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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
12 December 2013 (12.12.2013)

WIPO | PCT

(10) International Publication Number
WO 2013/182910 A2

(51) International Patent Classification: Not classified

(21) International Application Number:
PCT/IB2013/001741

(22) International Filing Date:
5 June 2013 (05.06.2013)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/655,747 5 June 2012 (05.06.2012) US

(71) Applicant: **CELLECTIS** [FR/FR]; 8 Rue De La Croiz Jarry, F-75013 Paris (FR).

(72) Inventors; and

(71) Applicants : **JUILLERAT, Alexandre** [CH/FR]; 38, Rue De L'Amiral Mouchez, F-75014 Paris (FR). **DUCHATEAU, Philippe** [FR/FR]; Bateau Fawen, Quai aux Dames, F-91210 Draveil (FR).

(74) Agent: **CABINET ORES**; 36, rue de St Pétersbourg, F-75008 Paris (FR).

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

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

Published:

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

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

(54) Title: NEW TRANSCRIPTION ACTIVATOR-LIKE EFFECTOR (TALE) FUSION PROTEIN

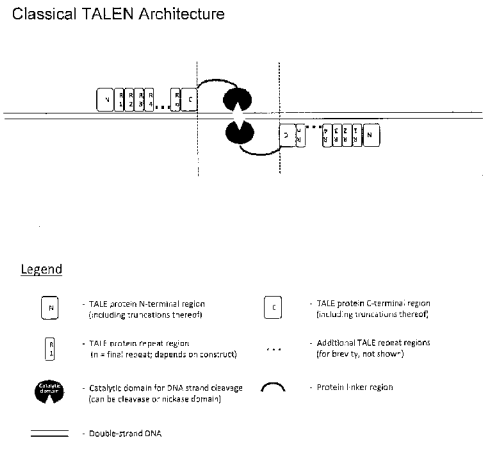


Figure 1

(57) Abstract: The present invention relates to Transcription Activator-Like Effector (TALE) derived proteins that allow to efficiently target and/or process double stranded nucleic acid sequences. The proteins of the invention are typically chimeric protein monomers composed of a core scaffold comprising Repeat Variable Dipeptide regions (RVDs) having binding specificity to a DNA target sequence, to which is fused a catalytic domain to its N-terminal. This later catalytic domain, which can be a monomer of a nuclease, is placed at this position to possibly interact with another catalytic domain fused to another TAL monomer, such that, when said monomers are binding to their respective target DNA sequences, both catalytic domains form a catalytic entity likely to process DNA in the proximity of these target sequences. This new TAL architecture makes it possible to target only one DNA strand, which is not the case, for instance, with classical TALEN architectures. The present invention also relates to vectors encoding such proteins and compositions or kits in which Transcription Activator-Like Effector (TALE) proteins of the present invention are used.

WO 2013/182910 A2

New transcription Activator-Like Effector (TALE) fusion proteins

Field of the invention

The present invention relates to Transcription Activator-Like Effector (TALE) derived proteins that allow to efficiently target and/or process double stranded nucleic acid sequences. The proteins of the invention are typically chimeric protein monomers composed of a core scaffold comprising Repeat Variable Dipeptide regions (RVDs) having binding specificity to a DNA target sequence, to which is fused a catalytic domain to its N-terminal. This later catalytic domain, which can be a monomer of a nuclease, is placed at this position to possibly interact with another catalytic domain fused to another TAL monomer, such that, when said monomers are binding to their respective target DNA sequences, both catalytic domains form a catalytic entity likely to process DNA in the proximity of these target sequences. By contrast to classical TALEN architectures, this new TAL architecture makes it possible to target only one DNA strand. The present invention also relates to vectors encoding such proteins and compositions or kits in which Transcription Activator-Like Effector (TALE) proteins of the present invention are used.

Background of the invention

TAL effector DNA binding domains have been derived from of a recently discovered new class of proteins AvrBs3 originating from the plant pathogen *Xanthomonas*, which act as Transcription Activators in plant cells during the process of infection (Kay et al. 2007). These AvrBs3 proteins have been found to activate the transcription of some specific genes involved into the infection process by binding to specific promoter sequences. Their binding domains are composed of an array of motifs of 33-35 amino acids repeats. These repeats differ essentially by the residues 12 and 13 (di-residues), named RVDs (repeat variable diresidues), which constitute the TAL effector DNA binding domain of these proteins. The study of the RVDs in relation with the natural promoter DNA sequences targeted by these AvrBs3 proteins has revealed in 2009 that there was a specific correlation between the RVDs found within the TAL effector DNA binding domain and the nucleic acids present in the promoter sequences. As a result, a code has been established between amino acids and DNA sequences, so that it is now possible, by following said code, to engineer TAL effector DNA binding domains by assembly of selected RVDs to target specific DNA sequences (Boch, Scholze et al. 2009; Moscou and Bogdanove 2009).

Engineered TAL effector DNA binding domains have since been used with success, especially as a tool in genome engineering by fusion of a catalytic domain with nuclease activity to the C-terminal of AvrBs3-like proteins. Such fusions form endonucleases with sequence specificity called TALE nuclease or TALEN (WO 2011/072246). These TALEN have the ability to create double strand breaks at proximity of selected DNA target sequences. TALEN have been shown to be active to various extents in cell-based assays in yeast, mammalian cells and plants (Christian, Cermak et al. 2010; Li, Huang et al. 2010; Cermak, Doyle et al. 2011; Geissler, Scholze et al. 2011; Huang, Xiao et al. 2011; Li, Huang et al. 2011; Mahfouz, Li et al. 2011; Miller, Tan et al. 2011; Morbitzer, Elsaesser et al. 2011; Mussolino, Morbitzer et al. 2011; Sander, Cade et al. 2011; Tesson, Usal et al. 2011; Weber, Gruetzner et al. 2011; Zhang, Cong et al. 2011; Deng, Yan et al. 2012; Li, Piatek et al. 2012; Mahfouz, Li et al. 2012).

Up to now, researchers have classically used TALEN architecture based on the fusion of a FokI catalytic head (CH) to C-terminal truncated forms of the wild type protein AvrBs3. Despite truncation of the N-terminal domain has also been proven to be functional to a certain extend (more precisely at position 152), the N-terminal domain of TAL effector DNA binding domains have not been used for fusion with active proteins (N-terminal part of AvrBs3 has not been reported to display any real functionality).

Meanwhile, current C-terminal TALEN fusions using FokI as a catalytic head (classical TALEN architecture) suffers major drawbacks limiting the fields of its possible applications. Indeed, the FokI catalytic head requires TALEN dimerization to be active, which requires two TAL monomers facing each other on the two opposite DNA strands to recompose an active molecule (Christian, Cermak et al., 2010). In addition, the N-terminus toward C-terminus orientation of a TAL has to follow the 5' toward 3' orientation of one DNA strand to allow binding ((Deng, Yan et al. 2012; Mak, Bradley et al. 2012)). More importantly, the targeted sequences have to start with a thymine base (T) for an effective binding by the first RVDs of the protein located at the N-terminal domain of the TAL (Boch, Scholze et al. 2009; Moscou and Bogdanove 2009). This "T" requirement significantly reduces the possibilities of targeting any nucleotide sequence into the genome.

Given the above requirements for targeting DNA sequences into the genome, the inventors have developed new TALEN scaffolds, which unexpectedly allow targeting a broader spectrum of nucleic acid sequences more easily to target than using classical TALEN architecture.

Brief summary of the invention

In a general aspect, the present invention relates to Transcription Activator-Like Effector (TALE) monomers, to which are fused catalytic domains, especially into their N-terminals. These catalytic domains, when associated with a catalytic domain of another TAL monomer, form a chemical or enzyme entity that allows the processing of DNA. This processing may be to modify the structure of the DNA molecule by cleavage, base replacement or chemical reaction. It may be also to regulate gene expression.

As a particular embodiment, the TAL-derived monomers have two catalytic domains respectively fused to their N and C-terminals, so that multiple monomers can be associated together all along a DNA sequence in order to create different catalytic entities intervening at different DNA locations, The catalytic entities become active preferably concomitantly, upon multimer formation upon binding to their respective DNA targets (Figures 4 and 5).

By contrast to the classical TALEN architecture, the polypeptides according to the invention, allow to efficiently target sequences that are all located on either DNA strand. In particular, they make possible to target only one DNA strand.

Also, the TAL-derived monomers according to the invention allow to target nucleic acid sequences that are not targetable with the classical TALEN architecture, especially double strand DNA sequences comprising fewer T bases or none on one strand, such as those composed of highly repetitive sequences of G, A and C triplets (e.g. expansion triplets).

The present invention also concerns different methods in which these polypeptides can be advantageously used, as well as nucleic acids and vectors encoding such polypeptides.

Brief description of the figures and tables

In addition to the preceding features, the invention further comprises other features which will emerge from the description which follows, as well as to the appended drawings. A more complete appreciation of the invention and many of the attendant advantages thereof will be readily obtained as the same becomes better understood by reference to the following Figures in conjunction with the detailed description below.

Figure 1: Schematic representation of the classical TALEN architecture as described in the international application WO2011/072246.

Figure 2: Schematic representation of a new TALEN architecture according to the invention as described herein

Figure 3: Schematic of Head (N-terminal)/Tail (C-terminal) protein configuration according to the invention.

