Title: CHIMERIC ENDONUCLEASES AND USES THEREOF

Abstract: The invention relates to chimeric endonucleases, comprising a endonuclease and a heterologous DNA binding domain comprising one or more Zn$_2$C$_6$ zinc fingers, 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 an endonuclease and a heterologous DNA binding domain comprising one or more $z_{12}C_6$ zinc fingers, as well as methods of targeted integration, targeted deletion or targeted mutation of polynucleotides using such 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 "meganuclease" or "homing endonucleases" and are frequently associated with parasitic DNA elements, such as group 1 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 (US2011/0071083). Most meganucleases do also act on sequences similar to their optimal binding site, which may lead to unintended or even detrimental off-target effects. Despite the fact, that several approaches have already been taken to avoid enhance the efficiency of meganuclease induced homologous recombination e.g. by fusing nucleases 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), there is still a need in the art to develop meganucleases having high induction rates of homologous recombination and/or a high specificity in regard to 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 LAGLIDADG endonuclease and at least one a heterologous DNA binding domain comprising one or more Zn\(\eta\)\(\alpha\)\(\kappa\) zinc fingers. Preferably the chimeric endonuclease comprises at least one LAGLIDADG endonuclease comprising an amino acid sequence having at least 80% amino acid sequence identity to a polypeptide described by any one of SEQ ID NOs: 1, 2, 3, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 161 or 165. Preferably having at least 80% amino acid sequence identity to a polypeptide described by SEQ ID NO: 1, 2 or 3. In one preferred embodiment, the chimeric endonuclease comprises at least one LAGLIDADG endonuclease, which is an engineered or optimized endonuclease or an optimized version of an engineered endonuclease, preferably an optimized endonuclease or an optimized version of an engineered endonuclease. In a further embodiment, the chimeric endonuclease comprises a heterologous DNA binding domain comprising one or more Zn\(\eta\)\(\alpha\)\(\kappa\) zinc fingers derived from a transcription factor. In an preferred embodiment, the chimeric endonuclease comprises a heterologous DNA binding domain comprising at least one polypeptide having at least 80% amino acid sequence identity to a polypeptide described by any one of SEQ ID NOs: 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, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120 or 121. The chimeric endonucleases described herein may or may not comprise a linker to connect at least one endonuclease with at least one heterologous DNA binding domain. Preferably the linker (synonymous "linker polypeptide") consists of at least 3 amino acids and wherein the at least one third of the amino acids in the amino acid sequence of this linker polypeptide are glycine or serine or alanine or a combination of glycine, serine and alanine. Preferably the chimeric endonucleases comprise at least one NLS-sequence and or a Seel 11 or Sec1V secretion signal. One embodiment of the invention provides chimeric endonucleases, wherein the DNA binding activity of the heterologous DNA binding domain is inducible, preferably being inducible by expression of a second monomer of a dimeric or heterodimeric LAGLIDADG endonuclease. The invention provides further for isolated polynucleotides compris-
ing a nucleotide sequence, which codes for a chimeric endonuclease of the invention. Preferably the isolated polynucleotide is codon optimized or has a low content RNA instability motives 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, has a low content of restriction sites, or has a low content of RNA secondary structures or has any combination of those features. Another embodiment of the invention is an expression cassette comprising an isolated polynucleotide as described above in functional combination with a promoter and a terminator sequence.

Further embodiment of 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 a LAGLIDADG endonuclease and a recognition sequence of a heterologous DNA binding domain comprising one or more zinc fingers. Preferably the recognition sequence of the heterologous DNA binding domain can be bound by at least one DNA binding domain comprising an amino acid sequence described by any one of SEQ ID NOs: 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, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121.

The invention does also provide for isolated polynucleotides comprising a chimeric recognition sequence, preferably the chimeric recognition sequence is comprised in an expression cassette, or close to the 5’- or 3’-end or close to both ends, wherein the expression cassette comprises a promoter, a terminator and a sequence capable to be expressed by the promoter. Preferably the sequence capable to be expressed codes for a marker gene.

The invention does further provide a a chimeric recognition sequence comprising a DNA recognition sequence of l-Scel and a recognition sequence, which can be bound by at least one DNA binding domain comprising an amino acid sequence described by any one of SEQ ID NOs: 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, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, being directly connected or being connected by a sequence of 1 to 10 nucleotides. In one embodiment the chimeric recognition sequence comprises a DNA recognition sequence of l-Scel and a recognition sequence of AlcR, or AlcR (1-60) being directly connected or being connected by a sequence of 1 to 10 nucleotides. In one embodiment of the invention, the isolated polynucleotide comprises a chimeric recognition sequence, comprising a polynucleotide sequence as described by any one of SEQ ID NO: 13, 14, 15, 16, 43, 44, 45 or 46.

Other embodiments of the invention are vectors, host cells or non human organisms comprising a polynucleotide coding for a chimeric endonuclease, or an isolated polynucleotide coding for a chimeric endonuclease or an isolated polynucleotide comprising an chimeric recognition sequence, or an expression cassette comprising a polynucleotide coding for a chimeric endonuclease or an chimeric recognition sequence, and vectors, host cells or non human organisms comprising a combination of the chimeric endonucleases, isolated polynucleotides and expression cassettes described above. 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 in 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: a) providing at least one heterologous DNA binding domain 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 integration 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 model of a chimeric nuclease comprising l-Scel as N-terminal and amino acids 1 to 60 of AICR as a C-terminal domain.

Figure 2 is a graphical representation of the experimental results of Examples 10 and 20d as described herein. Thereby providing a comparison of the capability of wildtype l-Scel and three
different variants of chimeric endonucleases to induce homologous recombination in plants. The frequency of induced homologous recombination is represented by the percentage of plants showing GUS-activity after homologous recombination (% blue plants). Figure 2 does also comprise the polynucleotide sequence of the DNA recognition sequence of wildtype l-Scel, called wt target site, as well as the polynucleotide sequence of the chimeric recognition site of the chimeric endonucleases (l-Scel-AlcR(1-60), l-Scel#2-AlcR(1-60) and l-Scel#1-AlcR(1-60)). Figure 2 provides further the amino acid sequence showing the C-terminus of Seel, comprising different mutations, the lysine (L) used as linker and the first six amino acids of the former N-terminus of AlcR (1-60) used to create the chimeric endonucleases. The different mutations on the former C-terminus of wt l-Scel change the wildtype amino acid sequence "TISSETFLK" to "TIKSEETFLK" in the chimeric endonuclease l-Scel#1-AlcR(1-60) and to "ALIANQAFKL" in the chimeric endonuclease l-Scel#2-AlcR(1-60).

Figure 3 depicts a sequence alignment of different l-Scel homologs, wherein 1 is SEQ ID NO: 1, 2 is SEQ ID NO: 122, 3 is SEQ ID NO: 123, 4 is SEQ ID NO: 124, 5 is SEQ ID NO: 125.

Figure 4 depicts a sequence alignment of different l-Crel homologs, wherein 1 is SEQ ID NO: 126, 2 is SEQ ID NO: 127, 3 is SEQ ID NO: 128, 4 is SEQ ID NO: 129, 5 is SEQ ID NO: 130.

Figures 5a to 5c depicts a sequence alignment of different Pl-Scel homologs, wherein 1 is SEQ ID NO: 145, 2 is SEQ ID NO: 146, 3 is SEQ ID NO: 147, 4 is SEQ ID NO: 148, 5 is SEQ ID NO: 149.

Figure 6 depicts a sequence alignment of different l-Ceul homologs, wherein 1 is SEQ ID NO: 131, 2 is SEQ ID NO: 132, 3 is SEQ ID NO: 133, 4 is SEQ ID NO: 134, 5 is SEQ ID NO: 135.

Figure 7 depicts a sequence alignment of different l-Chul homologs, wherein 1 is SEQ ID NO: 136, 2 is SEQ ID NO: 137, 3 is SEQ ID NO: 138, 4 is SEQ ID NO: 139, 5 is SEQ ID NO: 140.

Figure 8 depicts a sequence alignment of different l-Dmol homologs, wherein 1 is SEQ ID NO: 141, 2 is SEQ ID NO: 142, 3 is SEQ ID NO: 143, 4 is SEQ ID NO: 144.

Figure 9 depicts a sequence alignment of different l-Msol homologs, wherein 1 is SEQ ID NO: 150 and 2 is SEQ ID NO: 151.

Figure 10 shows a sequence alignment of different \( \text{\textit{zeta}}_2\text{C}_6 \) domains homologous to the DNA binding domain of AlcR (AlcR 1 to 60). The Consensus sequence shows amino acids being conserved in those homologs (AlcR 1 to 60 consensus sequence). Sequence No. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24 and 25 refer to SEQ ID NO: 70, 71, 72, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93 and 94, respectively.
Figure 11 shows a sequence alignment of different ZnCC6 domains homologous to the DNA binding domain of AflR. The Consensus sequence shows amino acids being conserved in those homologs (AflR consensus sequence). Sequence No. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 and 12 refer to SEQ ID NO: 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68 and 69, respectively.

Figure 12 shows a sequence alignment of different ZnCC6 domains homologous to the DNA binding domain of Hap1. The Consensus sequence shows amino acids being conserved in those homologs (Hap1 consensus sequence). Sequence No. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 and 13 refer to SEQ ID NO: 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106 and 107, respectively.

Figure 13 shows a sequence alignment of different ZnCC6 domains homologous to the DNA binding domain of Leu3. The Consensus sequence shows amino acids being conserved in those homologs (Leu3 consensus sequence). Sequence No. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13 and 14 refer to SEQ ID NO: 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120 and 121, respectively.

DETAILED DESCRIPTION OF THE INVENTION

The invention provides chimeric LAGLIDADG endonucleases, comprising at least one LAGLIDADG endonuclease and at least one heterologous DNA binding domain comprising one or more ZnCC6 zinc fingers.

LAGLIDADG endonucleases

LAGLIDADG endonucleases useful in the invention 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 one or several amino acid changes in their LAGLIDADG motif. LAGLIDADG endonucleases comprising only one LAGLIDADG motif act usually 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 from polynucleotides of organisms mentioned for exemplary purposes in Table 1 to 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 InterPro database, e.g. InterPro accession number IPR004860.

The term LAGLIDADG endonucleases shall also encompass artificial homo- and heterodimeric LAGLIDADG endonucleases, which can be created 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 l-Dmo I as one domain can be found for example 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.

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/I, l-Sce I, l-Chu I, l-Dmo I, l-Cre I, l-Csm I, Pl-Sce I, Pl-Tli I, Pl-Mtu I, Pl-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, 1-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, 1-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 prefered 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, 1-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 prefered are l-Dmo I, l-Cre I, l-Sce I, 1-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 I-Sce I and homologs of I-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.

Preferred monomeric LAGLIDADG endonucleases are: l-Cre, l-Sce I, l-Chu, l-Dmo I, l-Csm I, Pl-Sce I, Pl-Tli I, Pl-Mtu I, l-Sce II, l-Sce III, HO, Pi-Civ I, Pi-Ctr I, Pl-Aae I, Pl-Bsu I, Pl-Dha I, Pl-Dra I, Pl-Mav I, Pl-Mch I, Pi-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, Pi-Msm I, Pi-Mth I, Pl-Mtu I, Pl-Mxe I, Pi-Npu I, Pi-Pfu I, Pi-Rma I, Pi-Spb I, Pl-Ssp I, Pl-Fac I, Pl-Mja I, Pi-Pho I, Pi-Tag I, Pi-Thy I, Pi-Tko I, and Pl-Tsp 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.

Preferably, I-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 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 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; 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.

Still more preferred monomeric LAGLIDADG endonucleases are: l-Dmo I, l-Sce I, and l-Chu 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 LAGLIDADG endonucleases are the LAGLIDADG endonucleases mentioned in Tables 1 to 6 and homologs of these having at least 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.

