



(51) International Patent Classification:

C12N 9/22 (2006.01) C12N 15/52 (2006.01)  
C12N 15/113 (2010.01) C12N 15/85 (2006.01)

(21) International Application Number:

PCT/US2015/064352

(22) International Filing Date:

7 December 2015 (07.12.2015)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

62/088,440 5 December 2014 (05.12.2014) US

(71) Applicant: APPLIED STEMCELL, INC. [US/US]; 521 Cottonwood Dr, Suite #111, Milpitas, CA 95035 (US).

(72) Inventor: TSAI, Ruby, Yanru; 521 Cottonwood Dr, Suite #111, Milpitas, CA 95035 (US).

(74) Agent: ZHU, James; Jun He Law Offices P.C., 2275 East Bayshore Road Suite 101, Palo Alto, CA 94303 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM,

DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, 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).

Declarations under Rule 4.17:

— as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))

Published:

— with international search report (Art. 21(3))  
— with sequence listing part of description (Rule 5.2(a))

(54) Title: SITE-DIRECTED CRISPR/RECOMBINASE COMPOSITIONS AND METHODS OF INTEGRATING TRANSGENES

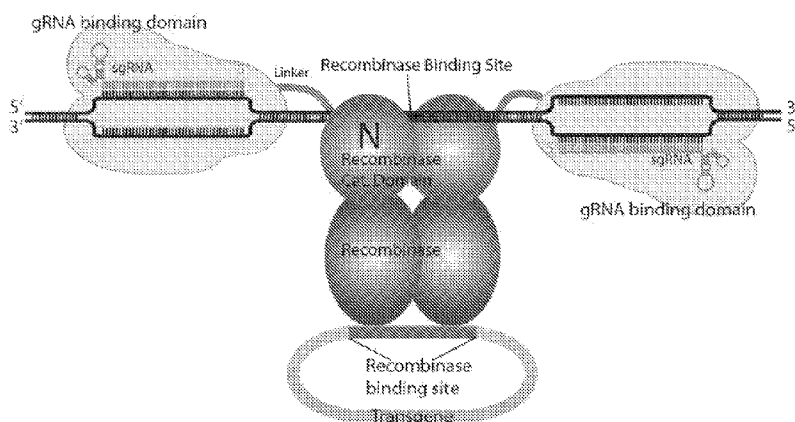


FIG. 1

(57) Abstract: Disclosed herein are targeted chimeric polypeptides, compositions thereof, expression vectors, and methods of use thereof, for the generation of transgenic cells, tissues, plants, and animals. The compositions, vectors, and methods of the present invention are also useful in gene therapy and cell therapy techniques. The chimeric polypeptide includes a CRISPR-Cas domain and a recombinase domain.

WO 2016/090385 A1

## SITE-DIRECTED CRISPR/RECOMBINASE COMPOSITIONS AND METHODS OF INTEGRATING TRANSGENES

---

### CROSS-REFERENCE TO RELATED APPLICATIONS

5 [0001] This application claims priority to U.S. provisional patent application no. 62/088,440, filed December 5, 2014, the disclosure of which is incorporated herein by reference.

### FIELD OF THE INVENTION

10 [0002] The present invention generally relates to compositions and methods used for integrating transgenes into the genome of a cell.

### BACKGROUND OF THE INVENTION

[0003] RNA-guided Cas9 nucleases derived from clustered regularly interspaced short palindromic repeats (CRISPR)-Cas systems have provided a versatile tool for editing the genome of diverse organisms. However, current technologies based on CRISPR-Cas system have limited ability of inserting large DNA fragments and are unable to perform homology-based editing such as targeted transgene insertion in non-dividing cells (e.g., neurons) or in cells with DNA homologous recombination deficiency. Therefore, there remains a need for new genome engineering technologies that are affordable, easy to set up and capable of editing genome in non-dividing cells or cells with DNA homologous recombination deficiency.

15  
20

### BRIEF SUMMARY OF THE INVENTION

[0004] Disclosed herein are targeted chimeric polypeptides, including compositions thereof, expression vectors, and methods of use thereof, for the generation of transgenic cells, tissues, plants, and animals. The compositions, vectors, and methods of the present invention are also useful in gene therapy and cell therapy techniques.

25

[0005] In one aspect, the present disclosure relates to a chimeric polypeptide. The polypeptide includes: 1) a CRISPR-Cas domain and 2) a recombinase domain. In various embodiments, the CRISPR-Cas domain is capable of binding to a guide RNA, which is capable of hybridizing to a target DNA sequence, and the recombinase domain comprises a site-specific recombinase, a mutant thereof or a fragment thereof.

30

[0006] In some embodiments, the CRISPR-Cas domain comprises a Type-II Cas9 protein (SEQ ID NO.: 1) or a fragment thereof. In certain embodiments, the CRISPR-Cas

domain does not have endonuclease activity. In certain embodiments, the CRISPR-Cas domain is a catalytically dead Cas9 (dCas9). In certain embodiments, the CRISPR-Cas domain comprises an amino acid sequence of SEQ ID 1.