Figure 4: Schematic representation of different configurations made possible by associating several Tal derived monomers according to the invention. A – Tail/Tail orientation: The TAL monomers interact through their C-terminal fused catalytic domains, whereas there are binding to sequences located on each DNA strands. **B – Head/Head orientation**: The TAL monomers interact through their N-terminal fused catalytic domains. **C – Tail/Head orientation**: The TAL monomers interact through their N and C-terminal fused catalytic domains, while they are binding to target sequences that are located on the same DNA strand. **D – Tail/Head multimerization**: Several TAL monomers interact through their N and C-terminal fused catalytic domains, creating multiple catalytic entities, all the monomers being located on the same DNA strand. **E** – Same situation as in D, an example of DNA sequence is provided showing that the other DNA strand does not provide a T base, and therefore could not be targeted by using classical TALEN. T represents the base that must be present in the DNA target sequence for it to be properly recognized by the first RVD of the TAL binding domain.

Figure 5: Double head architecture and example of TAL multimerization using the monomers according to the invention.

Table 1: Cleavage activity of Tail/Tail configuration (see example 1) on pseudo-palindromic sequences targets (two identical recognition sequences are placed facing each other on both DNA strands separated by a spacer ranging from 5 to 40 bps in our yeast SSA assay previously described (International PCT Applications WO 2004/067736 and in (Epinat, Arnould et al. 2003; Chames, Epinat et al. 2005; Arnould, Chames et al. 2006; Smith, Grizot et al. 2006) at 37°C. +/- represent a barely detectable activity, + a low activity, ++ a medium activity and +++ a high activity. n.d. indicates that no activity was detected.

Table 2: Cleavage activity of Head/Head configuration (see example 1) on pseudo-palindromic sequences targets (two identical recognition sequences are placed facing each other on both DNA strands separated by a spacer ranging from 5 to 35 bps in our yeast SSA assay previously described (International PCT Applications WO 2004/067736 and in (Epinat, Arnould et al. 2003; Chames, Epinat et al. 2005; Arnould, Chames et al. 2006; Smith, Grizot et al. 2006) at 37°C. +/-

represent a barely detectable activity, + a low activity, ++ a medium activity and +++ a high activity. n.d. indicates that no activity was detected.

Table 3: Cleavage activity of Head/Tail configuration (see example 2) on either targets containing two identical GCT repeated sequences, on the same strand, separated by a spacer of 12, 15, 18, 21, 24, 27 or 30 bps or a target composed exclusively of GCT repetitions in our yeast SSA assay previously described (International PCT Applications WO 2004/067736 and in (Epinat, Arnould et al. 2003; Chames, Epinat et al. 2005; Arnould, Chames et al. 2006; Smith, Grizot et al. 2006) at 37°C. + represent a low activity, ++ a medium activity and +++ a high activity.

Detailed description of the invention

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

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

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

Mammalian Cells (J. H. Miller and M. P. Calos eds., 1987, Cold Spring Harbor Laboratory); Immunochemical Methods In Cell And Molecular Biology (Mayer and Walker, eds., Academic Press, London, 1987); Handbook Of Experimental Immunology, Volumes I-IV (D. M. Weir and C. C. Blackwell, eds., 1986); and Manipulating the Mouse Embryo, (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y., 1986).

In a general aspect, the present invention relates to polypeptides that allow to efficiently target and/or process nucleic acids.

In a particular aspect, the present invention relates to polypeptides derived from Transcription Activator-Like Effector derived proteins, under the form of chimeric proteins, having an improved design or architecture, which allows to efficiently target and/or process DNA sequences that are not targetable with the classical TALEN architecture.

According to a first embodiment, the invention provides monomers derived from a Transcription Activator-Like Effector (TALE) comprising a core scaffold comprising Repeat Variable Dipeptide regions (RVDs) having binding specificity to a DNA target sequence and a catalytic domain fused to the N-terminal of said core scaffold. This fusion is generally operated using an appropriate flexible peptide linker. Said catalytic domain is preferably chosen in relation to another catalytic domain fused to another TAL-derived monomer, called counterpart domain.

According to one aspect of the invention, this counterpart domain preferably interacts with the first catalytic domain, so as to form a catalytic entity. Said catalytic domain or said catalytic entity has the ability to process DNA, which means that it can have an activity on the physical structure of the DNA, like for instance a nuclease activity, or an effect on gene expression, like for instance by inhibiting or enhancing transcription.

As a preferred embodiment, nuclease activity is provided by said catalytic domain. As an example an enzyme having type IIS nuclease activity, such as Fok1 (SEQ ID NO.81) can be fused to the N-terminal of the TAL derived scaffold. Enzymes, like Fok1 have the advantage of being active under dimer form, so that a first Fok1 monomer can be fused to the first TAL scaffold, and a second Fok1 monomer to a second TAL scaffold, in order to form a Fok1 homodimer as an active catalytic entity having nuclease activity. Such a catalytic domain comprises an amino acid sequence having at least 80%, preferably at least 90%, more preferably at least 95% identity with SEQ ID NO: 81.

As another example, monomers of endonucleases can be fused to TAL scaffolds, especially from meganucleases, to form a catalytic entity having nuclease activity. Thereby, the catalytic domain

can be a monomer that will form a heterodimer or homodimer catalytic entity having nuclease activity, when put into contact with its counterpart.

A protein monomer according to the invention can also have a catalytic entity emitting a recordable reporter signal.

In agreement with what precedes, the invention is more particularly drawn to a monomer comprising:

- A core scaffold comprising Repeat Variable Dipeptide regions (RVDs) having binding specificity to a DNA target sequence; and
- At least two catalytic domains respectively fused to the N-terminal and C-terminals of said core scaffold, wherein at least one of said catalytic domain, when in contact with a counterpart catalytic domain, form a catalytic entity to process DNA, upstream or downstream of said target sequence.

According to such embodiment, the N-terminal and C-terminals are both fused with catalytic domains, which does not mean that both catalytic domains are identical or display the same activities.

According to another aspect of the invention, the counterpart domain does not directly interact with the catalytic domain, but contributes to an additional or a further step of the DNA processing. In such case, the activities of both catalytic domains can be simultaneous or take place at different times. Nevertheless, as per the invention, those different activities are drawn to the same DNA region preferably located between the target sequence of each TAL-derived monomers. For instance, the first catalytic domain can have an endonuclease activity on this region and the second catalytic domain can have a nikase, so that NHEJ activity on said DNA region is increased.

Since the catalytic domain generally derives from an enzyme, the polypeptides according to the invention can be a protein encoded by a nucleic acid or a vector encoding such a polypeptide, which are also part of the invention.

Compared to classical TALEN architecture based on AvrBs3, the polypeptides of the invention, which are preferably monomers, display a new structure by fusion of said catalytic domain to the N-terminal of a TAL derived scaffold. Such monomers according to the invention have the ability to form multimers which can bind to any DNA strand of a double stranded DNA. In particular, it can bind to only one strand of said double stranded DNA, which was not possible to the classical TALEN architecture. This possibility had not been previously foreseen because classical TALEN

architecture has derived from the protein AvrBs3, which does not naturally comprise a catalytic domain in its N-terminal part. Also, in classical TALEN architecture, the nucleases have been so far fused to the C-terminal part of AvrBs3 in order to help or at least preserve dimerization of the whole protein.

Interestingly, it has been found by the present inventors that fusion to the N-terminal part of the TAL derived Scaffold could produce monomers that do not require dimerization of the core scaffold to be functional. Accordingly, when the catalytic domains are fused to the N-terminal part of the monomers, the core scaffolds may not need to assemble for obtaining an activity. Only the catalytic domains fused to the TAL derived scaffolds, should these catalytic domains have to dimerize to form an active catalytic entity (such as of Fok1), then need to dimerize. As a result, the TAL core scaffolds apparently undergo less steric constraints when they bind to target sequences, and thus, said target sequences can be defined on either DNA strand, which was not possible with classical TALEN architecture. According to a preferred method, the monomers bind to the same DNA strand.

Accordingly, the present invention more particularly concerns methods to process a double stranded DNA sequence into a cell by expressing the monomers previously described, and also methods combining such monomers to more efficiently target and process such DNA sequence. Several combinations of the monomers are illustrated in Figure 4 to target and/or process double stranded DNA. Said methods typically comprise transfecting a cell with one or more polynucleotide(s) expressing one or several monomer(s) as previously defined.

As a preferred embodiment of the invention, the monomers are fused to catalytic domains inducing double strand break (DSB) into the DNA between the sequences targeted by the RVDs of said two monomers. By this way, the method of the invention can be used for modifying genomic DNA, and can comprise additional steps of genetic engineering, in particular steps using homologous recombination techniques.

This method is particularly suited when one of the two DNA strand sequences is devoid of T or comprises fewer T, making it difficult to target with classical TALEN.

Such a method generally comprises at least one of the following steps:

- Identifying said double stranded DNA sequence in a cell ;
- Transfecting said cell with one or two nucleic acid encoding at least two protein monomers each comprising:

- A core scaffold comprising Repeat Variable Dipeptide regions (RVDs) having binding specificity to a DNA target sequence on one strand of said double stranded DNA; and
- A catalytic domain; the catalytic domain of the first monomer being fused to the N-terminal of its core scaffold, and the catalytic domain of the second monomer being fused to the C-terminal of its core scaffold, said two catalytic domains, when they are in contact, forming a catalytic entity being able to process DNA between the respective target sequences of said two monomers.

- expressing said protein monomers into said cell;

such that said monomers bind their respective target sequences and process DNA between said target sequences through the combination of their catalytic domains.