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 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 have no effect on its DNA-binding-affinity, its dimer formation affinity or which 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, KD (e.g., the KD of l-Cre I for the WT DNA recognition sequence is approximately 0.1 nM). As used herein, a meganuclease has "altered" binding affinity if the KD 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, KD. As used herein, a meganuclease has "altered" affinity for dimer formation, if the KD of the recombinant meganuclease monomer or the recombinant LAGLIDADG endonuclease monomer for a reference meganuclease monomer or for a refer-
ence 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 mutating 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 for example, but not excluding others;: A36G, L40M, L40V, 141 S, 141 N, 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 nucleases.

In one embodiment, the LAGLIDADG endonuclease is selected from the group comprising: l-Sce 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-T7 I, 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>UniProt 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>122</td>
<td>98</td>
</tr>
<tr>
<td>063264</td>
<td>Z. bisporus</td>
<td>123</td>
<td>72</td>
</tr>
<tr>
<td>Q34839</td>
<td>K. thermotolerans</td>
<td>124</td>
<td>71</td>
</tr>
</tbody>
</table>
Tabel 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>126</td>
<td>100</td>
</tr>
<tr>
<td>Q8SMM1</td>
<td>C. lunzensis</td>
<td>127</td>
<td>56</td>
</tr>
<tr>
<td>Q8SML7</td>
<td>C. olivieri</td>
<td>128</td>
<td>58</td>
</tr>
<tr>
<td>Q1KVQ8</td>
<td>S. obliquus</td>
<td>129</td>
<td>49</td>
</tr>
</tbody>
</table>

Tabel 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>145</td>
<td>100</td>
</tr>
<tr>
<td>Q874G9</td>
<td>S. cerevisiae</td>
<td>146</td>
<td>99</td>
</tr>
<tr>
<td>Q874F9</td>
<td>S. pastorianus</td>
<td>147</td>
<td>97</td>
</tr>
<tr>
<td>Q8J0H1</td>
<td>S. cariocanus</td>
<td>148</td>
<td>87</td>
</tr>
<tr>
<td>Q8J0G4</td>
<td>Z. bailii</td>
<td>149</td>
<td>61</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>131</td>
<td>100%</td>
</tr>
<tr>
<td>Q8WKZ1</td>
<td>C. echinozygotum</td>
<td>132</td>
<td>63%</td>
</tr>
<tr>
<td>Q8WL12</td>
<td>C. elongatum</td>
<td>133</td>
<td>58%</td>
</tr>
<tr>
<td>Q8WL11</td>
<td>A. stipitatus</td>
<td>134</td>
<td>55%</td>
</tr>
<tr>
<td>Q8WKX7</td>
<td>C. monadina</td>
<td>135</td>
<td>51%</td>
</tr>
</tbody>
</table>

Table 5: Exemplary homologs of l-Chul, which can be cloned from other organisms are described in Table 1.

<table>
<thead>
<tr>
<th>Uni-Prot Accession Nr.</th>
<th>Organism</th>
<th>SEQ ID NO:</th>
<th>Amino Acid Sequence Identity to l-Chul</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q53X18</td>
<td>C. humicola</td>
<td>136</td>
<td>100%</td>
</tr>
<tr>
<td>Q8WL03</td>
<td>C. zebra</td>
<td>137</td>
<td>67%</td>
</tr>
<tr>
<td>Q8WKX6</td>
<td>C. monadina</td>
<td>138</td>
<td>62%</td>
</tr>
<tr>
<td>Q8WL10</td>
<td>A. stipitatus</td>
<td>139</td>
<td>58%</td>
</tr>
<tr>
<td>Q8SMI6</td>
<td>N. aquatica</td>
<td>140</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><em>D. mobilis</em></td>
<td>141</td>
<td>100%</td>
</tr>
<tr>
<td>Q6L6Z4</td>
<td><em>Thermoproteus sp.</em></td>
<td>142</td>
<td>51%</td>
</tr>
<tr>
<td>Q6L6Z5</td>
<td><em>Thermoproteus sp.</em></td>
<td>143</td>
<td>50%</td>
</tr>
<tr>
<td>A3MXB6</td>
<td><em>P. calidifontis</em></td>
<td>144</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 (SEQ ID NO: 1) for homologs described in Table 1, l-Crel (SEQ ID NO: 126) for homologs described in Table 2, Pl-Scel (SEQ ID NO: 145) for homologs described in Table 3, l-Ceil (SEQ ID NO: 131) for homologs described in Table 4, l-Chul (SEQ ID NO: 136) for homologs described in Table 5, or l-Dmol (SEQ ID NO: 141) for homologs described in Table 6. The l-Msol homolog described by SEQ ID NO: 151, might have a different enzymatic activity, DNA-binding-affinity, dimer formation affinity or changes in its DNA recognition sequence, when compared to the reference endonuclease of l-Msol as described by SEQ ID NO: 150. Another preferred endonuclease is l-Ani (SEQ ID NO: 161), preferably comprising the activity enhancing mutations: F13Y and S111Y, or F13Y, S111Y and K222R, or F13Y, I55V, F91I, S92T and S111Y.

Accordingly, in one embodiment of the invention the chimeric endonuclease comprises at least one LAGLIDADG endonuclease comprising an amino acid sequence having at least 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98% or 99% amino acid sequence identity to a polypeptide described by any one of SEQ ID NOs: 1, 2, 3, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 161 or 165.

In another embodiment of the invention the chimeric endonuclease comprises at least one LAGLIDADG endonuclease comprising an amino acid sequence having at least 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98% or 99% amino acid sequence identity to a polypeptide described by any one of SEQ ID NOs: 126, 127, 128, 129, or 130.

In another embodiment of the invention the chimeric endonuclease comprises at least one LAGLIDADG endonuclease comprising an amino acid sequence having at least 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98% or 99% amino acid sequence identity to a polypeptide described by any one of SEQ ID NOs: 131, 132, 133, 134 or 135.

In another embodiment of the invention the chimeric endonuclease comprises at least one LAGLIDADG endonuclease comprising an amino acid sequence having at least 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98% or 99% amino acid sequence identity to a polypeptide described by any one of SEQ ID NOs: 136, 137, 138, 139 or 140.
In another embodiment of the invention the chimeric endonuclease comprises at least one LAGLIDADG endonuclease comprising an amino acid sequence having at least 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98% or 99% amino acid sequence identity to a polypeptide described by any one of SEQ ID NOs: 141, 142, 143, or 144.

In another embodiment of the invention the chimeric endonuclease comprises at least one LAGLIDADG endonuclease comprising an amino acid sequence having at least 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98% or 99% amino acid sequence identity to a polypeptide described by any one of SEQ ID NOs: 145, 146, 147, 148 or 149.

In another embodiment of the invention the chimeric endonuclease comprises at least one LAGLIDADG endonuclease comprising an amino acid sequence having at least 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98% or 99% amino acid sequence identity to a polypeptide described by any one of SEQ ID NOs: 150 or 151.

In another embodiment of the invention the chimeric endonuclease comprises at least one LAGLIDADG endonuclease comprising an amino acid sequence having at least 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98% or 99% amino acid sequence identity to a polypeptide described by any one of SEQ ID NOs: 161, preferably comprising the activity enhancing mutations: F13Y and S111Y, or F13Y, S111Y and K222R, or F13Y, I55V, F91Y, S92T and S111Y.

In another embodiment of the invention the chimeric endonuclease comprises at least one LAGLIDADG endonuclease comprising an amino acid sequence having at least 80% amino acid sequence identity to a polypeptide described by SEQ ID NO: 1, 2, 3 or 165.

Preferred are LAGLIDADG endonucleases for which exact protein crystal structures have been determined, like l-Dmo i, H-Drc i, l-Sce i, l-Crc 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 and which can easily be modeled on crystal structures of l-Dmo i, H-Drc i, l-Sce i, l-Crc i. One example, of an endonuclease, which can be modeled on the crystal structure of l-Crc i, is l-Mso i (Chevalier et al., Flexible DNA Target Site Recognition by Divergent Homing Endonuclease Isoschizomers l-Crel and l-Msol, J. Mol. Biol. (2003) 329, pages 253-269).

Another way to create homologs of LAGLIDADG endonucleases is to mutate the amino acid sequence of a LAGLIDADG endonuclease in order to modify its DNA binding affinity, its dimer formation affinity or to change its DNA recognition sequence.

Homologs of LAGLIDADG endonucleases, which have been mutated in order to modify their DNA binding affinity, its dimer formation affinity or to change its DNA recognition sites 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. The determination of protein structure as well as sequence alignments of homologs of LAGLIDADG 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.


Engineered versions of l-Scel, l-Crel, l-Msol and l-Ceul 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 hystidine.

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-
sequence are disclosed for example in WO07/047859 and WO09/076292.

For example, an important DNA recognition site of l-Scel has the following sequence (described by SEQ ID NO: 12):

5
sense: 5' - T T A C C C T G T T A T C C C T A G - 3'
base position:  1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
antisense 3' - A T G G A C A T A G G A T C - 5'

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, E66.
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, L98.

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 keep 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, C193, V193, A193, T193, S193.

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 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 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 LAGLIDADG endonucleases the following mutations should be avoided:


For l-Crel: Q47E,
for l-Ceu l E66Q,
for l-Msol D22N,
for PI-Scel mutations in D218, D229, D326 or T341.

Engineered endonuclease variants of l-Anil (SEQ ID NO: 161), 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-Anil 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 or cryptic splice sites or cryptic miRNA targets 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 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 glycine as the second N-terminal amino acid, or
- any combination of a), b), c) d), e) and f) when compared to the amino acid sequence of the non optimized nuclease.

PEST Sequences are defined as hydrophilic stretches of at least 12 amino acids length with 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 LAGLIDAG endonuclease is deleted, it is important to retain the LAGLIDAG 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, LAYWFMDGGK, KTIIPNVLVENYLTPMSLAYWFMDGGK, KPIIYIDSMSYLIFYNLK, KLPNTISSETFLK or TISSETFLK,

or which does not comprise an amino acid sequence described by the sequence: HVCLLYDQWVLSPPH, LAYWFMDGGK, KPIIYIDSMSYLIFYNLK, KLPNTISSETFLK or TISSETFLK,

or which does not comprise an amino acid sequence described by the sequence: HVCLLYDQWVLSPPH, LAYWFMDGGK, KLPNTISSETFLK or TISSETFLK,

or which does not comprise an amino acid sequence described by the sequence: LAYWFMDGGK, KLPNTISSETFLK or TISSETFLK,

or which does not comprise an amino acid sequence described by the sequence: KLPNTISSETFLK or TISSETFLK,

In one embodiment the optimized nuclease is l-Scel, 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 l-Scel or its homologs having at least 49%, 51%, 58%, 60%, 70%, 80%, 85%, 90%, 92%, 93%, 94%, 95%, 96%, 98% or 99% sequence identity on amino acid level and having 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 l-Scel 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 TISSETFLK at the C-terminus.

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

<table>
<thead>
<tr>
<th>Wildtype and optimized l-Scel</th>
<th>Amino Acid Sequence on C-terminus</th>
</tr>
</thead>
<tbody>
<tr>
<td>l-Scel wildtype</td>
<td>TISSETFLK</td>
</tr>
<tr>
<td>l-Scel -1</td>
<td>TISSETFL</td>
</tr>
<tr>
<td>l-Scel -2</td>
<td>TISSETF</td>
</tr>
<tr>
<td>l-Scel -3</td>
<td>TISSET</td>
</tr>
<tr>
<td>l-Scel -4</td>
<td>TISSE</td>
</tr>
<tr>
<td>l-Scel -5</td>
<td>TISS</td>
</tr>
<tr>
<td>l-Scel -6</td>
<td>TIS</td>
</tr>
<tr>
<td>l-Scel -7</td>
<td>T</td>
</tr>
<tr>
<td>l-Scel -8</td>
<td>T</td>
</tr>
<tr>
<td>l-Scel -9</td>
<td>all 9 amino acids on C-terminus of wt l-Scel 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: 37), or AIANQAFLK (SEQ ID NO: 38).

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 M202K, M202H and M202R.

