Msol homologs

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>70</th>
<th>89</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MTKNLTQPEAAAYIAGFLDGGSITYAKLIAPPDSDKXYQXQVSEALTIFS</td>
<td>QRXDFPPYLDYDIYDQGLRGNLHRONGSDKRGGDIADYIGSITTH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>MKLESTLAAAYIAGFLDGGSITYAKLISRPDXAVIKXYQVSEFSQXQYLDYDENALDKGSLRKDRGGDIADYITTPAGA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consensus</td>
<td>TL T AAYIAGFLDGGSITYAKLI RPDY IKYQISLAIISP QRKDRF YLDYDILK G LRKDRGGDIADYITH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>90</th>
<th>100</th>
<th>110</th>
<th>120</th>
<th>130</th>
<th>140</th>
<th>150</th>
<th>160</th>
<th>170</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>LSIILPDLLPYLRITKQKANRLHIINLYPQAQKNFSKFLDLVKEIKYDVMQNLNIRADIKSTNYDRRKLEPLKAKHEISSSP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>LSIILPHLLPYLRITKQKANQLHIINQYPAACKNHEFLSLVKEKVDQNLNIRADPLPKRNYQSLLEEFQTACHQISSSP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consensus</td>
<td>LSIILP LLPYLRITKQKAN ILHIIN YP A KN FL LVKIVD IQNLNKR DE KATNY LLEEF AGKI SSP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 8: Sequence alignment of different TetR homologs

Consensus
Figure 9a: Sequence alignment of HTH domains of different TetR homologs

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>61</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>KLQ PNT VIRA LDLL NEVG VDG LTR KLA ERL GVQ PAP YWH FRN KR AL LD L A E M L A E N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>KLQ PNT VIR A LD IN EVG VEA LTTR KLA KRL GVQ RPY WHF RN KR AL LD L A E M L A E -</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>KLQ KGT VJA AL DELL NEVG MSL TTTR KLA ERL KVQ QPAP YWHF QNKR AL LD L A E M L A E -</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>KLQRE AVIR TALE LLEL DVGMGE LTTR KLA ERL GVQ PAP YWHF RNKR AL LD L A E M L TIN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>KLQ RAVI QTLALE LLEL NEVG ESG LTR KLA ERL GVQ PAP YWHF RNKR VLLLD LAE TIAE -</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Consensus: (1) KLQ RVIR ALEL NEVG MEGL TR LK LQ ERL GVQ PAP YWHF RNKR AL LD L A E M L A E N

Figure 9b: Sequence alignment of HTH domains of different ArcR homologs

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>74</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>KFA KLF I KILC PCQVGYQ DEFY IE IKQLNM S DMA GS RTAGI TIHELF DKDV KHD KNWL VS K L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>KLaketk KLILC PCQVGYD DEFYE I KHFM TIQL LS DMA GS R ET TSH I INEL KEEK ILFR KN S KNWL VS K DL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>KLaketk KLILC PCQVGYD DEFYE I KHFM TIQL LS DMA GS R ET TSH I INEL KEEK ILFR KN S KNWL VS K DL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>KLaketk KLILC PCQVGYD NEFY IE IKQMI TQL LS DMA GS R ET TGI HI I NEL KEEK ILFR KS N KNWL LS KEL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>KLaketk KLILC PCQVGYD NEFY IE IKQMI TQL LS DMA GS R ET TGI HI I NEL KEEK ILFR KS N KNWL LS KEL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Consensus: (1) KLaketk KLILC PCQVGYD DEFYE I KHFM TIQL LS DMA GS R ET TGI HI I NEL KEEK ILFR KS N KNWL VS K DL
Figure 10a: Sequence alignment of HTH domains of different LacR homologs

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>58</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MKFVTLYDVAEQYAGVSYTQTVSFVVQASHVSAKRKEVEAMAELNYIPNRMAQQLAG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>MKFVTLYDVAEQYAGVSYTQTVSFVVQASHVSAKRKEVEAMAELNYIPNRMAQQLAG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>MKFVTLYDVAEQYAGVSYTQTVSFVVQASHVSAKRKEVEAMAELNYIPNRMAQQLAG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>KPIPLLHDDAVYAGVSYQTQSVLNNQRHAIRCSIQRHKEQAMAALNYIPNRMAQQLAG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>CRTTTTLRVDARGVSYQTQSVRLNNERKETVLRAMALHYIPNRGAQQLAG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consensus</td>
<td>MKPVTLYDVAEYAGVSYQTQTVSFVVQASHVSAKRKEVEAMAELNYIPNRMAQQLAG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 10b: Sequence alignment of HTH domains of different MerR homologs

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>70</th>
<th>80</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MQFREGLEADCGVKNTEIVYRGLLPEPRTEKGYRMSQQTVDLRNFVTRMKQLGFTLNEIDYHLGVVDRDEAKCR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>MQFREGLEADCGVKNTEIVYRGLLPEPRTEKGYRMSQQTVDLRNFVTRMKQLGFTLNEIDYHLGVVDRDEAKCR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>MQFREGLEADCGVKNTEIVYRGLLPEPRTEKGYRMSQQTVDLRNFVTRMKQLGFTLNEIDYHLGVVDRDEAKCR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>MQFREGLEADCGVKNTEIVYRGLLPEPRTEKGYRMSQQTVDLRNFVTRMKQLGFTLNEIDYHLGVVDRDEAKCR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>MQFREGLEADCGVKNTEIVYRGLLPEPRTEKGYRMSQQTVDLRNFVTRMKQLGFTLNEIDYHLGVVDRDEAKCR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>MQFREGLEADCGVKNTEIVYRGLLPEPRTEKGYRMSQQTVDLRNFVTRMKQLGFTLNEIDYHLGVVDRDEAKCR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consensus</td>
<td>MQFREGLEADCGVKNTEIVYRGLLPEPRTEKGYRMSQQTVDLRNFVTRMKQLGFTLNEIDYHLGVVDRDEAKCR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 11: Sequence alignment of HTH domains of different MarA homologs
Figure 12: Sequence alignment of different MarA homologs