Preferably, the Repeat Variable Dipeptide regions (RVDs) of both monomers are chosen to bind target sequences, which are located on the same DNA strand, preferably the DNA strand that comprises more T bases than the other. It may happen, for instance, that a nucleic acid sequence can be devoid of T or show T only at inappropriate positions, making it more difficult to target using classical TALEN architecture. This may be the case in particular when it comprises multiple repeats or trinucleotide repeats expansion. A DNA sequence may also contain a low number of T bases in a genomic region to be processed, which then reduces the number of manageable targets. The method according to the invention aims to resolve such situations, by allowing to reaching more putative targets into a genome.

Accordingly, the method of the invention may include the steps of providing a cell containing a DNA sequence showing a low occurrence of T, generally less than 20 %, more generally less than 10 %, more generally less than 5 %, and even less than 1 or 2 %, in view of its processing as previously described.

Also, the method of the invention allow to process a double stranded DNA sequence comprising putative trinucleotide repeats expansion devoid of T; or displaying a lower occurrence of T on one of its strand, said method comprising at least one of the following steps:

- Providing a cell which may contain a double stranded DNA sequence comprising such putative trinucleotide repeats expansion devoid of T on one of its strand;
- Determining a threshold number of trinucleotide repeats under which no processing is desired;
- Transfecting said cell with one or two nucleic acid encoding at least two protein monomers each comprising:

- A core scaffold comprising Repeat Variable Dipeptide regions (RVDs) having binding specificity to said number of trinucleotide repeats; and
- A catalytic domain; the catalytic domain of the first monomer being fused to the N-terminal of its core scaffold, and the catalytic domain of the second monomer being fused to the C-terminal of its core scaffold, said two catalytic domains, when they are in contact, forming a catalytic entity being able to process DNA between the respective target sequences of said two monomers.

- Inducing expression of said chimeric protein monomer encoded by said nucleic acid;

such that said protein monomers bind and process DNA between said trinucleotide repeats DNA target sequences when said repeats are equal or above said threshold number.

The activity of the above new TALEN monomers on multiple repeats is also tunable by modifying the number of RVDs (e.g. sequences containing a number of repetitions under a threshold value are protected from cleavage using TALEN containing long arrays of RVDs and in contrary short arrays of RVDs allows targeting smaller number of repetitions).

According to another aspect of the invention, the previously described monomers may be used in a method to measure the number of putative repeated sequences within a double stranded genome DNA sequence, in particular repeated sequences devoid of T on one strand of said genome sequence, said method comprising at least one of the following steps :

- providing said DNA genome in a cell
- Transfecting said cell with one or two nucleic acids encoding at least two protein monomers each comprising:
  - A core scaffold comprising Repeat Variable Dipeptide regions (RVDs) having binding specificity to said repeated DNA sequence (target sequence); and
  - A catalytic domain; the catalytic domain of the first monomer being fused to the N-terminal of its core scaffold, and the catalytic domain of the second monomer being fused to the C-terminal of its core scaffold, said two catalytic domains, when they are in contact, forming a catalytic entity being able to produce a reporter signal.
- Inducing expression of said protein monomers encoded by said nucleic acid(s);
- Recording said reporter signal intensity;
- Deducing from said reporter signal intensity the number of putative repeated sequences within said DNA sequence into the genome of said cell;
- Optionally, comparing this reporter signal intensity with the reporter signal intensity recovered from a control cell.

The methods of the invention may be applied to any types of cells, especially eukaryotic cells including mammalian and plant cells.

The DNA to be processed is preferably chromosomal DNA. The invention could also be applied to other nucleic acids, such as RNA.

Considering the above methods, especially the multimerization process into which several monomers of the invention are used in combination, one aspect of the invention concerns a set or a kit of at least two protein monomers as previously defined. It also concerns kit of polynucleotides or vectors encoding such monomers.

In these sets or kits, said first and second monomers preferably comprise:

- A core scaffold comprising Repeat Variable Dipeptide regions (RVDs) having binding specificity to a DNA target sequence; and
- A catalytic domain; the catalytic domain of the first monomer being fused to the N-terminal of its core scaffold, and the catalytic domain of the second monomer being fused to the C-terminal of its core scaffold, said two catalytic domains, when they are in contact, forming a catalytic entity being able to process DNA between the respective target sequences of said two monomers.

As previously mentioned, the DNA target sequences of said two monomers can be advantageously located on the same DNA strand of a given double stranded DNA.

Said set of kit can comprise additional monomers, especially a third monomer in order to bridge at least two, preferably at least three, target sequences, the number of monomers being in theory not limited. Numerous monomers may be used to ensure, for instance, the deletion or the silencing of a whole region in a genome. The present monomers may also be used to process methylated sequences, which are often more difficult to process. In particular, this may be used to increase the probability of DSB event in a region where the chromosomal DNA is difficult to access.

As an alternative, several monomers may be assembled into a single chain polypeptide, thereby being expressed by a single polynucleotide.

Other definitions:

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

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

- DNA or nucleic acid processing activity refers to a particular or given enzymatic activity conferred by a catalytic domain onto the nucleic acid structure or onto the expression of genes, directly on indirectly. Said DNA or nucleic acid processing activity can refer to a cleavage activity, either a cleavase activity either a nickase activity, more broadly a nuclease activity but also a polymerase activity, a transcriptional activity, a kinase activity, a phosphatase activity, a methylase activity, a topoisomerase activity, an integrase activity, a transposase activity, a ligase, a helicase or recombinase activity as non-limiting examples.

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

- by "peptide linker" or "peptidic linker" it is intended to mean a peptide sequence which allows the connection of different monomers or different parts comprised in a fusion protein such as between a TALE DNA binding domain and a protein domain in a chimeric protein or a polypeptide according to the present invention and which allows the adoption of a correct conformation for said chimeric protein activity and/or specificity. Peptide linkers can be of various sizes, from 3 amino acids to 50 amino acids as a non limiting indicative range. Peptide linkers can also be qualified as structured or unstructured. Peptide linkers can be qualified as active linkers when they comprise active domains that are able to change their structural conformation under appropriate stimulation.

- by "subdomain" it is intended a protein subdomain or a protein part that interacts with another protein subdomain or protein part to form an active entity and / or a catalytic active entity bearing nucleic acid or DNA processing activity of said chimeric protein or polypeptide according to the invention.

- by "exogenous sequence" it is intended to mean a DNA construct comprising a first and second portion that are homologous to regions 5' and 3' of a DNA target in situ. The DNA construct also comprises a third portion positioned between the first and second portion which comprise some homology with the corresponding DNA sequence in situ or alternatively comprise no homology with the regions 5' and 3' of the DNA target in situ. Following cleavage of the DNA target, a homologous recombination event is stimulated between the genome containing the targeted gene comprised in the locus of interest and the repair matrix, wherein the genomic sequence containing the DNA target is replaced by the third portion of the repair matrix and a variable part of the first and second portions of the repair matrix.

- by "DNA target", "DNA target sequence", "target DNA sequence", "nucleic acid target sequence", "target sequence" , is intended a polynucleotide sequence which can be bound by the TALE DNA binding domain that is included in the proteins of the present invention. It refers to a specific DNA location, preferably a genomic location in a cell, but also a portion of genetic material that can exist independently to the main body of genetic material such as plasmids, episomes, virus, transposons or in organelles such as mitochondria or chloroplasts as non-limiting examples. The nucleic acid target sequence is defined by the 5' to 3' sequence of one strand of said target, as indicated for SEQ ID NO: 83 to 89 in table 3 as a non-limiting example. Generally, the DNA target is adjacent or in the proximity of the locus to be processed either upstream (5' location) or downstream (3' location). In a preferred embodiment, the target sequences and the proteins are designed in order to have said locus to be processed located between two such target sequences. Depending on the catalytic domains of the proteins, the target sequences may be distant from 5 to 50 bases (bp), preferably from 10 to 40 bp, more preferably from 15 to 30, even more preferably from 15 to 25 bp. These later distances define the spacer referred to in the description and the examples. It can also define the distance between the target sequence and the nucleic acid sequence being processed by the catalytic domain on the same molecule.

- By " delivery vector" or " delivery vectors" is intended any delivery vector which can be used in the present invention to put into cell contact (i.e "contacting") or deliver inside cells or subcellular compartments agents/chemicals and molecules (proteins or nucleic acids) needed in the present invention. It includes, but is not limited to liposomal delivery vectors, viral delivery vectors, drug delivery vectors, chemical carriers, polymeric carriers, lipoplexes, polyplexes, dendrimers, microbubbles (ultrasound contrast agents), nanoparticles, emulsions or other appropriate transfer vectors. These delivery vectors allow delivery of molecules, chemicals, macromolecules (genes, proteins), or other vectors such as plasmids, peptides developed by Diatos. In these cases, delivery vectors are molecule carriers. By "delivery vector" or "delivery vectors" is also intended delivery methods to perform transfection.