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 TIK-SETFL 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 TIASETFL 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 homologs of l-Scel having at least 49%, 51%, 58%, 60%, 70%, 80%, 85%, 90%, 92%, 93%, 94%, 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 l-Scel 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: 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, 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 l-Scel 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: l-Scel -1, l-Scel -2, l-Scel -3, l-Scel -4, l-Scel -5, l-Scel -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 l-Scel, 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) l-Scel -1, l-Scel -2, l-Scel -3, l-Scel -4, l-Scel -5, l-Scel -6, l-Scel -7, l-Scel -8 and l-Scel -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 comprise at least one heterologous DNA binding domain comprising one or more Zη2Cβ zinc fingers.

Zη2Cβ zinc fingers form a unique group of DNA binding domains, which are more or less exclusively found in transcription factors of yeasts and fungi. They are characterized by a shared structure in which two Zinc ions are complexed by an amino acid motif described by the general formula:

-Cys-(X)2-Cys-(X)6-Cys-(X)4i,-Cys-(X)2-Cys-(X)14,-Cys-, 

wherein Cys stands for Cysteine and X for any amino acid.

Mutational analysis of the 6 Cysteines, which are spaced by two, six, five to forty one, two and six to eight other amino acids, has shown that all of them are necessary for complexation of the two Zinc ions, which in turn will facilitate correct folding of a cloverleaf-shaped structure. Basic residues usually predominate at the first, third, fourth, and sixth residues between the second and third cysteines. Nonconservative mutations in either of the third or the fourth of these residues frequently abolish the DNA binding capacity of this DNA binding domain. Amino acid residues between the second and third cysteines are usually basic, in particular at the first, third, fourth, and sixth position. Structural studies have shown that these basic residues frequently form contacts to the DNA. The loop region between the third and fourth Cysteine shows variability in length and sequence and is known as a variable subregion (see Figures 3 to 6). Mutations in the variable subregion generally show little effect, but some mutations reduce function.

In this variable subregion a proline residue is found in many cases one or two residues N-terminal to the fourth cysteine. This proline it thought to support a turn of the amino acid chain, which is necessary for correct folding, however the conserved proline is not absolutely required and can in many cases be replaced for example with leucine, glutamine, or arginine, in particu-
lar, if other prolines are present in the variable subregion.

Because of their structure, Ζ η 2 zinc fingers can be easily distinguished from other zinc comprising DNA binding domains e.g. of the C2H2- or the CCHC-type, which are disclosed for example in WO07/014275, WO08/076290, WO08/076290 or WO03/062455.

Ζ η 2 zinc fingers bind in the majority of cases to DNA binding sites having trinucleotides of guanine and cysteine in their core region, e.g. CGG or CGA; however, other terminal trinucleotides are found in some binding sites, such as GGG or GGA, or even TCC, TCG, GCC or GCA.

Many transcription factors comprising Ζ η 2 zinc fingers in or as their DNA binding domains have been described in the art, e.g. in WO 02/24865. Non limiting examples of Ζ η 2 zinc finger comprising transcription factors are proteins described by SEQ ID NO: 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 177, 178, 179, 180, 181, 182, 183, 184 and 185.

Preferred heterologous DNA binding domains comprise transcription factors or their DNA binding domains comprising Ζ η 2 zinc fingers domains, which form additional contacts via a N-terminal or C-terminal repeats of basic amino acids like argine or lysine (R or K in one letter code) and/or have the capacity to bind as monomers to DNA. Examples for those type of Ζ η 2 zinc finger transcription factors are AfIR, ArgR, Hap1 or Leu3.

In an preferred embodiment, the heterologous DNA binding domain of the chimeric LAGLIDADG endonuclease comprises AlcR as described by SEQ ID NO: 6 and homologs of AlcR 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 some embodiments of the invention it will be of advantage to use the full length sequence of the transcription factor comprising one or more Ζ η 2 zinc finger or a large fragment thereof, e.g. in cases where the DNA binding activity of the transcription factor or its fragment is inducible. For example, the DNA binding activity of AlcR is inducible by ethanol, acetaldehyde, threonine, ethylamine, propan-1-ol and butan-2-ol.

Accordingly, in one embodiment of the invention, at least one heterologous DNA binding domain of the chimeric LAGLIDADG endonuclease comprises a Ζ η 2 zinc finger transcription factor, or a larger fragment thereof.

A larger fragment of Ζ η 2 zinc finger transcription factor, means a fragment of at least 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95%, 97% or 98% of the amino acid sequence of the wild-type Ζ η 2 zinc finger transcription factor.

In one embodiment of the invention, at least one heterologous DNA binding domain of the chimeric LAGLIDADG endonuclease comprises a Ζ η 2 zinc finger transcription factor having an inducible DNA binding activity.

In one embodiment of the invention, at least one heterologous DNA binding domain of the chimeric LAGLIDADG endonuclease comprises AlcR, AfIR, Hap1, Leu3, or a homolog of any one of 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 one embodiment of the invention, at least one heterologous DNA binding domain of the chimeric LAGLIDADG endonuclease comprises a larger fragment of AlcR, AfIR, Hap1, Leu3, or a homolog of any one of 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.

However, in other cases it will be preferable to use one or several small heterologous DNA binding domains. Accordingly in one embodiment of the invention, the heterologous DNA binding domain comprises more or less only the DNA binding domain of a Zn\(\eta_2\)C\(\delta\) zinc finger comprising transcription factor i.e. a DNA binding domain fragment.

Accordingly, in an equally preferred embodiment, the heterologous DNA binding domain of the chimeric LAGLIDADG endonuclease comprises a DNA binding domain fragment of AlcR (AlcR 1 to 60) as described by SEQ ID NO: 70 and homologs of AlcR (1-60) 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.

The DNA binding domain fragment of AlcR belongs to a group of homologs, which can be described by the consensus sequence shown in Figure 3. Exemplary members of this group comprise an amino acid sequence as described by any one of SEQ ID NOs: 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, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120 and 121 or homologs of any one of 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.

Further DNA binding domain fragments of Zn\(\eta_2\)C\(\delta\) zinc finger comprising transcription factors, as well as their consensus sequences are described by Figures 4, 5 and 6.

Accordingly in another embodiment of the invention, the heterologous DNA binding domain of the chimeric LAGLIDADG endonuclease comprises an amino acid sequence selected from the group of sequences described by SEQ ID NO: 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, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120 and 121 or homologs of any one of 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 chimeric endonuclease comprises a heterologous DNA binding domain comprising at least one polypeptide having at least 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99% amino acid sequence identity to a polypeptide described by any one of SEQ ID NOs: 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68 or 69.
In another preferred embodiment, the chimeric endonuclease comprises a heterologous DNA binding domain comprising at least one polypeptide having at least 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99% amino acid sequence identity to a polypeptide described by any one of SEQ ID NOs: 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93 or 94.

In another preferred embodiment, the chimeric endonuclease comprises a heterologous DNA binding domain comprising at least one polypeptide having at least 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99% amino acid sequence identity to a polypeptide described by any one of SEQ ID NOs: 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106 or 107.

In another preferred embodiment, the chimeric endonuclease comprises a heterologous DNA binding domain comprising at least one polypeptide having at least 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99% amino acid sequence identity to a polypeptide described by any one of SEQ ID NOs: 108, 109, 110, 11, 112, 113, 114, 115, 116, 117, 118, 119, 120 or 121.

In another preferred embodiment, the chimeric endonuclease comprises a heterologous DNA binding domain comprising at least one polypeptide having at least 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99% amino acid sequence identity to a polypeptide described by any one of SEQ ID NOs: 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74 or 75.

In a further embodiment of the invention, the heterologous DNA-binding domain is selected from the group consisting of: AlcR, 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 AlcR, 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.

A preferred DNA binding domain fragment are amino acids 1 to 60 of AlcR or its homologs 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.

**Preparation of Chimeric LAGLIDADG endonucleases:**
LAGLIDADG endonucleases and heterologous DNA binding domains can be combined in many alternative ways.

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

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 LAGLIDADG 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 LAGLIDADG endonuclease. It is also possible to make alternating combinations of LAGLIDADG endonucleases and heterologous DNA binding domains.

In case the chimeric endonuclease comprises more than one LAGLIDADG endonuclease or more than one heterologous DNA binding domain, it is possible to use several copies of the same heterologous DNA binding domain or LAGLIDADG endonuclease or to use different heterologous DNA binding domains or LAGLIDADG 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 LAGLIDADG endonuclease or by reducing the number 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). of the entire amino acid sequence of the chimeric endonuclease.

In one embodiment the chimeric LAGLIDADG endonucleases are combinations of: l-Scel and AlcR, or l-Scel and AlcR (1 to 60), or l-Crel and AlcR, or l-Crel and AlcR (1 to 60), or l-Msol and AlcR, or l-Msol and AlcR (1 to 60), wherein AlcR, or AlcR (1 to 60) are fused N- or C-terminal to l-Scel, l-Crel or l-Msol and wherein l-Scel, l-Crel, l-Msol, AlcR, AlcR (1 to 60), include their homologs 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.

A preferred embodiment is a N- or C-terminal fusion of AlcR or amino acids 1 to 60 of AlcR (AlcR 1 to 60) with l-Scel.

Even more preferred is a C-terminal fusion of AlcR or amino acids 1 to 60 of AlcR with l-Scel.

Preferred examples are C-terminal fusion of AlcR or amino acids 1 to 60 of AlcR with l-Scel comprising only one lysine as linker sequence, e.g. as described by VC-SAH 48, 49, 50 and 51, having the amino acid sequence described by SEQ ID No: 7, 8, 8, and 10.

The chimeric endonucleases can be constructed in many ways and combinations. Examples are given by the following structure. N-terminus- LAGLIDADG endonuclease- Zn2C6 zinc finger - C-terminus,

N-terminus- Zn2C6 zinc finger - LAGLIDADG endonuclease - C-terminus
N-terminus- Zn2C6 zinc finger - LAGLIDADG endonuclease - Zn2C6 zinc finger - C-terminus,
other combinations will be possible, wherein one chimeric endonuclease might comprise one or
more Zri2C6 zinc finger at the N- or C-terminus.

In another embodiment of the invention the chimeric LAGLIDADG endonucleases have the following structure:

<table>
<thead>
<tr>
<th>Line</th>
<th>Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>N-terminus-l-Scel - AlcR - C-terminus, or</td>
</tr>
<tr>
<td></td>
<td>N-terminus-l-Scel - AlcR (1 to 60) - C-terminus, or</td>
</tr>
<tr>
<td>10</td>
<td>N-terminus-l-Crel - AlcR - C-terminus, or</td>
</tr>
<tr>
<td></td>
<td>N-terminus-l-Crel - AlcR (1 to 60) - C-terminus, or</td>
</tr>
<tr>
<td></td>
<td>N-terminus- 1Msol - AlcR - C-terminus, or</td>
</tr>
<tr>
<td>15</td>
<td>N-terminus-l-Msol - AlcR (1 to 60) - C-terminus, or</td>
</tr>
</tbody>
</table>

The chimeric LAGLIDADG 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.

Chimeric LAGLIDADG endonucleases having a nuclear localization signal are for example described by SEQ ID NOs: 8, 10, 50, 51, 52, 53.

In a preferred embodiment, the sequences encoding the chimeric LAGLIDADG endonucleases are modified by insertion of an intron sequence. This prevents expression of a functional enzyme in procaryotic 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 chimeric LAGLIDADG endonuclease can be modified by adding a Sec IV secretion signal to the N-, or C-Terminus of the chimeric LAGLIDADG 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 chimeric LAGLIDADG endonuclease can be modified by adding a Sec III secretion signal to the N-, or C-Terminus of the chimeric LAGLIDADG endonuclease. Suitable Sec III 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 chimeric LAGLIDADG 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 in WO 00/02996 and WO05/085417 included herein by reference.

If a Sec IV secretion signal is added to the chimeric LAGLIDADG endonuclease and the chimeric LAGLIDADG endonuclease is intended to be expressed for example in Agrobactenum rhizogenes or in Agrobactenum tumefaciens, it is of advantage to adapt the DNA sequence coding for the chimeric LAGLIDADG endonuclease to the codon usage of the expressing organism. Preferably the chimeric LAGLIDADG endonuclease 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 LAGLIDADG endonuclease does not have a DNA recognition sequence or less preferred DNA recognition sequence in the Agrobacterium genome. In case the chimeric LAGLIDADG 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 LAGLIDADG endonuclease and the heterologous DNA binding domain are connected via a linker polypeptide (linker).