Consensus (90)  

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>1</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
</tr>
<tr>
<td>2</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
</tr>
<tr>
<td>3</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
</tr>
<tr>
<td>4</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
</tr>
<tr>
<td>5</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
</tr>
<tr>
<td>6</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
</tr>
<tr>
<td>7</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
</tr>
<tr>
<td>8</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
<td>SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMNHQETSPFPLHNLNN</td>
</tr>
</tbody>
</table>

Consensus (90) SNEPILYLAERYGFESQQTTRFPKNVDFPPHYRITMGESRYLPLNHAA
**INTERNATIONAL SEARCH REPORT**

**A. CLASSIFICATION OF SUBJECT MATTER**

See extra sheet

According to International Patent Classification (IPC) or to both national classification and IPC

**B. FIELDS SEARCHED**

Minimum documentation searched (classification system followed by classification symbols)

IPC: C12N 9/17; C12N 15/17

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

WPI; EPODOC; CNKI; CNPAT; NCBI PubMed; GOOGLE Scholar; ISI Web of Knowledge: nuclease, endonuclease, meganuclease, homing, transcription factor, binding, recognition, cleav+, domain?, sequence?, region?, site?, polypeptide?, peptide?, nucleic acid?, chimeric, chimera?, fus+, hybrid?; See HTH, tet, arc, homologous recombination, GUS

**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

<table>
<thead>
<tr>
<th>Category</th>
<th>Citation of document, with indication, where appropriate, of the relevant passages</th>
<th>Relevant to claim No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>WO 2004/03 1346 A2 (FRED HUTCHINSON CANCER RESEARCH CENTER) 15 Apr. 2004 (15.04.2004) page 3, lines 13-15; page 4, lines 3-12; page 5, lines 9-30; page 11, line 18 to page 13, line 25; page 14, lines 3-24; page 17, line 20 to page 18, line 9; page 34, line 1 to page 37, line 27</td>
<td>1-23</td>
</tr>
<tr>
<td>A</td>
<td>US 7008780 B2 (MASSACHUSETTS INSTITUTE OF TECHNOLOGY) 07 Mar. 2006 (07.03.2006) the whole document</td>
<td>1-28</td>
</tr>
</tbody>
</table>

* Further documents are listed in the continuation of Box C.  
See patent family annex.

**Date of the actual completion of the international search**  
02 Mar. 2011 (02.03.2011)

**Date of mailing of the international search report**  
17 Mar. 2011 (17.03.2011)

**Name and mailing address of the ISA/CN**  
The State Intellectual Property Office, the P.R.China
6 Xitucheng Rd., Jimen Bridge, Haidian District, Beijing, China 100088
Facsimile No. 86-10-62019451

**Authorized officer**  
TANG Li

**Telephone No.** (86-10) 82245036

Form PCT/ISA/210 (second sheet) (July 2009)
## DOCUMENTS CONSIDERED TO BE RELEVANT

<table>
<thead>
<tr>
<th>Category</th>
<th>Citation of document, with indication, where appropriate, of the relevant passages</th>
<th>Relevant to claim No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>LANGDON, Robert C et al. A chimeric activator of transcription that uses two DNA-binding domains to make simultaneous contact with pairs of recognition sites. Molecular Microbiology. 2001, Vol.41, No. 4, pages 885-896</td>
<td>1-28</td>
</tr>
<tr>
<td>Patent Documents referred in the Report</td>
<td>Publication Date</td>
<td>Patent Family</td>
</tr>
<tr>
<td>-----------------------------------------</td>
<td>-----------------</td>
<td>---------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AU 2003290518 A8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CA 22091 83 A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AU 4895096 A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AU 719001 B</td>
</tr>
<tr>
<td></td>
<td></td>
<td>KR 987000992 A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>US 2003 126624 A1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>JP 3817739 B2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>US 20071 50973 A1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>US 7485441 B2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>US 7763446 B2</td>
</tr>
</tbody>
</table>
**INTERNATIONAL SEARCH REPORT**

| Box No. I | Nucleotide and/or amino acid sequence(s) (Continuation of item iteml.c of the first sheet) |

1. With regard to any nucleotide and/or amino acid sequence disclosed in the international application, the international search was carried out on the basis of:

   a. a sequence listing filed or furnished
      - □ on paper
      - ☒ in electronic form

   b. time of filing or furnishing
      - ☒ contained in the applicant as filed
      - □ filed together with the application in electronic form
      - □ furnished subsequently to this Authority for the purposes of search

2. ☐ In addition, in the case that more than one version or copy of a sequence listing has been filed or furnished, the required statements that the information in the subsequent or additional copies is identical to that in the application as filed or does not go beyond the application as filed, as appropriate, were furnished.

3. Additional comments:

---

Form PCT/ISA/210 (continuation of first sheet (1)) (July 2009)
Continuation of: CLASSIFICATION OF SUBJECT MATTER

C12N 9/22 (2006.01) i
C12N 15/10 (2006.01) i
C12N 15/52 (2006.01) i
C12N 15/62 (2006.01) i
C12N 15/82 (2006.01) i