The terms "vector" or "vectors" refer to a nucleic acid molecule capable of transporting another nucleic acid to which it has been linked. A "vector" in the present invention includes, but is not limited to, a viral vector, a plasmid, a RNA vector or a linear or circular DNA or RNA molecule which may consists of a chromosomal, non chromosomal, semi-synthetic or synthetic nucleic acids. Preferred vectors are those capable of autonomous replication (episomal vector) and/or expression of nucleic acids to which they are linked (expression vectors). Large numbers of suitable vectors are known to those of skill in the art and commercially available. One type of preferred vector is an episome, i.e., a nucleic acid capable of extra-chromosomal replication. Preferred vectors are those capable of autonomous replication and/or expression of nucleic acids to which they are linked. Vectors capable of directing the expression of genes to which they are operatively linked are referred to herein as "expression vectors. A vector according to the present invention comprises, but is not limited to, a YAC (yeast artificial chromosome), a BAC (bacterial artificial), a baculovirus vector, a phage, a phagemid, a cosmid, a viral vector, a plasmid, a RNA vector or a linear or circular DNA or RNA molecule which may consist of chromosomal, non chromosomal, semi-synthetic or synthetic DNA. In general, expression vectors of utility in recombinant DNA techniques are often in the form of "plasmids" which refer generally to circular double stranded DNA loops which, in their vector form are not bound to the chromosome. Large numbers of suitable vectors are known to those of skill in the art. Vectors can comprise selectable markers, for example: neomycin phosphotransferase, histidinol dehydrogenase, dihydrofolate reductase, hygromycin phosphotransferase, herpes simplex virus thymidine kinase, adenosine deaminase, glutamine synthetase, and hypoxanthine-guanine phosphoribosyl transferase for eukaryotic cell culture; TRP1 for *S. cerevisiae*; tetracyclin, rifampicin or ampicillin resistance in *E. coli*. Preferably said vectors are expression vectors, wherein a sequence encoding a polypeptide of interest is placed under control of appropriate transcriptional and translational control elements to permit production or synthesis of said polypeptide. Therefore, said polynucleotide is comprised in an expression cassette. More particularly, the vector comprises a replication origin, a promoter operatively linked to said encoding polynucleotide, a ribosome binding site, a RNA-splicing site (when genomic DNA is used), a polyadenylation site and a transcription termination site. It also can comprise an enhancer or silencer elements. Selection of the promoter will depend upon the cell in which the polypeptide is expressed. Suitable promoters include tissue specific and/or inducible promoters. Examples of inducible promoters are: eukaryotic metallothionine promoter which is induced by increased levels of heavy metals, prokaryotic lacZ promoter which is induced in response to isopropyl-β-D-thiogalacto-pyranoside (IPTG) and eukaryotic heat shock promoter which is induced by increased temperature. Examples of tissue specific promoters are skeletal muscle creatine kinase, prostate-specific antigen (PSA), α-antitrypsin protease, human

surfactant (SP) A and B proteins, β-casein and acidic whey protein genes. Delivery vectors and vectors can be associated or combined with any cellular permeabilization techniques such as sonoporation or electroporation or derivatives of these techniques.

- Viral vectors include retrovirus, adenovirus, parvovirus (e. g. adenoassociated viruses), coronavirus, negative strand RNA viruses such as orthomyxovirus (e. g., influenza virus), rhabdovirus (e. g., rabies and vesicular stomatitis virus), paramyxovirus (e. g. measles and Sendai), positive strand RNA viruses such as picornavirus and alphavirus, and double-stranded DNA viruses including adenovirus, herpesvirus (e. g., Herpes Simplex virus types 1 and 2, Epstein-Barr virus, cytomegalovirus), and poxvirus (e. g., vaccinia, fowlpox and canarypox). Other viruses include Norwalk virus, togavirus, flavivirus, reoviruses, papovavirus, hepadnavirus, and hepatitis virus, for example. Examples of retroviruses include: avian leukosis-sarcoma, mammalian C-type, B-type viruses, D type viruses, HTLV-BLV group, lentivirus, spumavirus (Coffin, J. M., Retroviridae: The viruses and their replication, In Fundamental Virology, Third Edition, B. N. Fields, et al., Eds., Lippincott-Raven Publishers, Philadelphia, 1996).

- By "lentiviral vector" is meant HIV-Based lentiviral vectors that are very promising for gene delivery because of their relatively large packaging capacity, reduced immunogenicity and their ability to stably transduce with high efficiency a large range of different cell types. Lentiviral vectors are usually generated following transient transfection of three (packaging, envelope and transfer) or more plasmids into producer cells. Like HIV, lentiviral vectors enter the target cell through the interaction of viral surface glycoproteins with receptors on the cell surface. On entry, the viral RNA undergoes reverse transcription, which is mediated by the viral reverse transcriptase complex. The product of reverse transcription is a double-stranded linear viral DNA, which is the substrate for viral integration in the DNA of infected cells.

- By "integrative lentiviral vectors (or LV)", is meant such vectors as non limiting example, that are able to integrate the genome of a target cell.

- At the opposite by "non integrative lentiviral vectors (or NILV)" is meant efficient gene delivery vectors that do not integrate the genome of a target cell through the action of the virus integrase.

- Inducible promoters may be induced by pathogens or stress, more preferably by stress like cold, heat, UV light, or high ionic concentrations (reviewed in Potenza C et al. 2004, In vitro Cell Dev Biol 40:1-22). Inducible promoter may be induced by chemicals (reviewed in (Moore, Samalova et al. 2006); (Padidam 2003); (Wang, Zhou et al. 2003); (Zuo and Chua 2000).

- By cell or cells is intended any prokaryotic or eukaryotic living cells, cell lines derived from these organisms for *in vitro* cultures, primary cells from animal or plant origin.

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

- In the frame of the present invention, "eukaryotic cells" refer to a fungal, plant or animal cell or a cell line derived from the organisms listed below and established for in vitro culture. More preferably, the fungus is of the genus *Aspergillus, Penicillium, Acremonium, Trichoderma, Chrysoporium, Mortierella, Kluyveromyces* or *Pichia*; More preferably, the fungus is of the species *Aspergillus niger, Aspergillus nidulans, Aspergillus oryzae, Aspergillus terreus, Penicillium chrysogenum, Penicillium citrinum, Acremonium Chrysogenum, Trichoderma reesei, Mortierella alpine, Chrysosporium lucknowense, Kluyveromyces lactis, Pichia pastoris* or *Pichia ciferrii*. More preferably the plant is of the genus *Arabidospis, Nicotiana, Solanum, lactuca, Brassica, Oryza, Asparagus, Pisum, Medicago, Zea, Hordeum, Secale, Triticum, Capsicum, Cucumis, Cucurbita, Citrullis, Citrus, Sorghum*; More preferably, the plant is of the species *Arabidospis thaliana, Nicotiana tabaccum, Solanum lycopersicum, Solanum tuberosum, Solanum melongena, Solanum esculentum, Lactuca saliva, Brassica napus, Brassica oleracea, Brassica rapa, Oryza glaberrima, Oryza sativa, Asparagus officinalis, Pisum sativum, Medicago sativa, zea mays, Hordeum vulgare, Secale cereal, Triticum aestivum, Triticum durum, Capsicum sativus, Cucurbita pepo, Citrullus lanatus, Cucumis melo, Citrus aurantifolia, Citrus maxima, Citrus medica, Citrus reticulata*.

More preferably the animal cell is of the genus Homo, *Rattus, Mus, Sus, Bos, Danio, Canis, Felis, Equus, Salmo, Oncorhynchus, Gallus, Meleagris, Drosophila, Caenorhabditis*; more preferably, the animal cell is of the species *Homo sapiens, Rattus norvegicus, Mus musculus, Sus scrofa, Bos taurus, Danio rerio, Canis lupus, Felis catus, Equus caballus, Salmo salar, Oncorhynchus mykiss, Gallus gallus, Meleagris gallopavo, Drosophila melanogaster, Caenorhabditis elegans*.

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

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

- by "mutation" is intended the substitution, deletion, insertion of one or more nucleotides/amino acids in a polynucleotide (cDNA, gene) or a polypeptide sequence. Said mutation can affect the coding sequence of a gene or its regulatory sequence. It may also affect the structure of the genomic sequence or the structure/stability of the encoded mRNA.

- In the frame of the present invention, the expression "double-strand break-induced mutagenesis" (DSB-induced mutagenesis) refers to a mutagenesis event consecutive to an NHEJ event following an endonuclease-induced DSB, leading to insertion/deletion at the cleavage site of an endonuclease.

- By "gene" is meant the basic unit of heredity, consisting of a segment of DNA arranged in a linear manner along a chromosome, which codes for a specific protein or segment of protein. A gene typically includes a promoter, a 5' untranslated region, one or more coding sequences (exons), optionally introns, a 3' untranslated region. The gene may further comprise a terminator, enhancers and/or silencers.

- As used herein, the term "locus" is the specific physical location of a DNA sequence (e.g. of a gene) on a chromosome. The term "locus" usually refers to the specific physical location of a polypeptide or chimeric protein's nucleic target sequence on a chromosome. Such a locus can comprise a target sequence that is recognized and/or cleaved by a polypeptide or a chimeric protein according to the invention. It is understood that the locus of interest of the present invention can not only qualify a nucleic acid sequence that exists in the main body of genetic material (i.e. in a chromosome) of a cell but also a portion of genetic material that can exist independently to said main body of genetic material such as plasmids, episomes, virus, transposons or in organelles such as mitochondria or chloroplasts as non-limiting examples.

- By "fusion protein" is intended the result of a well-known process in the art consisting in the joining of two or more genes which originally encode for separate proteins or part of them, the translation of said "fusion gene" resulting in a single polypeptide with functional properties derived from each of the original proteins.

- By "chimeric protein" according to the present invention is meant any fusion protein comprising at least one RVD to bind a nucleic acid sequence and one protein domain to process a nucleic acid target sequence within or adjacent to said bound nucleic acid sequence.