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 GGSGGGSG, or GGSGGSGGG, or GSGSGGSG.

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 the LAGLIDADG endonuclease and the respective DNA recognition site, as well as the heterologous DNA-binding domain and the respective DNA binding site. Suitable bioinformatic tools are for example described in Desjarlais & Berg (1994), PNAS, 90, 2256 to 2260 and in Desjarlais & Berg (1994), PNAS, 91, 11099 to 11103.

DNA recognition sequences of chimeric endonucleases (chimeric recognition sequences):

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 DNA binding sequences of the heterologous DNA binding domains used. It is obvious, 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: "^" indicates the cleavage site of the sequence specific DNA-endonuclease within a DNA recognition sequence, the letter N represents any nucleotide, and can be replaced by A, T, G or C).

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>Organism of origin</th>
<th>DNA recognition sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>I-Crel</td>
<td>Chlamydomonas reinhardtii</td>
<td>5'-CAAAACGTCGTGAGACAGTTTC ^3'</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 recognition 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 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 regions 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.

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: I-Scel - AlcR or I-Scel AlcR (1 to 60), preferably having an amino acid sequence described by SEQ ID NO: 7, 8, 9, 10, 50, 51, 52 and 53.

I-SCEL AlcR or
I-SCEL AlcR (1 to 60) target site 1 cgtgcggatctaggataaaggttaat (SEQ ID NO: 13)
I-SCEL AlcR or
I-SCEL AlcR (1 to 60) target site 2 cgtgcggatctaggataaaggttaat (SEQ ID NO: 14)
I-SCEL AlcR or
I-SCEL AlcR (1 to 60) target site 3 cgtgcggatcgctaggataaaggttaat (SEQ ID NO: 15)
I-SCEL AlcR or
I-SCEL AlcR (1 to 60) target site 4 cgtgcggatcgctaggataaaggttaat (SEQ ID NO: 16)

A chimeric endonuclease having the structure: AlcR(1 to 60)-I-SCEL, preferably having an amino acid sequence described by SEQ ID NO: 54, 55 and 56

AlcR (1-60) I-SCEL or
AlcR-I-SCEL target site 1 cgtgcggatcattacccgttacctccta (SEQ ID NO: 43)
AlcR (1-60) l-Scel or
AlcR-l-Scel target site 2 cgtgcggatcnnattaccctgttatcccta (SEQ ID NO: 44)
AlcR (1-60) l-Scel or
AlcR-l-Scel target site 3 cgtgcggatcnnattaccctgttatcccta (SEQ ID NO: 45)
AlcR (1-60) l-Scel or
AlcR-l-Scel target site 4 cgtgcggatcnnattaccctgttatcccta (SEQ ID NO: 46)

Examples of recognition sequences of heterologous DNA binding domains are:
AlcR and AlcR (1-60) 5'- WGC GG-3'
AfIR 5'-TCGNNNNNCGA-3' (SEQ ID NO: 164)
Hap1 5'-CGGNNNTA-3'
Leu3 5'-RGCCG-3'
wherein A stands for adenine, G for guanine, C for cytosine, T for thymine, W for adenine or thymine, R for guanine or adenine and N for adenine or guanine or cytosine or thymine.

Zn2C6 zinc finger domains homologous to the Zn2C6 zinc finger domains of AlcR, AfIR, Hap1, Leu3 comprising transcription factors and fragments thereof comprising the Zn2C6 zinc fingers will bind to the same or very similar binding sites, like AlcR (1-60) binds to the same or similar recognition sequences as AlcR.

Polynucleotides:
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: 2, 3, 5, 7, 8, 9, 10, 50, 51, 52, 53, 54, 55, and 56 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: 2, 3, 5, 7, 8, 9, 10, 50, 51, 52, 53, 54, 55 and 56.

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 contend of cryptic splice sites, or has a low contend of cryptic polyA sites, or has a low contend of cryptic miRNA targets, 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) pro-moter (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):637-6469).

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 recognition sequence of an endonuclease and a recognition sequence of a heterologous DNA binding domain.
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 I-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, scArc or AlcR.

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

In one embodiment the isolated polynucleotide comprise a sequence of a DNA recognition site or a chimeric recognition site selected from the group comprising: SEQ ID NO: 13, 14, 15, 16, 26, 27, 28, 29, 33, 34, 35, 36, 43, 44, 45 and 46.

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

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 example, 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, identification, or selection of transformed cells, tissues or organism (e.g., plants). The terms "sequence 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 selectable marker confers to the cell or organism a phenotype resulting in a growth or viability difference. The selectable marker may interact with a selection agent (such as a herbicide or antibiotic or pro-drug) to bring about this phenotype. A screenable marker confers to the cell or organism 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 antibiotic, 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 elements of the cytokinin or hormone biosynthesis leading to the production of a plant hormone e.g., 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 selected 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 genome. Screenable marker sequences include but are not limited to reporter genes (e.g. luciferase, glucuronidase, chloramphenicol acetyl transferase (CAT, etc.). Preferred marker sequences include but shall not be limited to:

i) Negative selection marker
As a rule, negative selection markers are useful for selecting cells which have successfully undergone 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 example a herbicide such as phosphonothricin, glyphosate or bromoxynil), a metabolism inhibitor such as 2-deoxyglucose-6-phosphate (WO 98/45456) or an antibiotic such as, for example, tet-
racyclin, 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 transformed 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 phosphinotricin acetyltransferases (PAT), which acetylates the free amino group of the glutamine synthase inhibitor phosphinotricin (PPT) and thus brings about detoxification of PPT (de Block et al. (1987) EMBO J 6:2513-2518) (also referred to as Bialophs- 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-),
- acetalactate 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 phosphotransferase 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 toruloides) 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 isopentenytransferase 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 pigments (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 Symposium, 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 tyrosine to give DOPA and dopaquinone which subsequently form melanine, which is readily detectable,
- 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. 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, Saxifragaceae, 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, Orchidaeae 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 oleracea 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 monocotyledonous 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 Haematococcus, Phaeodactylum 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 nucleic 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 Cloning Techniques, Methods in Enzymology, Vol. 152, Academic Press, hic., San Diego, CA (Berger); Current Protocols in Molecular Biology, F. M. Ausubel et al., eds., Current Protocols, a joint venture between Greene Publishing Associates, Inc. and John Wiley & Sons, Inc. (1998 Supplement), T. Maniatis, E.F. Fritsch and J. Sambrook, Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY (1989), in T.J. Silhavy, M.L. Berman 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 generated 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 transfect 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 11:1553-1558; and Becker D et al. (1994) Plant J 5:299-307. These methods involve penetration of cells by small particles with the nucleic acid either within the matrix of small beads or particles, or on the surface. The biolistic PDS-1000 Gene Gun (Biorad, Hercules, CA) uses helium pressure to accelerate 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 plant protoplasts are also discussed in Lazzeri P (1995) Methods Mol Biol 49:95-106. Preferred general methods which may be mentioned are the calcium-phosphate-mediated transfection,
the DEAE-dextran-mediated transfection, the cationic lipid-mediated transfection, electroporation, 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 Horticulturae 55:5-36 and Potrykus (1990) CIBA Found Symp 154:198.


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), electroporation, the incubation of dry embryos in DNA-containing solution, sonication and microinjection, 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 electroporation of DNA into plant cells, the plasmid used does not need to meet any particular requirement. 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 genome of the plant cell.

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 Agrobacterium. 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 79-1 81), 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 16:917-918). 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.

The use of T-DNA for the transformation of plant cells has been studied and described intensively (EP-A1 120 516; Hoekema, In: The Binary Plant Vector System, Offset-drukkerij Kanters B. V., Alblasserdam, Chapter V; Fraley et al. (1985) Crit Rev Plant Sci 4:1-45 and An et al. (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 nucleases 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 nucleases 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, Macmillian 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 benzamide, 8-hydroxy-2-methylquinazolin-4-one (NU1025), 1,1-b-dihydro-(2H)benzopyrano(4,3,2-de)isoquinolin-3-one (GPI 6150), 5-aminoisoquinolin-3-one, 3,4-dihydro-5-(4-(1-piperidiny)butoxy)-1-(2H)-isoquinolinolone, 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 effi-
cacy 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 LAGLIDADG endonucleases:

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

The method comprises the steps of:

a. providing at least one LAGLIDADG 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 LAGLIDADG endonuclease or LAGLIDADG 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 LAGLIDADG endonuclease coding regions of step b) and all heterologous DNA binding domains of step c),
e. expressing a chimeric LAGLIDADG endonuclease from the translational fusion created in step d),
f. testing the chimeric LAGLIDADG 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 LAGLIDADG 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 LAGLIDADG nuclease and the at least one heterologous DNA binding site were arranged in the translational fusion, and testing the chimeric LAGLIDADG endonuclease for cleaving activity on a polynucleotide having potential DNA recognition sites and potential recognition sites for the LAGLIDADG nucleases and heterologous DNA binding domains comprised by the chimeric LAGLIDADG endonuclease and selecting at least one polynucleotide that is cut by the chimeric LAGLIDADG endonuclease.

The method can also be used to design a chimeric LAGLIDADG endonuclease for cleaving activity on a preselected polynucleotide, by first providing a polynucleotide having a specific sequence, thereafter selecting at least one LAGLIDADG 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 LAGLIDADG endonuclease and the at least one heterologous DNA binding domain, expressing the chimeric LAGLIDADG endonuclease encoded by said translational fusion and testing the chimeric LAGLIDADG endonuclease of cleavage activity on the preselected polynucleotide sequence, and selecting a chimeric LAGLIDADG endonuclease having such cleavage activity.

This method can be used to design a chimeric LAGLIDADG 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 LAGLIDADG endonuclease comprising the nuclease and the heterologous DNA binding domain. Alternatively, this method can also be used to create a chimeric LAGLIDADG 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, it is possible to identify a LAGLIDADG 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 LAGLIDADG endonuclease comprising the identified LAGLIDADG endonuclease and the heterologous DNA binding domain, it will be possible to test the chimeric LAGLIDADG 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 LAGLIDADG endonucleases and recognition sites of DNA binding proteins like ζηεε transcription factors. Further, it is possible to mutate the amino acid sequence of LAGLIDADG endonucleases, like l-Scel, l-Crel, l-Dmol or l-Msol to create new binding and DNA cleavage activity. By creating chimeric LAGLIDADG endonucleases comprising endonucleases like l-Scel, l-Crel, l-Dmol or l-Msol and heterologous DNA binding domains it is possible to create chimeric LAGLIDADG endonucleases which will bind and cleave such preselected polypeptides.

Methods for homologous recombination and targeted mutation using chimeric LAGLIDADG 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 LAGLIDADG endonuclease or an expression cassette coding for a chimeric LAGLIDADG 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: 13, 14, 15, 16, 26, 27, 28, 29, 43, 44, 45 or 46.

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: SEQ ID NO: 13, 14, 15, 16, 26, 27, 28, 29, 43, 44, 45 or 46.

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: 13, 14, 15, 16, 26, 27, 28, 29, 43, 44, 45 or 46.

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.

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.

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, preferably selected from the group of sequences described by SEQ ID NO: 13, 14, 15, 16, 26, 27, 28, 29, 43, 44, 45 or 46.

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, 5, 7, 8, 9, 10, 50, 51, 52, 53, 54, 55, and 56 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 de-
scribed 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" 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. A sequence stretch of "NNNNNNNNNN" in sequences of the sequence protocol is meant to be a placeholder for genes encoding the different versions of the sequence specific DNA-endonuclease. The different genes can be expressed from the Arabinose inducible pBAD promoter (Guzman et al., J Bacterioi 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 the placeholder is replaced by the sequence of I-Scel (SEQ ID NO: 18), was called VC-SAH40-4.