- By "protein domain" is meant the nucleic acid target sequence processing part of said chimeric protein according to the present invention. Said protein domain can provide any

catalytical activity (catalytic domain) as classified and named according to the reaction they catalyze [Enzyme Commission number (EC number) at http://www.chem.qmul.ac.uk/iubmb/enzyme/)]. Said protein domain can be a catalytically active entity by itself. Said protein domain can be a protein subdomain that needs to interact with another protein subdomain to form a dimeric protein domain active entity.

- By a "TALE-nuclease" (TALEN) is intended a fusion protein consisting of a DNA-binding domain derived from a Transcription Activator Like Effector (TALE) and one nuclease catalytic domain to cleave a nucleic acid target sequence. Said TALEN is a subclass of chimeric protein according to the present invention.

- By "variant(s)", it is intended a RVD variant, a chimeric protein variant, a DNA binding variant, a TALEN variant, a polypeptide variant obtained by replacement of at least one residue in the amino acid sequence of the parent molecule.

- By "functional mutant" is intended a catalytically active mutant of a protein or a protein domain; such mutant can have the same activity compared to its parent protein or protein domain or additional properties. This definition applies to chimeric proteins or protein domains that constitute chimeric proteins according to the present invention. Are also encompassed in the scope of this definition "derivatives" of these proteins or protein domains that comprise the entirety or part of these proteins or protein domains fused to other proteic or chemical parts such as tags, antibodies, polyethylene glycol as non-limiting examples.

- "identity" refers to sequence identity between two nucleic acid molecules or polypeptides. Identity can be determined by comparing a position in each sequence which may be aligned for purposes of comparison. When a position in the compared sequence is occupied by the same base, then the molecules are identical at that position. A degree of similarity or identity between nucleic acid or amino acid sequences is a function of the number of identical or matching nucleotides at positions shared by the nucleic acid sequences. Various alignment algorithms and/or programs may be used to calculate the identity between two sequences, including FASTA, or BLAST which are available as a part of the GCG sequence analysis package (University of Wisconsin, Madison, Wis.), and can be used with, e.g., default setting. Unless otherwise stated, the present invention encompasses polypeptides and polynucleotides sharing at least 70 %, generally at least 80 %, more generally at least 85 %, preferably at least 90 %, more preferably at least 95 % and even more preferably at least 97 % with those described herein.

The above written description of the invention provides a manner and process of making and using it such that any person skilled in this art is enabled to make and use the same, this

enablement being provided in particular for the subject matter of the appended claims, which make up a part of the original description.
As used above, the phrases "selected from the group consisting of," "chosen from," and the like include mixtures of the specified materials.

Where a numerical limit or range is stated herein, the endpoints are included. Also, all values and subranges within a numerical limit or range are specifically included as if explicitly written out.
The above description is presented to enable a person skilled in the art to make and use the invention, and is provided in the context of a particular application and its requirements. Various modifications to the preferred embodiments will be readily apparent to those skilled in the art, and the generic principles defined herein may be applied to other embodiments and applications without departing from the spirit and scope of the invention. Thus, this invention is not intended to be limited to the embodiments shown, but is to be accorded the widest scope consistent with the principles and features disclosed herein.

Having generally described this invention, a further understanding can be obtained by reference to certain specific examples, which are provided herein for purposes of illustration only.

Examples

Example 1: Activities of Tail/Tail and Head/Head protein configurations

Cloning of the "AvrBs3" RVD array in the TAL backbone

The amino acid sequences of the N-terminal, C-terminal domains and RVDS were based on the AvrBs3 TAL (ref: GenBank: X16130.1, SEQ ID NO: 1). The yeast expression TAL backbones used in these experiment (pCLS13597, SEQ ID NO: 2 and pCLS12843, SEQ ID NO: 3) were derived from the pCLS8422 (SEQ ID NO: 4) and pCLS8426 (SEQ ID NO: 5) respectively, where a second FokI Catalytic head was introduced by blunt end cloning in the EcoRV site and subsequent sequencing to validate the FokI orientation and sequence integrity. The cassette comprised between the NcoI and BamHI were further subcloned in pCLS7183 (SEQ ID NO: 6) leading to the final yeast cloning backbones. These backbone, pCLS13597 and pCLS12843, contain an additional N-terminal NLS sequence followed by an HA tag upstream the first FokI catalytic head. The C-terminal and the N-terminal domains (complete Nter domain, SEQ ID NO: 7, associated with pCLS12843 or delta152 truncated Nter, SEQ ID NO: 8, associated with pCLS13597) are separated by two BsmBI restriction sites. The RVD arrays (17.5 RVDs, SEQ ID NO: 9), targeting the AvrBs3 sequence (SEQ ID NO: 10)

was assembled in solution by creation of the four tetra-RVDs (coding for blocks 1-4, 5-8, 9-12 and 13-16) from di-RVDs. The two octa-RVDs (1-8 and 9-16) were assembled starting from the four tetra-RVDs to form the hexadeca-RVD (1-16). The final RVD array was prepared by assembly of the hexadeca-RVD with the terminal 1.5 RVD block. All the assembly steps were done using restriction enzymes SfaNI and BbvI and T4 DNA ligase, following classical molecular biology protocols.
The final array was then subcloned in both pCLS13597 and pCLS12843 using type IIs restriction enzymes BsmBI for the receiving plasmid and BbvI and SfaNI for the inserted RVD sequence leading to pCLS14333 (SEQ ID NO: 11) and pCLS12944 (SEQ ID NO: 12). DNA coding for the TAL was amplified in E. coli, recovered by standard miniprep techniques and sequenced to assess the integrity of the insert.

Activities in yeast

The activity of the two individual FokI catalytic heads and their dependence toward the spacer length were tested at 37°C and 30°C in our yeast SSA assay previously described (International PCT Applications WO 2004/067736 and in (Epinat, Arnould et al. 2003; Chames, Epinat et al. 2005; Arnould, Chames et al. 2006; Smith, Grizot et al. 2006) on pseudo-palindromic sequences; two identical recognition sequences are placed facing each other (Tail/Tail or Head/Head) on both DNA strands (SEQ ID NO: 16 to 51 (Tail/Tail) and 52 to 82 (Head/Head). All the yeast target reporter plasmids containing the TALEN DNA target sequences were constructed as previously described (International PCT Applications WO 2004/067736 and in (Epinat, Arnould et al. 2003; Chames, Epinat et al. 2005; Arnould, Chames et al. 2006; Smith, Grizot et al. 2006). TALEN cleavage activity levels, in yeast, of individual clones on the complete sets of targets are presented in table 1 (Tail/Tail) and table 2 (Head/Head).

Example 2: Activities of Head/Tail protein configuration

Cloning of the "GCT" RVD array in the TAL backbone

The amino acid sequences of the N-terminal, C-terminal domains and RVDS were based on the AvrBs3 TAL (ref: GenBank: X16130.1, SEQ ID NO: 1). The yeast expression TAL backbone used in these experiment (pCLS13597, SEQ ID NO: 2) was derived from the pCLS8422 (SEQ ID NO: 4), where a second FokI Catalytic head was introduced by blunt end cloning in the EcoRV site and subsequent sequencing to validate the FokI orientation and sequence integrity. The cassette comprised between the NcoI and BamHI was further subcloned in pCLS7183 (SEQ ID NO: 6)

leading to the final cloning yeast backbone. This backbone, pCLS13597, contains an additional N-terminal NLS sequence followed by an HA tag upstream the first FokI catalytic head. The C-terminal and the N-terminal domains are separated by two BsmBI restriction sites. The RVD arrays (SEQ ID NO: 13), targeting the repeated GCT sequences (SEQ ID NO: 14) was synthesized using a solid support method composed of consecutive restriction/ligation/washing steps. In brief the first block (coding for a di-RVD) was immobilized on a solid support through biotin/streptavidin interaction, the second bloc (tri-RVD) is then ligated to the first and after SfaNI digestion a third bloc (tri-RVD) is coupled. The process is repeated using tri- or di-RVD blocs upon obtaining of the desired RVD array. The product is cloned in a classical pAPG10 cloning plasmid for amplification in E. coli and sequencing. The RVD array was then subcloned in the pCLS13597 using type IIs restriction enzymes BsmBI for the receiving plasmid and BbvI and SfaNI for the inserted RVD sequence leading to pCLS14332 (SEQ ID NO: 15). DNA coding for the TAL was amplified in E. coli, recovered by standard miniprep techniques and sequenced to assess the integrity of the insert.

Activities in yeast

The activity of the pCLS14332 construct were tested at 37°C and 30°C in our yeast SSA assay previously described (International PCT Applications WO 2004/067736 and in Epinat, Arnould et al. 2003; Chames, Epinat et al. 2005; Arnould, Chames et al. 2006; Smith, Grizot et al. 2006) either on targets containing two identical CTG repeated sequences on the complementary strand (5'-3'), corresponding to the CAG repetition on the coding strand, separated by a spacer of 12, 15, 18, 21, 24, 27 or 30 bps (SEQ ID NO: 83 to 90) or a target composed exclusively of CTG repetitions on the complementary strand (5'-3'), corresponding to the CAG repetition on the coding strand, (SEQ ID NO : 90). All the yeast target reporter plasmids containing the TALEN DNA target sequences were constructed as previously described (International PCT Applications WO 2004/067736 and in (Epinat, Arnould et al. 2003; Chames, Epinat et al. 2005; Arnould, Chames et al. 2006; Smith, Grizot et al. 2006). TALEN cleavage activity levels, in yeast, of individual clones on the complete set of targets are presented in table 3.

List of cited references

Boch, J., H. Scholze, et al. (2009). "Breaking the code of DNA binding specificity of TAL-type III effectors." Science 326(5959): 1509-12.

Cermak, T., E. L. Doyle, et al. (2011). "Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting." Nucleic Acids Res 39(12): e82.

Christian, M., T. Cermak, et al. (2010). "Targeting DNA double-strand breaks with TAL effector nucleases." Genetics 186(2): 757-61.

Deng, D., C. Yan, et al. (2012). "Structural basis for sequence-specific recognition of DNA by TAL effectors." Science 335(6069): 720-3.

Geissler, R., H. Scholze, et al. (2011). "Transcriptional activators of human genes with programmable DNA-specificity." PLoS One 6(5): e19509.

Huang, P., A. Xiao, et al. (2011). "Heritable gene targeting in zebrafish using customized TALENs." Nat Biotechnol 29(8): 699-700.

Kay S. et al. (2007)."A bacterial Effector Acts as a Plant Transcription factor and Induces a Cell Size Regulator." Science 318:648-651.

Li, L., M. J. Piatek, et al. (2012). "Rapid and highly efficient construction of TALE-based transcriptional regulators and nucleases for genome modification." Plant Mol Biol 78(4-5): 407-16.

Li, T., S. Huang, et al. (2010). "TAL nucleases (TALNs): hybrid proteins composed of TAL effectors and FokI DNA-cleavage domain." Nucleic Acids Res 39(1): 359-72.

Li, T., S. Huang, et al. (2011). "Modularly assembled designer TAL effector nucleases for targeted gene knockout and gene replacement in eukaryotes." Nucleic Acids Res 39(14): 6315-25.

Mahfouz, M. M., L. Li, et al. (2012). "Targeted transcriptional repression using a chimeric TALE-SRDX repressor protein." Plant Mol Biol 78(3): 311-21.

Mahfouz, M. M., L. Li, et al. (2011). "De novo-engineered transcription activator-like effector (TALE) hybrid nuclease with novel DNA binding specificity creates double-strand breaks." Proc Natl Acad Sci U S A 108(6): 2623-8.

Mak, A. N., P. Bradley, et al. (2012). "The crystal structure of TAL effector PthXo1 bound to its DNA target." Science 335(6069): 716-9.

Miller, J. C., S. Tan, et al. (2011). "A TALE nuclease architecture for efficient genome editing." Nat Biotechnol 29(2): 143-8.

Morbitzer, R., J. Elsaesser, et al. (2011). "Assembly of custom TALE-type DNA binding domains by modular cloning." Nucleic Acids Res 39(13): 5790-9.

Moscou, M. J. and A. J. Bogdanove (2009). "A simple cipher governs DNA recognition by TAL effectors." Science 326(5959): 1501.

Mussolino, C., R. Morbitzer, et al. (2011). "A novel TALE nuclease scaffold enables high genome editing activity in combination with low toxicity." Nucleic Acids Res 39(21): 9283-93.

Sander, J. D., L. Cade, et al. (2011). "Targeted gene disruption in somatic zebrafish cells using engineered TALENs." Nat Biotechnol 29(8): 697-8.

Tesson, L., C. Usal, et al. (2011). "Knockout rats generated by embryo microinjection of TALENs." Nat Biotechnol 29(8): 695-6.

Weber, E., R. Gruetzner, et al. (2011). "Assembly of designer TAL effectors by Golden Gate cloning." PLoS One 6(5): e19722.

Zhang, F., L. Cong, et al. (2011). "Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription." Nat Biotechnol 29(2): 149-53.

Tail/Tail orientation		
spacer [bp]	pCLS12944	pCLS14333
5	+/-	n.d.
6	+	n.d.
7	n.d.	n.d.
8	+	+/-
9	++	+
10	+++	+++
11	+++	+++
12	+++	+++
13	+++	+++
14	+++	+++
15	+++	+++
16	+++	+++
17	+++	+++
18	+++	+++
19	+++	+++
20	+++	+++
21	+++	+++
22	+++	+++
23	+++	+++
24	+++	+++
25	+++	+++
26	+++	+++
27	+++	+++
28	+++	+++
29	+++	+++
30	+++	+++
31	+++	+++
32	+++	+++
33	+++	+++
34	+++	+++
35	+++	+++
36	+++	+++
37	+++	++
38	+++	+
39	+++	+
40	+++	n.d.

Table 1

Head/Head orientation		
spacer [bp]	pCLS12944	pCLS14333
5	++	n.d.
6	++	n.d.
7	++	n.d.
8	++	n.d.
9	++	n.d.
10	++	n.d.
11	++	n.d.
12	++	n.d.
13	++	n.d.
14	++	n.d.
15	+/-	n.d.
16	+	+/-
17	+	+/-
18	+	++
19	+	++
20	+	++
21	+	++
22	+	+++
23	+	+++
24	n.d.	+++
25	+/-	+++
26	n.d.	+++
27	+/-	+++
28	+/-	0,8
29	n.d.	n.d.
30	n.d.	n.d.
31	+	n.d.
32	n.d.	n.d.
33	+/-	++
34	+/-	+++
35	+/-	+++

Table 2

	pCLS14332
TiCAG	+++
TiCAG_sp12	+
TiCAG_sp15	++
TiCAG_sp18	+++
TiCAG_sp21	+++
TiCAG_sp24	+++
TiCAG_sp27	+++
TiCAG_sp30	+++

Table 3

CLAIMS

1) A monomer derived from a Transcription Activator-Like Effector (TALE) comprising:
   (a) A core scaffold comprising Repeat Variable Dipeptide regions (RVDs) having binding specificity to a DNA target sequence; and
   (b) At least one catalytic domain fused to the N-terminal of said core scaffold, wherein said catalytic domain, when in contact with a counterpart catalytic domain, form a catalytic entity to process nucleic acids adjacent to said target sequence.

2) A monomer derived from a Transcription Activator-Like Effector (TALE) according to claim 1, comprising:
   (c) A core scaffold comprising Repeat Variable Dipeptide regions (RVDs) having binding specificity to a DNA target sequence; and
   (d) At least two catalytic domains respectively fused to the N-terminal and C-terminal of said core scaffold, wherein at least one of said catalytic domain, when in contact with a counterpart catalytic domain, form a catalytic entity to process DNA adjacent to said target sequence.

3) A monomer according to claim 1 or 2, wherein at least one of said catalytic domains are fused to said core scaffold by a peptide linker.

4) A monomer according to any one of claim 1 or 3, wherein said catalytic domains are proteins.

5) A monomer according to any one of claims 1 to 4, wherein said catalytic entity has a nuclease activity.

6) A monomer according to any one of claim 1 to 5, wherein said catalytic and counterpart domains form a homodimer or heterodimer catalytic entity.

7) A monomer according to any one of claims 1 to 6, wherein said catalytic entity has a type IIS nuclease activity.

8) A monomer according to any one of claim 1 to 7, wherein said catalytic domain is a polypeptide comprising an amino acid sequence having at least 80%, preferably at least 90%, more preferably at least 95% identity with any of SEQ ID NO: 81.

9) A monomer according to claim 8, wherein said catalytic domain is FokI.

10) The monomer according to any one of claims 1 to 6, wherein said catalytic domain is a monomer of meganuclease.

11) A protein monomer according to any one of claim 1 to 4, wherein said catalytic entity emits a recordable reporter signal.

12) A set of at least two protein monomers, wherein said first and second monomers each comprise:

i. A core scaffold comprising Repeat Variable Dipeptide regions (RVDs) having binding specificity to a DNA target sequence; and

ii. A catalytic domain; the catalytic domain of the first monomer being fused to the N-terminal of its core scaffold, and the catalytic domain of the second monomer being fused to the C-terminal of its core scaffold, said two catalytic domains, when they are in contact, forming a catalytic entity being able to process DNA between the respective target sequences of said two monomers.

13) A set of at least two protein monomers according to claim 12, wherein the DNA target sequences of said two monomers are located on the same DNA strand of a double stranded DNA.

14) A set of at least two monomers according to claim 12 or 13, wherein at least one of said monomer is a monomer according to any one of claims 1 to 11.

15) A set of at least two monomers according to any one of claims 12 to 14, wherein said catalytic entity induces double strand break (DSB) between the target sequences of said two monomers.

16) A set of at least two monomers according to any one of claims 12 to 15, wherein said DNA target sequences are located in a DNA sequence comprising multiple repeats.

17) A set of at least two monomers according to any one of claims 12 to 16, wherein said DNA target sequences are located in a DNA sequence including methylated sequences.

18) A set of at least two monomers according to any one of claims 12 to 17, wherein said set comprises at least a third monomer as defined in claims 1 to 11.

19) A single chain formed by a set of at least two monomers as defined in any one of claims 12 to 18.

20) A polynucleotide comprising a coding sequence for the polypeptide monomers of any one of claims 1 to 19.

21) A composition comprising protein monomers according to any one of claims 1 to 19 or a polynucleotide according to claim 20.

22) A method to process a double stranded DNA sequence into a cell, wherein said method comprises transfecting said cell with a polynucleotide according to claim 20 and expressing the monomers encoded by said polynucleotide into said cell.