Example 1b: I-Scel - AlcR fusion constructs
In Gene 73 (2), 385-396 (1988) Felenbok et al. described the AlcR Protein as a transcriptional activator in A. nidulans. The AlcR encoding sequence was fused to the C terminus of the I-Scel
sequence, with a single lysine as linker. 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 AlcR. AlcR function can be regulated by the addition of ethanol. 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-SAHI-40. The sequence of the construct is identical to the sequence of construct I, whereas the sequence "NNNNNNNNNN" was replaced by the sequence described by SEQ ID NO: 19.

A similar construct was generated, which in addition to the latter contains a NLS sequence. The resulting plasmid was called VC-SAHI-50-37. The sequence of the construct is identical to the sequence of construct I, whereas the sequence "NNNNNNNNNN" was replaced by the sequence described by SEQ ID NO: 20.

The first 60 amino acids of AlcR represent the DNA binding domain of the protein, so another construct was generated, where only those first 60 amino acids are fused to the C-terminus of l-Scel to create an l-Scel - AlcR (1-60) fusion. The resulting plasmid was called VC-SAHI-49-1. The sequence of the construct is identical to the sequence of construct I, whereas the sequence "NNNNNNNNNN" was replaced by the sequence described by SEQ ID NO: 21.

A similar construct was generated, which in addition to the latter contains a NLS sequence. The resulting plasmid was called VC-SAHI-48-8. The sequence of the construct is identical to the sequence of construct I, whereas the sequence "NNNNNNNNNN" was replaced by the sequence described by SEQ ID NO: 22.

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: 23 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: 24), was called VC-SAHI-6-1. A control plasmid without a target site was called VC-SAHI-7-1 (SEQ ID NO 25)

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

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

In Structure 9, 827-36 (2001) Cahuzac et al. described the DNA binding domain of AlcR in complex with its cognate recognition sequence. Based on this information, combined target sites were generated, that consist of the target site of the nuclease l-Scel and AlcR. 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 plas-
mids were called VC-SA56-1, VC-SA57-2, VC-SA58-2, VC-SA59-1. 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: 26, NO: 27, NO: 28, NO: 29 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 cassettes 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 OD600 =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 μl of each aliquot were plated on LB solid media, supplemented with Ampicillin and Kanamycin. The plates were incubated over night at 37 ºC 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-AlcR fusions: E. coli/growth assay indicates endonuclease activity (enzymatic activity) against the respective target sites.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>VC-SA7-1</td>
<td>++</td>
<td>++</td>
<td>++</td>
<td>++</td>
<td>++</td>
</tr>
<tr>
<td>VC-SA6-1</td>
<td>_</td>
<td>+</td>
<td>x</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 $OD_{600}$ 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. VIS-02]). 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 nptll 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: 31 shows a sequence stretch of "NNNNNNNNNNNN". 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: l-Scel - AlcR fusion constructs

The sequence stretch of "NNNNNNNNNNN" of "Construct IV" was separately replaced by genes encoding three different versions of l-Scel-AlcR fusions, described in Example 1b. The resulting plasmids were called VC-SAH91-1 (NLS - l-Scel - AlcR(1-60)), VC-SAH92-1 (l-Scel - AlcR(1-60)), VC-SAH103-3 (NLS - l-Scel - AlcR) and VC-SAH104-22 (l-Scel - AlcR).

Example 7: Constructs harboring nuclease recognition sequences/target sites to monitor nuclease 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: 32. 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 et al. (1987) EMBO J 6:3901-3907.

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

Combined target sites were generated, that consist of the target site of the nuclease I-Scel and AlcR. 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 I-Scel fusion protein.

The resulting plasmids were/are called VC-SAH52-2 1, VC-SAH 111, VC-SAH 112, VC-SAH55-22. The sequence of the constructs is identical to the sequence of construct V, whereas the sequence "NNNNNNNNNNN" was replaced by the sequences described by SEQ ID NO: 33, NO: 34, NO: 35, NO: 36 respectively.

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

Plasmids VC-SAH87-4, VC-SAH91-1, VC-SAH92-1, VC-SAH 103-3, VC-SAH 105, VC-SAH 140, VC-SAH 139-20, VC-SAH89-10, VC-SAH 90 were/ are transformed into A. thaliana according to the protocol described in Example 5. Selected transgenic lines (T1 generation) were/ are grown in the greenhouse and some flowers were/ are used for crossings (see below).

Example 9: Transformation of constructs harboring combined target sites to monitor recombination into A. thaliana

Plasmids VC-SAH52-21, VC-SAH1 11, VC-SAH 112, VC-SAH55-22, VC-SAH 113, VC-SAH 114, VC-SAH 115, VC-SAH16-4, VC-SAH17-8, VC-SAH18-7 and VC-SAH19-15 were/ are transformed into A. thaliana according to the protocol described in Example 5. Selected transgenic lines (T1 generation) were/ are grown in the greenhouse and some flowers were 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 end-
donuclease were/ 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 AlcR fusions, transgenic lines of Arabidopsis harboring the T-DNA of the nuclease encoding constructs VC-SAH91-1 and VC-SAH87-4 were crossed with lines of Arabidopsis harboring the T-DNA of constructs VC-SAH52-21, VC-SAH55-22 and VC-SCB734-4, harboring the target sites. To visualize the activity of additional AlcR fusions, transgenic lines of Arabidopsis harboring the T-DNA of the nuclease encoding constructs VC-SAH91-1, VC-SAH92-1 and VC-SAH103-3 were crossed with lines of Arabidopsis harboring the T-DNA of constructs VC-SCB743-4, VC-SAH52-21 and VC-SAH55-22, harboring the target sites.

F1 seeds of the crosses were harvested. The seeds were surface sterilized and grown on medium A supplemented with the respective antibiotics and/or herbicides. Leaves were/ 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 can 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 were 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.

RESULTS:
Three independent lines harbouring the T-DNA of the nuclease encoding construct VC-SAH91-1 (NLS-I-Scel - AlcR(1-60)) were crossed with tree independent lines harbouring the T-DNA of the nuclease encoding construct VC-SAH743-4 (native l-Scel site) and with tree independent lines harbouring the T-DNA of the nuclease encoding construct VC-SAH55-22 (target site l-Scel - AlcR).

Leaves were harvested and used for histochemical GUS staining. The combination of NLS - l-Scel - AlcR(1-60), alternatively called I-Scel - AlcR(1-60), with the combined target site (SAH55-22 also called #55) resulted in 49% blue plants, whereas the crossing to the native l-Scel site resulted only in 3% blue plants.

In contrast, the native version of the nuclease I-Scel, alternatively called wt I-Scel, showed only very little recombinant activity (0% for the native (wt) target site and 8% for the native target site comprised in the combined target site SAH55-22).

See Figure 2

Example 20: stabilized versions of l-Scel fused to AlcR
Example 20 a): Constructs harboring sequence specific DNA-endonuclease expression cas-
settes, encoding the C terminal shortened version of l-Scel fused to AlcR for expression in E. coli.

To generate l-Scel variants with an increased specificity and stability, fusion proteins were generated between AlcR and versions of l-Scel, where the proposed C-terminal PEST sequence (amino acids 228-236) is altered. Two different alterations of the C terminus were chosen, in the first one (C term mod #1) the C terminus was replaced by the SEQ ID NO: 37, in the second one (C term mod #2) the C terminus was replaced by the SEQ ID NO: 38. The resulting plasmids are based on the construct described in Example 1a), where the placeholder was replaced by genes encoding the different l-Scel fusions. In the plasmid named VC-SAH128-3, NNNNNN was replaced by NLS - l-Scel C term mod #1 AlcR (1-60) (SEQ ID NO:39). In the plasmid named VC-SAH129-1 , NNNNNN was replaced by NLS - l-Scel C term mod #2 AlcR (1-60) (SEQ ID NO:40).

Also plasmids with full length versions of AlcR are generated. These are named VC-SAH130-30, where NNNNNN was replaced by NLS - l-Scel C term mod #1 AlcR (SEQ ID NO:41) and VC-SAH131-6, where NNNNNN was replaced by NLS - l-Scel C term mod #2 AlcR (SEQ ID NO:42).

Similar constructs were generated, which encode a fusion of AlcR(1-60) to the N terminus of l-Scel, which has a deletion of the C terminal PEST sequence (amino acids 228-236).

Tree variants were generated: VC-SAH 186-5 a direct fusion, VC-SAH 185-1 a version with one amino acid linker and VC-SAH187-10 a version with a linker of 3 amino acids (SEQ ID NO: 47, 48 and 49).

To tests these N terminal fusions of AlcR to l-Scel, additional target sites had to be created. The resulting target vectors were the following:

VC-SAH181-1 (CGTGCGGATCATACCTCGATTACCTCCCTA) (SEQ ID NO:43)
VC-SAH182-2 (CGTGCGGATCNATTACCTCGATTACCTCCCTA) (SEQ ID NO:44)
VC-SAH183-3 (CGTGCGGATCNATTACCTCGATTACCTCCCTA) (SEQ ID NO:45)
VC-SAH184-2 (CGTGCGGATCNATTACCTCGATTACCTCCCTA) (SEQ ID NO:46)

Example 20 b): Demonstration of the endonuclease activity in E. coli

The nuclease versions with a C terminal fusion of AlcR, encoded by plasmids described in example 20a, were cotransformed in E. coli with vectors VC-SAH56-1, VC-SAH57-2, VC-SAH 58-2, VC-SAH59-1, encoding the combined target sites. The activity and specificity of these versions of l-Scel was analyzed as described in Example 3 and 4.

RESULTS:

In E. coli, the C terminal l-Scel - AlcR Fusions described in Example 20a) behaved comparable to VC-SAH48 to VC-SAH51, which are shown in Example 4, tablel.

VC-SAH128-3, VC-SAH1 29-1, VC-SAH1 30-30, SAH131-6 did all cut the combined target site encoded by VC-SAH59-1 with highest efficiency. Whereas the activity on the native l-Scel target site was much lower than by the native nuclease.
Example 20 c): Constructs harboring sequence specific DNA-endonuclease expression cassettes, encoding the C terminal shortened version of l-Scel fused to AlcR for expression in A. thaliana

The versions of l-Scel -AlcR fusions described in Example 20a were cloned into Construct IV. In the plasmid named VC-SAH126-1 , the placeholder was replaced by NLS - l-Scel C term mod #1 AlcR (1-60) (SEQ ID NO:39). In the plasmid named VC-SAH127-1 , NNNNNNN was replaced by NLS - l-Scel C term mod #2 AlcR (1-60) (SEQ ID NO:40). Also plasmids with full length versions of AlcR were generated. These are named VC-SAH137-1, where NNNNNNN was replaced by NLS - l-Scel C term mod #1 AlcR (SEQ ID NO:41) and VC-SAH138-2, where NNNNNNN was replaced by NLS - l-Scel C term mod #2 AlcR (SEQ ID NO:42).

Example 20 d): Demonstration of the endonuclease activity in A. thaliana

Plasmids SAH126-1 and VC-SAH 127-1 were transformed in A. thaliana as described in Example 5). Plants were crossed with lines carrying the T-DNA with a reporter construct encompassing cognate target sites, as described in Example 7b). Similarly, VC-SAH137-1 and VC-SAH138-2 were transformed in A. thaliana. Plants were crossed with lines carrying the T-DNA with a reporter construct encompassing cognate target sites, as described in Example 7b). Activity and specificity of these versions of the nuclease is analyzed as described in Example 10. 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 AlcR fusions, transgenic lines of Arabidopsis harboring the T-DNA of the nuclease encoding constructs SAH126-1 and VC-SAH127-1 were crossed with lines of Arabidopsis harboring the T-DNA of constructs VC-SAH55-22 and VC-SCB734-4, harboring the target sites.

F1 seeds of the crosses were harvested. The seeds were surface sterilized and grown on medium A supplemented with the respective antibiotics and/or herbicides. Leaves were 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 was observed by comparing these results with control crosses. All transgenic lines of Arabidopsis harboring the T-DNA of constructs encoding the different fusions of l-Scel tested so far, were also 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 was analyzed for fully blue seedlings.
RESULTS:

Three independent lines harbouring the T-DNA of the nuclease encoding construct SAH126-1 (NLS - l-Scel C term mod #1 AlcR (1-60)) and three independent lines harbouring the T-DNA of the nuclease encoding construct SAH127-1 (NLS - l-Scel C term mod #2 AlcR (1-60)) were crossed with tree independent lines harbouring the T-DNA of the nuclease encoding construct VC-SAH743-4 (native l-Scel site) and with tree independent lines harbouring the T-DNA of the nuclease encoding construct VC-SAH55-22 (target site l-Scel - AlcR).

Leaves were harvested and used for histochemical GUS staining. The combination of NLS - l-Scel C term mod #1 - AlcR(1-60), alternatively called l-Scel#1 - AlcR(1-60), with the combined target site (SAH55-22 alternatively called #55) resulted in 100% blue plants, whereas the crossing to the native l-Scel gave 0% blue plants.

The combination of NLS - l-Scel C term mod #2 - AlcR(1-60), alternatively called l-Scel#2 - AlcR(1-60), with the combined target site (SAH55-22) resulted in 76% blue plants, whereas the crossing to the native l-Scel gave 0% blue plants.

See Figure 2
1. A chimeric endonuclease comprising at least one LAGLIDADG endonuclease and at least one heterologous DNA binding domain comprising one or more Zn\textsubscript{2}C\textsubscript{6} zinc fingers.

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

3. A chimeric endonuclease as claimed in claim 1, wherein at least one LAGLIDADG endonuclease, which is an engineered or optimized endonuclease or an optimized version of an engineered endonuclease.

4. A chimeric endonuclease as claimed in any one of claims 1 to 3, comprising a heterologous DNA binding domain comprising one or more Zn\textsubscript{2}C\textsubscript{6} zinc fingers derived from a transcription factor.

5. A chimeric endonuclease as claimed in any one of claims 1 to 4, wherein the heterologous DNA binding domain comprises at least one polypeptide having at least 80% amino acid sequence identity to a polypeptide described by any one of SEQ ID NOs: 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, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120 or 121.

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

7. A chimeric endonuclease as claimed in any one of claims 1 to 6, wherein the endonuclease having DNA double strand break inducing activity and the heterologous DNA binding domain are connected via a linker polypeptide.

8. A chimeric endonuclease as claimed in any one of claims 1 to 7, wherein the linker polypeptide, consists of at least 3 amino acids and wherein the at least one third of the amino acids in the amino acid sequence of this linker polypeptide are glycine or serine or alanine or a combination of glycine, serine and alanine.

9. A chimeric endonuclease as claimed in any one of claims 1 to 8, comprising at least one NLS-sequence.

10. 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.
11. A chimeric endonuclease as claimed in any one of claims 1 to 10, wherein the DNA double strand break inducing activity of the endonuclease is inducible by expression of the second monomer of a dimeric or heterodimeric LAGLIDADG endonuclease.

12. A chimeric endonuclease as claimed in any one of claims 1 to 11, further comprising at least one Sec t or SecIV secretion signal.

13. An isolated polynucleotide comprising a nucleotide sequence, which codes for a chimeric endonuclease as claimed in any one of claims 1 to 12.

14. An isolated polynucleotide comprising a nucleotide sequence, as claimed in claim 13, 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).

15. An expression cassette comprising an isolated polynucleotide as claimed in claim 13 or 15 in functional combination with a promoter and an terminator sequence.

16. 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 a LAGLIDADG endonuclease and
   b. a recognition sequence of a heterologous DNA binding domain comprising one or more zinc fingers.

17. An isolated polynucleotide comprising a chimeric recognition sequence as claimed in claim 16, wherein the recognition sequence of a heterologous DNA binding domain can be bound by at least one DNA binding domain comprising an amino acid sequence described by any one of SEQ ID NOs: 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, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121.

18. An isolated polynucleotide comprising a chimeric recognition sequence as claimed in any one of claims 16 or 17, comprising,
   a. a DNA recognition sequence of l-Scel,
   b. a recognition sequence of AlcR or AlcR (1-60) being directly connected or being connected by a sequence of 1 to 10 nucleotides.
19. An isolated polynucleotide comprising a chimeric recognition sequence as claimed in any one of claims 16 to 19, comprising a polynucleotide sequence as described by any one of SEQ ID NO: 13, 14, 15, 16, 43, 44, 45 or 46.

20. 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 12, or
   b. an isolated polynucleotide as claimed in claims 13 or 14, or
   c. an expression cassette as claimed in claim 15, or
   d. an isolated polynucleotide comprising a chimeric recognition sequence as claimed in any one of claims 16 to 19, or
   e. any combination of a), b), c) and d).

21. A non-human organism as claimed in claim 20, wherein the non-human organism is a plant.

22. 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 endonuclease 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).

23. 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 16 to 19 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 claimed in any one of claims 1 to 12 or an expression cassette as claimed in claim 15,
   e. combining the polynucleotides of b), c) and the chimeric endonuclease of d) in said cell and
24. A method for homologous recombination of polynucleotides as claimed in claim 23 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).

25. 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 as claimed in any one of claims 1 to 12 being able to cleave the chimeric 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 growing cells comprising mutated polynucleotides.

26. A method for homologous recombination or targeted mutation as claimed in any one of claims 23 to 25, 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 Sec11 or SecIV peptide fused to the chimeric endonuclease.
Figure 1: Model of a chimeric nuclease comprising I-Scel as N-terminal and AlcR (1-60) as a C-terminal domain.
Figure 2:

Direct comparison of favorite variants in planta

% blue plants

<table>
<thead>
<tr>
<th>target site</th>
<th>wt</th>
<th>#55</th>
</tr>
</thead>
<tbody>
<tr>
<td>nuclease</td>
<td>I-SceI</td>
<td>I-SceI – AlcR(1-60)</td>
</tr>
<tr>
<td>wt</td>
<td>wt I-SceI</td>
<td>I-SceI – AlcR(1-60)</td>
</tr>
</tbody>
</table>

wt target site: attacacctgtatcccta
#55 target site: attacacctgtatccctagcggtccgcacg

I-SceI#1 – AlcR(1-60)  KPYLFPQMMYKLFENTKSEITFLKLMAKTRR
I-SceI#2 – AlcR(1-60)  KPYLFPQMMYKLFENAQAPFLKLMAKTRR
Figure 3: Sequence alignment of different I-Scel homologs

Consensus

1 2 3 4 5
1 MK IQNQIMNLGPGSKLLKEYKSQLT LT EQ EAGILIGDAYIE RSRDEGTYCMQ FEWKKNAYIDHVCLLYDEWVLSPHKKE
2 MK IQNQIMNLGPGSKLLKEYKSQLT LT EQ EAGILIGDAYIE RSRDEGTYCMQ FEWKKNAYIDHVCLLYDEWVLSPHKKE
3 MK IQNQIMNLGPGSKLLKEYKSQLT LT EQ EAGILIGDAYIE RSRDEGTYCMQ FEWKKNAYIDHVCLLYDEWVLSPHKKE
4 MTMYIQNQIMNLGPGSKLLKEYKSQLT LT EQ EAGILIGDAYIE RSRDEGTYCMQ FEWKKNAYIDHVCLLYDEWVLSPHKKE
5 MK IQNQIMNLGPGSKLLKEYKSQLT LT EQ EAGILIGDAYIE RSRDEGTYCMQ FEWKKNAYIDHVCLLYDEWVLSPHKKE

Consensus

1 2 3 4 5
1 RVNLGKNLVLITWGAQTFKHA FPNKLA LFNNMK K I NNLVENYTPLSAYWFMDDGKWDYNKMS NKSIVLNTQ PT EEVEYLI
2 RVNLGKNLVLITWGAQTFKHA FPNKLA LFNNMK K I NNLVENYTPLSAYWFMDDGKWDYNKMS NKSIVLNTQ PT EEVEYLI
3 RVNLGKNLVLITWGAQTFKHA FPNKLA LFNNMK K I NNLVENYTPLSAYWFMDDGKWDYNKMS NKSIVLNTQ PT EEVEYLI
4 RVNLGKNLVLITWGAQTFKHA FPNKLA LFNNMK K I NNLVENYTPLSAYWFMDDGKWDYNKMS NKSIVLNTQ PT EEVEYLI
5 RVNLGKNLVLITWGAQTFKHA FPNKLA LFNNMK K I NNLVENYTPLSAYWFMDDGKWDYNKMS NKSIVLNTQ PT EEVEYLI

Consensus

1 2 3 4 5
1 KGLRNKFLNCYVKINNNKPIIYIDGMSYTLFHYNLKPYLIFEOMYKLPTNTSSSTFLK---
2 KGLRNKFLNCYVKINNNKPIIYIDGMSYTLFHYNLKPYLIFEOMYKLPTNTSSSTFLK---
3 KGLRNKFLNCYVKINNNKPIIYIDGMSYTLFHYNLKPYLIFEOMYKLPTNTSSSTFLK---
4 KGLRNKFLNCYVKINNNKPIIYIDGMSYTLFHYNLKPYLIFEOMYKLPTNTSSSTFLK---
5 KGLRNKFLNCYVKINNNKPIIYIDGMSYTLFHYNLKPYLIFEOMYKLPTNTSSSTFLK---

Consensus

1 2 3 4 5
1 GLN KFNLNC MKFKNKKPIIYIPS SY IFYNLI PYIIPEMYKYLPI S
Figure 4: Sequence alignment of different I-Crel homologs

```
(1)  1  10  20  30  40  50  60  70  89
1   (1) -------- MNTKYKEFLLYLAGFVGGDGALIAQIKPNSYKFKHOLSEAPFQTQKORRWFEDIRLVEIGVGYRDRGDYS
2   (1) -------- MNKFTPDQLELAGLIDGDIAQIVYSEDKYTDQGFLQTTVCOLKRWLEELQKEIIAGSVDVRDTVS
3   (1) -------- MKDQEDLELTTYLAGFIDGDSTFAQESINLKYKTYQCGTTLKXLLFRLEHGDVDHRKDAFSYLV
4   (1) MTNNMQQNKKKIKNKKDALTTYLAGFDIDGSIAIQIVRHKYQCTSVGQKTQKEHFEIQLQSEhvGSSVRNYNLDC
5   (1) --MKNINSTRYSHLTNEQKATLYAGFIDCDCGMAIQTVREDPAFVYQNTLQGQRTSKEHELKEFLAEGYVYWSRN

Consensus (1) M L D LTYLAGFDGDGSIIAQIV R DYKFQIQRLTVQITQLTRK FLE I DEIG G VRKR TVSDYVLTE
```

```
(90)  90  100  110  120  130  140  150  160  178
1   (90) IMLHNPLTGQFPLKLQKQANLVLKIEWRLPSKEXSPDKHELQCTWDQTAALND-SSTKTTTEVRAVLDSSEKSSP-----
2   (90) TSNVKKPLELDPHELKKQKANLVLKIEWRLPSKEXSPDKHELQCTWDQTAALND-TKRYTAELVAKLKEKECVVVPETSEET
3   (90) PPEVQKLQFPLKLQKQANLVLKIEWRLPSKEXSPDKHELQCTWDQTAALND-SKKEHTTSSVWMSLGHELPEKVSKEVNPV
4   (90) PIVWEGLLLLQFPLKLQKQANLVLKIEWRLPSKEXSPDKHELQCTWDQTAALND-SKSHITAEVETTIREEGLI------
5   (90) ANIVQELPLLQFPLKLQKQANLVLKIEWRLPSKEXSPDKHELQCTWDQTAALND-TPKILENIVQKVAEBELSEDLQ------

Consensus (90) K VY LLLQFPLKLQKQANLVLKIEQLPSSK S D FLEC LVDQVA LND SK RK TAEVV A L L
```

```
(179)  179  193
1   (164) --------------
2   (168) NSGI--------
3   (168) ETSDLIEIEEPSSI
4   (168) --------------
5   (167) --------------

Consensus (179)
```
Figure 5a: Sequence alignment of different PI-Scel homologs