23) A method to process a double stranded DNA sequence, in particular a DNA sequence devoid of T on one of its strand, said method comprising:

(b) Identifying said double stranded DNA sequence in a cell ;

(c) Transfecting said cell with one or two nucleic acid encoding at least two protein monomers each comprising:

i. A core scaffold comprising Repeat Variable Dipeptide regions (RVDs) having binding specificity to a DNA target sequence on one strand of said double stranded DNA; and

ii. A catalytic domain; the catalytic domain of the first monomer being fused to the N-terminal of its core scaffold, and the catalytic domain of the second monomer being fused to the C-terminal of its core scaffold, said two catalytic domains, when they are in contact, forming a catalytic entity being able to process DNA between the respective target sequences of said two monomers.

(d) expressing said protein monomers into said cell; such that said monomers bind their respective target sequences and process DNA between said target sequences by contacting their catalytic domains.

24) A method according to claim 23, wherein said double stranded DNA is devoid of T on one of its DNA strand and said Repeat Variable Dipeptide regions (RVDs) are chosen to bind target sequences that are located on the complementary DNA strand comprising T bases.

25) A method according to claim 23 to process a double stranded DNA sequence comprising putative trinucleotide repeats expansion, said method comprising:

(a) Providing a cell which may contain a double stranded DNA sequence comprising such putative trinucleotide repeats expansion devoid of T on one of its strand;

(b) Determining a threshold number of trinucleotide repeats under which no processing is desired;

(c) Transfecting said cell with one or two nucleic acid encoding at least two protein monomers each comprising:

i. A core scaffold comprising Repeat Variable Dipeptide regions (RVDs) having binding specificity to said number of trinucleotide repeats; and

ii. A catalytic domain; the catalytic domain of the first monomer being fused to the N-terminal of its core scaffold, and the catalytic domain of the second monomer being fused to the C-terminal of its core scaffold, said two catalytic domains, when they are in contact, forming a catalytic entity being able to process DNA between the respective target sequences of said two monomers.

(d) Inducing expression of said chimeric protein monomer encoded by said nucleic acid; such that said protein monomers bind and process DNA between said trinucleotide repeats DNA target sequences when said repeats are equal or above said threshold number.

26) A method according to claim 25, wherein said nucleic acid basic pattern is n trinucleotides, wherein n is comprised between 2 and 15.

27) A method to modify a genome at a selected locus comprising the following steps:

(a) providing said genome in a cell;

(b) selecting two target sequences located upstream and downstream of a given locus of said genome;

(c) Transfecting said cell with one or two nucleic acids encoding at least two protein monomers each comprising:

i. A core scaffold comprising Repeat Variable Dipeptide regions (RVDs) having binding specificity respectively to said target sequences; and

ii. A catalytic domain; the catalytic domain of the first monomer being fused to the N-terminal of its core scaffold, and the catalytic domain of the second monomer being fused to the C-terminal of its core scaffold,

said two catalytic domains having an activity on the nucleic acids present at the locus when said core scaffolds bind to their target sequences,

(d) Inducing expression of said protein monomers encoded by said nucleic acid(s); such as to obtain processing of the genome at said locus by both catalytic domains.

28) A method according to any one of claims 23 to 27, wherein said catalytic entity or domain induces double strand break (DSB) between the target sequences of said two monomers.

29) A method to measure the number of putative repeated sequences within a double stranded genome DNA sequence, in particular repeated sequences devoid of T on one strand of said genome sequence, said method comprising :

(a) providing said DNA genome in a cell

(b) Transfecting said cell with one or two nucleic acids encoding at least two protein monomers each comprising:

I. A core scaffold comprising Repeat Variable Dipeptide regions (RVDs) having binding specificity to said repeated DNA sequence as a target sequence; and

II. A catalytic domain; the catalytic domain of the first monomer being fused to the N-terminal of its core scaffold, and the catalytic domain of the second monomer being fused to the C-terminal of its core scaffold, said two catalytic domains, when they are in contact, forming a catalytic entity being able to produce a reporter signal.

(c) Inducing expression of said protein monomers encoded by said nucleic acid(s);

(d) Recording said reporter signal intensity;

(e) Deducing from said reporter signal intensity the number of putative repeated sequences within said DNA sequence into the genome of said cell; and

(f) Optionally comparing this reporter signal intensity with the reporter signal intensity recovered from a control cell.

30) The method of any one of claims 22 to 29, wherein the cell is a eukaryotic cell.

31) The method of any one of claims 22 to 30, wherein the cell is a mammalian cell.

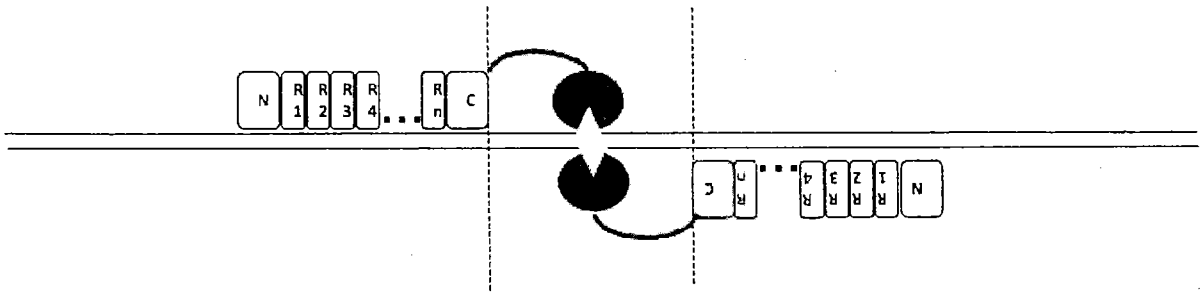
32) The method of any one of claims 22 to 31, wherein the cell is a plant cell.

33) The method of any one of claims 22 to 32, wherein said target DNA sequence is a chromosomal sequence.

34) The method of any one of claims 22 to 32, wherein said target DNA sequence is an episomal sequence.

35) The method of any one of claims 22 to 34, wherein said target DNA sequence is an organelle sequence.

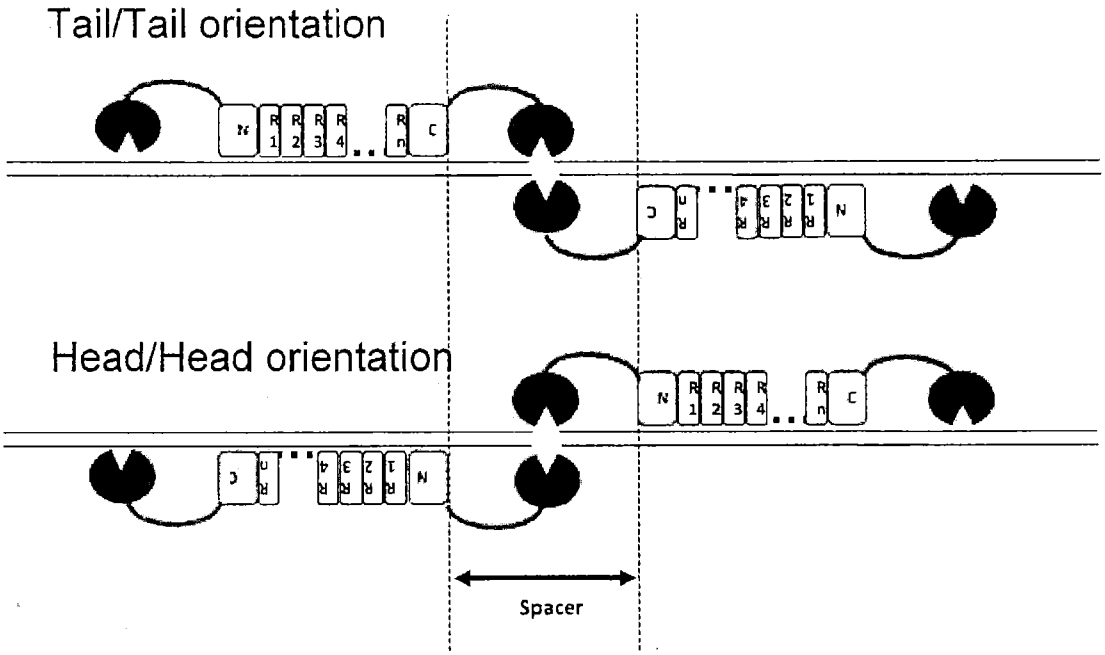
Classical TALEN Architecture



Legend

- N - TALE protein N-terminal region (including truncations thereof)
- C - TALE protein C-terminal region (including truncations thereof)
- R 1 - TALE protein repeat region (n = final repeat; depends on construct)
- ... - Additional TALE repeat regions (for brevity, not shown)
- Catalytic domain - Catalytic domain for DNA strand cleavage (can be cleavase or nickase domain)
- Protein linker region
- Double-strand DNA

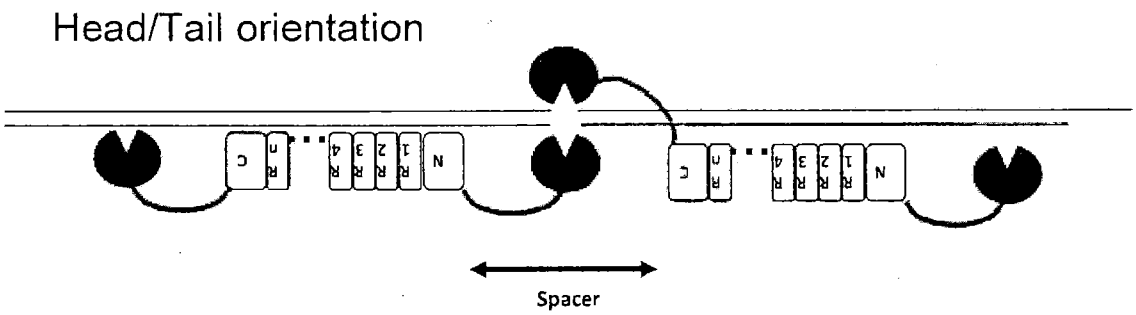
Figure 1



Legend

- N - TALE protein N-terminal region (including truncations thereof)
- C - TALE protein C-terminal region (including truncations thereof)
- R 1 - TALE protein repeat region (n = final repeat; depends on construct)
- ... - Additional TALE repeat regions (for brevity, not shown)
- Catalytic domain - Catalytic domain for DNA strand cleavage (can be cleavase or nickase domain)
- Protein linker region
- Double-strand DNA

Figure 2



Legend

- N - TALE protein N-terminal region (including truncations thereof)
- C - TALE protein C-terminal region (including truncations thereof)
- R 1 - TALE protein repeat region (n = final repeat; depends on construct)
- ... - Additional TALE repeat regions (for brevity, not shown)
- Catalytic domain - Catalytic domain for DNA strand cleavage (can be cleavase or nickase domain)
- Protein linker region
- Double-strand DNA

Figure 3

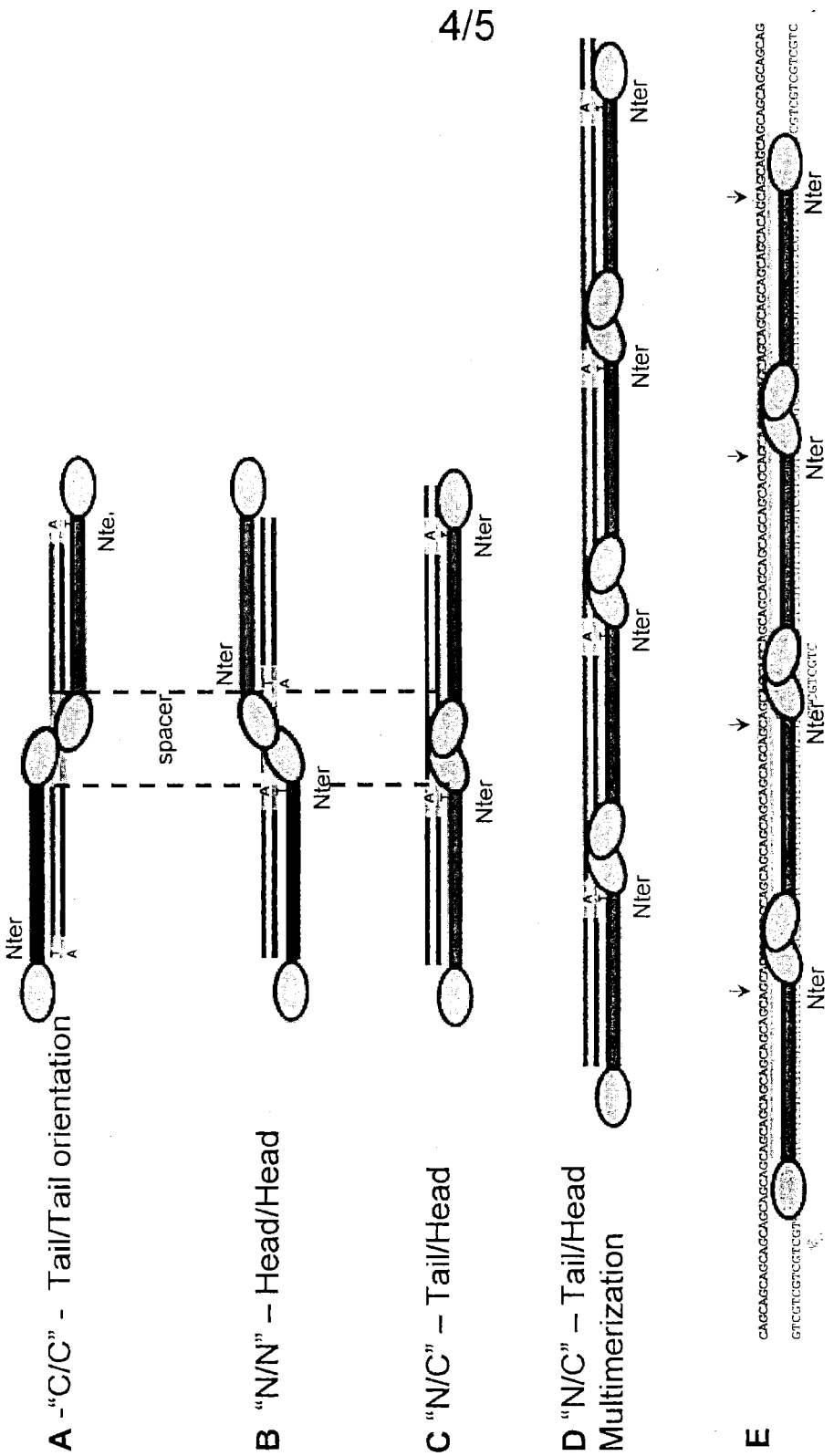

A -"C/C" - Tail/Tail orientation
B "N/N" – Head/Head
C "N/C" – Tail/Head
D "N/C" – Tail/Head Multimerization
E
Nter
spacer

Figure 4

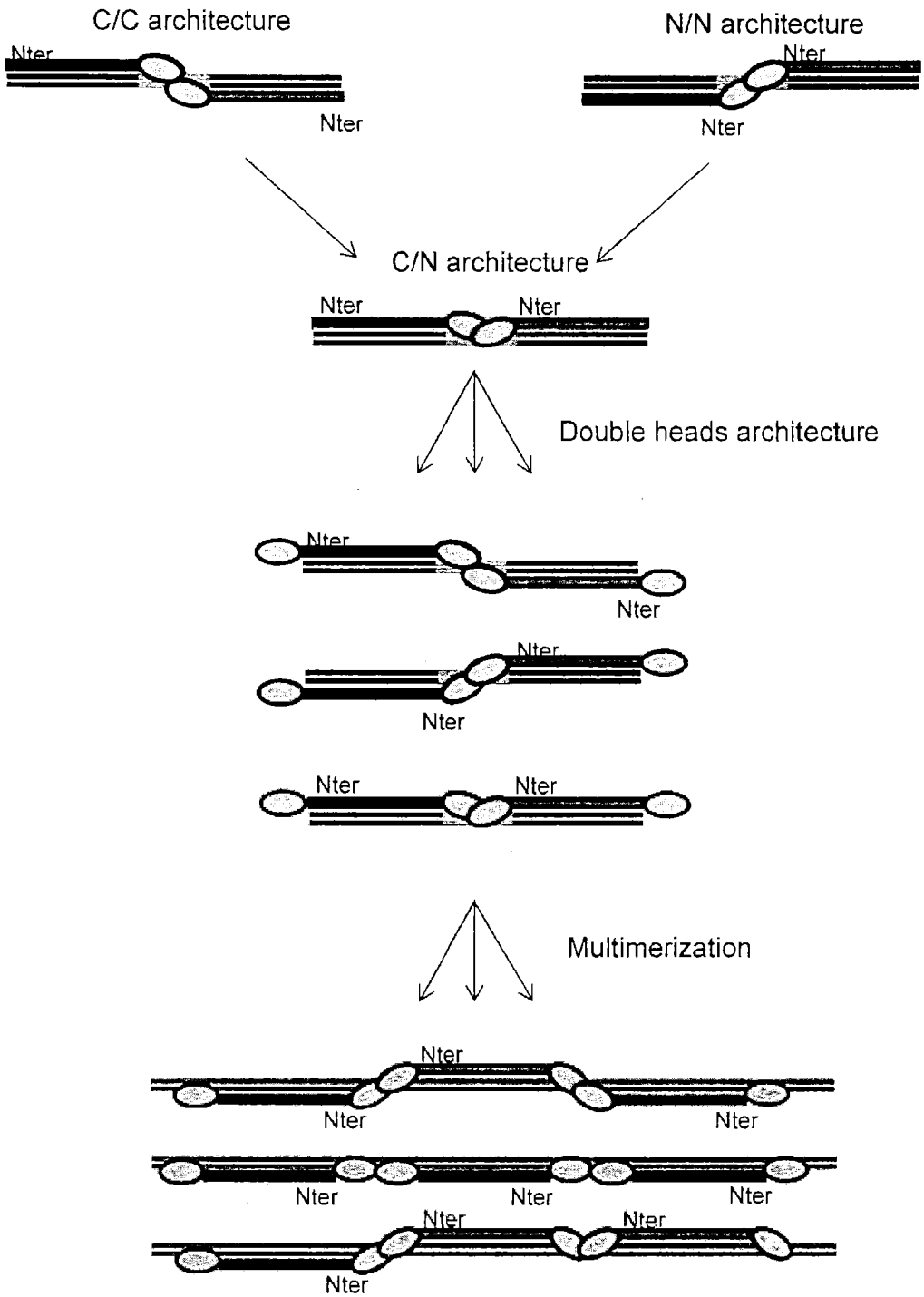

C/C architecture
Nter
Nter
N/N architecture
Nter
Nter
C/N architecture
Nter
Nter
Double heads architecture
Nter
Nter
Nter
Nter
Nter
Nter
Multimerization
Nter
Nter
Nter
Nter
Nter
Nter
Nter
Nter
Nter

Figure 5