<table>
<thead>
<tr>
<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>
<th>89</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) MAGAIENARKEIKRISLEDHAESYGAITYSVSGPVVIAEMNMGCMAMYELVKVGHDNLVGEVIRIDGDKATIQVYEETAGLTVGDPLVR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1) --------------------------GAITYSVSGPVVIAEMNMGCMAMYELVKVGHDNLVGEVIRIDGDKATIQVYEETAGLTVGDPLVR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1) --------------------------ISLEDHAESYGAITYSVSGPVVIAEMNMGCMAMYELVKVGHDNLVGEVIRIDGDKATIQVYEETAGLTVGDPLVR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1) --------------------------GAITYSVSGPVVIAEMNMGCMAMYELVKVGHDNLVGEVIRIDGDKATIQVYEETAGLTVGDPLVR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1) --------------------------GAITYSVSGPVVIAEMNMGCMAMYELVKVGHDNLVGEVIRIDGDKATIQVYEETAGLTVGDPLVR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Consensus (1) --------------------------GAITYSVSGPVVIAEMNMGCMAMYELVKVGHDNLVGEVIRIDGDKATIQVYEETAGLTVGDPLVR

<table>
<thead>
<tr>
<th>(89)</th>
<th>89</th>
<th>100</th>
<th>110</th>
<th>120</th>
<th>130</th>
<th>140</th>
<th>150</th>
<th>160</th>
<th>176</th>
</tr>
</thead>
<tbody>
<tr>
<td>(89) TGGKLSEVLPGLMELTIYDGIQRPLRAIKIESAQSIYIPRGIDTPALDRITKQFTPGKFQVGDHISGDIIYGSVFENSLISSKILLPF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(64) TGGKLSEVLPGLMELTIYDGIQRPLRAIKIESAQSIYIPRGIDTPALDRITKQFTPGKFQVGDHISGDIIYGSVFENSLISSKILLPF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(75) TGGKLSEVLPGLMELTIYDGIQRPLRAIKIESAQSIYIPRGIDTPALDRITKQFTPGKFQVGDHISGDIIYGSVFENSLISSKILLPF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Consensus (89) TGGKLSEVLPGLMELTIYDGIQRPLRAIKIESAQSIYIPRGIDTPALDRITKQFTPGKFQVGDHISGDIIYGSVFENSLISSKILLPF

<table>
<thead>
<tr>
<th>(177)</th>
<th>177</th>
<th>190</th>
<th>200</th>
<th>210</th>
<th>220</th>
<th>230</th>
<th>240</th>
<th>250</th>
<th>264</th>
</tr>
</thead>
<tbody>
<tr>
<td>(177) FRSQGTITWIAPAGEYLDEKILEVEFDGKSGFSLYETMFLRVRPVPTEKLSDAYPLLTGQRVLDDALPVCQGTTICIPAGFCGKCT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(152) FRSQGTITWIAPAGEYLDEKILEVEFDGKSGFSLYETMFLRVRPVPTEKLSDAYPLLTGQRVLDDALPVCQGTTICIPAGFCGKCT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(163) FRSQGTITWIAPAGEYLDEKILEVEFDGKSGFSLYETMFLRVRPVPTEKLSDAYPLLTGQRVLDDALPVCQGTTICIPAGFCGKCT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Consensus (177) FRSQGTITWIAPAGEYLDEKILEVEFDGKSGFSLYETMFLRVRPVPTEKLSDAYPLLTGQRVLDDALPVCQGTTICIPAGFCGKCT

<table>
<thead>
<tr>
<th>(265)</th>
<th>265</th>
<th>270</th>
<th>280</th>
<th>290</th>
<th>300</th>
<th>310</th>
<th>320</th>
<th>330</th>
<th>340</th>
<th>352</th>
</tr>
</thead>
<tbody>
<tr>
<td>(265) VISQSLSKNSDSSLTVGCFACKTNVLMADGSECEIDIECNQYVNGKMDGPRPREVIKLPRGRTMTYTVSVMQKSRKHKDSGSRVPE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(240) VISQSLSKNSDSSLTVGCFACKTNVLMADGSECEIDIECNQYVNGKMDGPRPREVIKLPRGRTMTYTVSVMQKSRKHKDSGSRVPE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(251) VISQSLSKNSDSSLTVGCFACKTNVLMADGSECEIDIECNQYVNGKMDGPRPREVIKLPRGRTMTYTVSVMQKSRKHKDSGSRVPE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Consensus (265) VISQSLSKNSDSSLTVGCFACKTNVLMADGSECEIDIECNQYVNGKMDGPRPREVIKLPRGRTMTYTVSVMQKSRKHKDSGSRVPE
Figure 5b: Sequence alignment of different PI-Scl homologs

Consensus (353) LLKFTCNATHELVVTRPRSLRRTIKGVEYFEVTIFEMQQKAPDGRIVELKVEVS KYPEGPERANELSHEYKASN-KAYF

Consensus (441) EWTIEARDLSLLGHSHVRKATQYTAPLRYNDDFFDYMQKSKFHLTIEGPKVLAYLLGWDIGDLSDRTFSVDSRSTSMLMERVIYEY

Consensus (529) EKLNLAEGKDRKEPQVAKTVNLSSKVVRGNGIRNNLNTELWDAIVGLFGLKGVKNI PSFLSTDINITRETFLEAGLILDSDGVVID

Consensus (617) EHKIKAIKTIIHTISVRDGLVSLARSLGLVVSVNAEPAKVDMNGT KHISYAITMSGGDVLNL SKCAGSKFRPAPAS FVRECRGF
Figure 5c: Sequence alignment of different PI-Scel homologs

Consensus

1  (703) YFEL0ELKEDDYGGITLSDSDDHQFLANQVHHNCGERNGEADVLMVEFPELEYSRMSGTEPIMKRTTLVANTSNPVAAARESIYT
2  (676) YFEL0ELKEDDYGGITLSDSDDHQFLANQVHHNCGERNGEADVLMVEFPELEYSRMSGTEPIMKRTTLVANTSNPVAAARESIYT
3  (689) YFEL0ELKEDDYGGITLSDSDDHQFLANQVHHNCGERNGEADVLMVEFPELEYSRMSGTEPIMKRTTLVANTSNPVAAARESIYT
4  (438) YFEL0ELKEDDYGGITLSDSDDHQFLANQVHHNCGERNGEADVLMVEFPELEYSRMSGTEPIMKRTTLVANTSNPVAAARESIYT
5  (440) YFEL0ELKEDDYGGITLSDSDDHQFLANQVHHNCGERNGEADVLMVEFPELEYSRMSGTEPIMKRTTLVANTSNPVAAARESIYT

Consensus

1  (793) GITLAEYFRDQKGNVSMIADSSSRWAEALREISGRLGEMPADQGFPAVLGAKLASFYERAGKAVALGSPDRTGSVISIVAASPGDF
2  (766) GITLAEYFRDQKGNVSMIADSSSRWAEALREISGRLGEMPADQGFPAVLGAKLASFYERAGKAVALGSPDRTGSVISIVAASPGDF
3  (777) GITLAEYFRDQKGNVSMIADSSSRWAEALREISGRLGEMPADQGFPAVLGAKLASFYERAGKAVALGSPDRTGSVISIVAASPGDF
4  (526) GITLAEYFRDQKGNVSMIADSSSRWAEALREISGRLGEMPADQGFPAVLGAKLASFYERAGKAVALGSPDRTGSVISIVAASPGDF
5  (528) GITLAEYFRDQKGNVSMIADSSSRWAEALREISGRLGEMPADQGFPAVLGAKLASFYERAGKAVALGSPDRTGSVISIVAASPGDF

Consensus

1  (831) EMIPPATLGLGITQVFQWGLDKKLAQRFHPSPINTSVSYSKTYNLNFYDSNYPEFQVLRDRMKE1LSNAAEELEQQVQLVGKSALSDS
2  (854) EMIPPATLGLGITQVFQWGLDKKLAQRFHPSPINTSVSYSKTYNLNFYDSNYPEFQVLRDRMKE1LSNAAEELEQQVQLVGKSALSDS
3  (865) EMIPPATLGLGITQVFQWGLDKKLAQRFHPSPINTSVSYSKTYNLNFYDSNYPEFQVLRDRMKE1LSNAAEELEQQVQLVGKSALSDS
4  (545) 
5  (543) 

Consensus

1  (969) KITLDVATLIKEDFLQQNGSYSYDAGCFIPMKTFDNMRAfSISHDEAQRAVANGANWKLADSGDVKHAVSSSKFEPFGKEEVHEG
2  (942) KITLDVATLIKEDFLQQNGSYSYDAGCFIPMKTFDNMRAfSISHDEAQRAVANGANWKLADSGDVKHAVSSSKFEPFGKEEVHEG
3  (952) KITLDVATLIKEDFLQQNGSYSYDAGCFIPMKTFDNMRAfSISHDEAQRAVANGANWKLADSGDVKHAVSSSKFEPFGKEEVHEG
4  (546) 
5  (543) 

Consensus

1  (1057) KITLDVATLIKEDFLQQNGSYSYDAGCFIPMKTFDNMRAfSISHDEAQRAVANGANWKLADSGDVKHAVSSSKFEPFGKEEVHEG
2  (1023) 
3  (1040) 
4  (546) 
5  (548) 

Consensus (1057)
Figure 6: Sequence alignment of different I-Ceul homologs
Figure 7: Sequence alignment of different I-Chu homologs

Consensus sequence:

MLTQQKDILFGSSLGDGNLTGEKPLCSENTLETSIVFDERTDKEVKRWFNL

Sequence 1:

MLTQQKDILFGSSLGDGNLTGEKPLCSENTLETSIVFDERTDKEVKRWFNL

Sequence 2:

MLTQQKDILFGSSLGDGNLTGEKPLCSENTLETSIVFDERTDKEVKRWFNL

Sequence 3:

MLTQQKDILFGSSLGDGNLTGEKPLCSENTLETSIVFDERTDKEVKRWFNL

Sequence 4:

MLTQQKDILFGSSLGDGNLTGEKPLCSENTLETSIVFDERTDKEVKRWFNL

Sequence 5:

MLTQQKDILFGSSLGDGNLTGEKPLCSENTLETSIVFDERTDKEVKRWFNL

Consensus:

MLTQQKDILFGSSLGDGNLTGEKPLCSENTLETSIVFDERTDKEVKRWFNL

Sequence 1 (90):

INPSLKKFDAMIFTYDQNTQKNKVDVP--VFYQFDTLTPQAYTFIYEDGAFLKINLSAMYCITESFSQGCTITIQFALRTLYNIDTTL

Sequence 2 (90):

INPSLKKFDAMIFTYDQNTQKNKVDVP--VFYQFDTLTPQAYTFIYEDGAFLKINLSAMYCITESFSQGCTITIQFALRTLYNIDTTL

Sequence 3 (90):

INPSLKKFDAMIFTYDQNTQKNKVDVP--VFYQFDTLTPQAYTFIYEDGAFLKINLSAMYCITESFSQGCTITIQFALRTLYNIDTTL

Sequence 4 (90):

INPSLKKFDAMIFTYDQNTQKNKVDVP--VFYQFDTLTPQAYTFIYEDGAFLKINLSAMYCITESFSQGCTITIQFALRTLYNIDTTL

Sequence 5 (90):

INPSLKKFDAMIFTYDQNTQKNKVDVP--VFYQFDTLTPQAYTFIYEDGAFLKINLSAMYCITESFSQGCTITIQFALRTLYNIDTTL

Consensus (90):

INPSLKKFDAMIFTYDQNTQKNKVDVP--VFYQFDTLTPQAYTFIYEDGAFLKINLSAMYCITESFSQGCTITIQFALRTLYNIDTTL

Sequence 1 (179)

TKKTLQDG----RIGYIAIEASSGAFREVIKFLVDGCMRYKVSDFNGHKL

Sequence 2 (179)

TKKTLQDG----RIGYIAIEASSGAFREVIKFLVDGCMRYKVSDFNGHKL

Sequence 3 (179)

TKKTLQDG----RIGYIAIEASSGAFREVIKFLVDGCMRYKVSDFNGHKL

Sequence 4 (179)

TKKTLQDG----RIGYIAIEASSGAFREVIKFLVDGCMRYKVSDFNGHKL

Sequence 5 (179)

TKKTLQDG----RIGYIAIEASSGAFREVIKFLVDGCMRYKVSDFNGHKL

Consensus (179):

TKKTLQDG----RIGYIAIEASSGAFREVIKFLVDGCMRYKVSDFNGHKL

Sequence 1 (252)

TKKTLQDG----RIGYIAIEASSGAFREVIKFLVDGCMRYKVSDFNGHKL

Sequence 2 (252)

TKKTLQDG----RIGYIAIEASSGAFREVIKFLVDGCMRYKVSDFNGHKL

Sequence 3 (252)

TKKTLQDG----RIGYIAIEASSGAFREVIKFLVDGCMRYKVSDFNGHKL

Sequence 4 (252)

TKKTLQDG----RIGYIAIEASSGAFREVIKFLVDGCMRYKVSDFNGHKL

Sequence 5 (252)

TKKTLQDG----RIGYIAIEASSGAFREVIKFLVDGCMRYKVSDFNGHKL

Consensus (252):

TKKTLQDG----RIGYIAIEASSGAFREVIKFLVDGCMRYKVSDFNGHKL
Figure 8: Sequence alignment of different I-Dmol homologs

Consensus

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MHNENVSGL</td>
<td>M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>M</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td>M</td>
<td>A</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td>M</td>
</tr>
</tbody>
</table>

Consensus

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MSVAILLGLIVDGGGLYALRYFGRTEYRVVTQDEAVVEKAVV</td>
<td>M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>M</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td>M</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td>M</td>
</tr>
</tbody>
</table>
Figure 9: Sequence alignment of different I-Msol homologs

Consensus

(1) TL T AAYAGFGLDGSIYAKLI RDFY IKYQISLAISP QRKDKF YLQDIYD L K G LRRKDRDGDIACYTI G H

(90) LSIILP LLYRLIRKKQAN ILHIIN YP A KN FL LVKIVD IQNLKK DE KATNY LLEEF AGKI SSP
Figure 10: Sequence Alignment of AlcR Homologs (AlcR 1 to AlcR 25)

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 |
| (1) MADTEFFQNSCPCPXGERCQDAPREN | (1) MDTQERFFQNSCPCPXGERCQDAPREN | (1) MEAHFQFFQNSCPCPXGERCQDAPREN | (1) MEPHEFFQNSCPCPXGERCQDAPREN | (1) MEDSFFQNSCPCPXGERCQDAPREN | (1) ERFFHFFQNSCPCPXGERCQDAPREN | (1) DSRFFYFQNSCPCPXGERCQDAPREN | (1) KRNNSCPCPXGERCQDAPREN | (1) KRNNSCPCPXGERCQDAPREN | (1) KRNNSCPCPXGERCQDAPREN | (1) KRNNSCPCPXGERCQDAPREN | (1) KRNNSCPCPXGERCQDAPREN | (1) KRNNSCPCPXGERCQDAPREN | (1) KRNNSCPCPXGERCQDAPREN | (1) KRNNSCPCPXGERCQDAPREN | (1) KRNNSCPCPXGERCQDAPREN | (1) KRNNSCPCPXGERCQDAPREN | (1) KRNNSCPCPXGERCQDAPREN | (1) KRNNSCPCPXGERCQDAPREN | (1) KRNNSCPCPXGERCQDAPREN | (1) KRNNSCPCPXGERCQDAPREN | (1) KRNNSCPCPXGERCQDAPREN | (1) KRNNSCPCPXGERCQDAPREN | (1) KRNNSCPCPXGERCQDAPREN | (1) KRNNSCPCPXGERCQDAPREN | (1) KRNNSCPCPXGERCQDAPREN |
| 10 | 20 | 30 | 40 | 50 | 60 | 70 | 85 | 90 | 100 | 110 | 120 | 130 | 140 | 150 | 160 | 170 | 180 | 190 | 200 | 210 | 220 | 230 | 240 | 250 |
| EANENGWVSCSNCFRNEDCTFNVLSQSK | SSNDYTITCSCNCFKYNECFNVLSQNFRA | RVLSAEINLIPCSCNCFKYNECFNVLSQNFRA | SVSGASAARQEAGNRRVLTESNLMPCSCNCFKYNECFNVLSQNFRA | STFSCNCTWKEFTCETFNVLSQSK | NTAARQAYNHSNSCNCFKYKFKTDFWLLSHIP | VAEAINKEVACSNCFKGTKCSFQYVLSS | DKPTVACCTMSLRSMSCTVAVLSAK | FDNVDSRSPSCYCAITNKCMEWSQSK | SLWDILRNSLRNLGADGANGVSLAEHLDEIDSRALCYCRLTRQCTTHWYPRQ | RLQGREDVPAHLNGHAAISLGS | NLFAAPETKQCTSCRETGCTMN | DKTPTVACCTMSLRSMSCTVAVLSAK | GQ | KFSTPCATCNAGRELGTCAVLASSH | NTAARQAYNHSNSCNCFKYKFKTDFWLLSHIP | GQ | KFSTPCATCNAGRELGTCAVLASSH | PCRNKSYNCTFOPSRKS | EAPCSNCRFNSQGLFSSVFKFR | VETCSNCRTNATDEPSVPMFKR | VETCSNCRTNATDEPSVPMFKR | GSSGELPPTLACSLCEKANNFKGPTULT | RCSCMNFQDCFTPYSQ | KRPSSETELQPCSYCTDKQFGMNWYQ | CSTQCGVLDCRGFRQ | PCRNKAGLGEYMPRSSKS |

Consensus:

RRQQ HSCD CRK KRCADAP
Figure 11: Sequence Alignment of AfR Homologs (AfR 1 to AfR 12)

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TRKLFESCISCSRSKVKCNKEKPTCSRRCVRGLPCEMVSRTGRTVRIG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>TRKLFESCMSGSKVKCNKEKPTCSRCAVRGLSCETISRVRTGRTVRIG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>RKLRSCHICANSKVKCNKEKPTCGRCRMRRRLPCRимвRSRRTGRT-----</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>RKLRSCTEASSKVKCGKEKPTCGRCRMRRGAKCTYMASRRTGRT-----</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>RKLRSCTSCASSKVRCTKEKPAACARCIERSGLACQYMVRMRGR-----</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>RKLRSCTSCASSKVRCTKEKPAACARCIERSGLACQYMVSKRMRGR-----</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>RKLRSCTSCASSKVRCTKEKPAACARCIERSGLACQYMVSKRMRGR-----</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>RKLRSCTSCASSKVRCTKEKPAACARCIERSGLACQYMVSKRMRGR-----</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>RKLRSCTSCASSKVRCTKEKPAACARCIERSGLACQYMVSKRMRGR-----</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>TSKRSDCHCIAISKVRCPKEKPTCARMKGRVCEYFVTKRPKR-----</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>QKLRFDCQSARSKVCTKEKPSRCETKIPCPRLQSRPR-----</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>KLRSESCHSSQKIRCTKERPACARCWNLGQYNISRRTGTR-----</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Consensus (1) RKLRSCTSCASSKVRCTKEKPTCARCI RGL C YMVSKR GR
Figure 12: Sequence Alignment of Hap1 Homologs (Hap1 1 to Hap1 13)

<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>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SPLLHMSSDSSKITKRRKRRIPPLSCTICRKRRKVCKDCLRPSHCQQCTKTGVAVHLCHYMEQTWAEEAEKELLK</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>--------SASNKVKKRRIPPLSCTICRKRRKVCKDCTRPCHCNQCTKTGVAVHLCHYMEQSWAEEAEKEM---</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>--------AQMRRKRRNRVPPLSCTICRKRRKVCKDCTRPHQCQCSKTGVAVHLCHYMEQTWAEEAEKELSK</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>------------MSSPTASKRRKRRIPPLSCTICRKRRKVCKDCTRPHCNQCTKTGVAVHLCHYMEQTWAEEAEKELSK</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>--------ANMKRKRKRIPPLSCTICRKRRKVCKDCTRPNCEQQCSTGVAVHLCHYMEQTWAEEAEKELSK</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>-----------KRRRKRVRPLSCTICRKRRKVCDGRPQCQCVKTVGVHLCHYMEQTWAEEAEKELSK</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>--------KVRKRRKRVPPLSCTICRKRRKVCDKSRNFCTQCVKTVAVHLCHYMEQAWAEEAEKELSK</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>--------KVRKRRKRIPPLSCTICRKRRKVCDTTRPHCNQCTKTGVAVHLCHYMEQNWADAEEKELSK</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>--------KVRKRRKRIPPLSCTICRKRRKVCDKTPHNQCIKTGVAVHLCHYMEQWACEGVKEKEM---</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>--------SERIKLKKRRRPVPSCTICRKRRKVKCDKXXPICTNCNKNGVPHLCCYYLEPSWAK</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>--------KRRQRRRPVPSCLNCKKRRKVKCDKFGXSCTGCIKNGVPHLCEYELEPVWKS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>--------QKRQRRRPVPSCLMCKRRKVDKAKPACGCACVNGVPHLCEYEPWASK</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>--------SKVRKRRRQSVPVLACIKRRKVKCDKGGKAPACGCGIRNGVPHLCEYTYPHWVDPK</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Consensus: KVRRKRNPVPSCTICRKRRKVCDKTRP CNQC KTGVAVHLCHYMEQTWAEERAEKEL K
Figure 13: Sequence Alignment of Leu3 Homologs (Leu3 1 to Leu3 14)
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 9/2; C12N 15/9

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)


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>1-21</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>1-26</td>
</tr>
</tbody>
</table>

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

* Special categories of cited documents:
  * A " document defining the general state of the art which is not considered to be of particular relevance
  * B " earlier application or patent but published on or after the international filing date
  * L " document which may throw doubts on priority claim (S) or which is cited to establish the publication date of another citation or other special reason (as specified)
  * O " document referring to an oral disclosure, use, exhibition or other means
  * P " document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

TANG Li

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 Xiucheng Rd., Jimen Bridge, Haidian District, Beijing, China 100088

Facsimile No. 86-10-62019451

Form PCT/ISA /210 (second sheet) (July 2009)
<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>1-26</td>
</tr>
<tr>
<td>A</td>
<td>LIPPOW, Shaun M et al. Creation of a type IIS restriction endonuclease with a long recognition sequence. Nucleic Acids Research. 20 March 2009, Vol.37, No. 9, pages 3061-3073</td>
<td>1-26</td>
</tr>
</tbody>
</table>
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
      - [x] in electronic form

   b. time of filing or furnishing
      - [x] 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:
<table>
<thead>
<tr>
<th>Patent Documents referred in the Report</th>
<th>Publication Date</th>
<th>Patent Family</th>
<th>Publication Date</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>AU 200329051 8 A8</td>
<td>07. 12. 2006</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CA 22091 83 A</td>
<td>11. 07. 1996</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AU 4895096 A</td>
<td>24. 07. 1996</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AU 719001 B</td>
<td>04. 05. 2000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>KR 987000992 A</td>
<td>30. 04. 1998</td>
</tr>
<tr>
<td></td>
<td></td>
<td>US 2003 126624 A 1</td>
<td>03. 07. 2003</td>
</tr>
<tr>
<td></td>
<td></td>
<td>JP 3817739 B2</td>
<td>06. 09. 2006</td>
</tr>
<tr>
<td></td>
<td></td>
<td>US 20071 50973 A 1</td>
<td>28. 06. 2007</td>
</tr>
<tr>
<td></td>
<td></td>
<td>US 7485441 B2</td>
<td>03. 02. 2009</td>
</tr>
<tr>
<td></td>
<td></td>
<td>US 7763446 B2</td>
<td>27. 07. 2010</td>
</tr>
</tbody>
</table>
INTERNATIONAL SEARCH REPORT

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