[0007] In some embodiments, the recombinase domain comprises a site-specific recombinase. In some embodiments, the site-specific recombinase is a tyrosine recombinase. Examples of the tyrosine recombinase include, but are not limited to, Cre, Flp and the lambda integrase. In some embodiments, the site-specific recombinase is a serine recombinase. Examples of the serine recombinases include gamma-delta resolvase, Tn3 resolvase (SEQ ID NO.: 2), Sin resolvase, Gin invertase, Hin invertase, Tn5044 resolvase, IS607 transposase and integrase. In some embodiments, the site-recombinase is an integrase selected from the group consisting of Bxb1, wBeta, BL3, phiR4, A118, TG1, MR11, phi370, SPBc, TP901-1, phiRV, FC1, K38, phiBT1 and phiC31 (SEQ ID NO. 3).

[0008] In certain embodiments, the recombinase domain comprises a mutant version or a fragment of a site-specific recombinase as disclosed above. In certain embodiments, the mutant has altered target specificity. In some embodiments, the mutant has reduced target specificity. In certain embodiments, the recombinase domain comprises an activated mutant of a recombinase, including but without limitation activated mutant of Tn3 and activated mutant of phiC31. In certain embodiments, the recombinase domain is a catalytic domain of an activated mutant of a recombinase. In certain embodiments, the recombinase domain comprises a catalytic domain of activated mutant of Tn3. In certain embodiments, the recombinase domain has an amino acid sequence selected from the group consisting of SEQ ID NO: 5 and the variants thereof as listed in Table 1.

[0009] In certain embodiments, the polypeptide further comprises a linker that links the CRISPR-Cas domain and the recombinase domain. In certain embodiments, the linker has an amino acid sequence of The-Ser .

[0010] In another aspect, the present application discloses a composition comprising the polypeptide as disclosed above, and the guide RNA. In certain embodiment, the composition further comprises a donor DNA including a first nucleotide sequence recognized by a site-specific recombinase and a transgene.

[0011] In yet another aspect, the present application discloses a composition comprising one or more vectors comprising (i) a first nucleotide sequence encoding the polypeptide as disclosed above; and (ii) a second nucleotide sequence encoding the guide RNA, wherein the first and second nucleotide are located on same or different vectors. In certain embodiments, the composition further comprises a donor DNA comprising a third

nucleotide sequence recognized by the polypeptide as disclosed above and a transgene. In certain embodiments, the composition further comprises a fourth nucleotide sequence encoding the site-specific recombinase.

[0012] Another aspect of the present application relates to a method of integrating a

5 transgene into the genome of a cell. In certain embodiments, the method comprises introducing into the cell the composition as disclosed above. In certain embodiments, the cell is a eukaryotic cell, e.g., a mammalian or human cell. In certain embodiments, the cell is a one-cell embryo.

[0013] These and other features, aspects, and advantages of the present invention will

10 become better understood with regard to the following description, appended claims and accompanying drawings.

#### BRIEF DESCRIPTION OF THE FIGURES

[0014] FIG. 1 is a schematic illustration of the complex of gRNA-directed Cas-

15 recombinase chimeric polypeptide (CasR complex) formed at the target sequence to mediate the recombination between the donor DNA and the target DNA. Typically, the target DNA include a sequence (the target site) similar to a recombinase recognition site. The complex comprises two chimeric polypeptides (CasR), each of which includes a CRISPR-Cas domain (gRNA binding domain) and a recombinase domain. The CRISPR-Cas domain is capable of  
20 binding to a guide RNA. The Cas domain and the recombinase domain are linked by a peptide linker. Upon binding to guide RNA(s) that hybridizes to sequences flanking the target site, two CasRs are recruited to the target site, respectively, wherein the recombinase domains form a dimer at the target site. A tetramer of recombinase or recombinase domain is then formed at the target site and activated by binding to its cognate sites, and mediates the  
25 recombination between the donor DNA and the target DNA.

[0015] FIG. 2A is another embodiment of the complex of site-directed recombinase formed at the target sequence to mediate the recombination between the donor DNA and the target DNA. The complex comprises a first CRISPR-Cas system guide RNA that hybridizes

30 hybridize with a second sequence in the target DNA and at least two site-directed modifying polypeptides comprising a guide-RNA-binding domain that is capable of binding to the first or second CRISPR-Cas system guide RNA, and a recombinase domain. The gRNA binding domain and recombinase domain are linked by a peptide linker. Upon binding to the guide

RNA through the guide-RNA-binding domain, the site-directed modifying polypeptides are recruited to the target sequence. A tetramer of recombinase or recombinase domain is formed at the target sequence and activated by binding to its cognate sites, and mediates the recombination between the donor DNA and the target DNA. The cognate sites are either located on a helper DNA or on a donor vector.

**[0016]** FIG. 2B illustrates another embodiment of the complex of site-directed recombinase formed at the target sequence to mediate the recombination between the donor DNA and the target DNA. Integrase is used as an example to illustrate how site-directed CRISPR/recombinase works for transgene insertion. Upon binding to the guide RNA through the guide-RNA-binding domain, the site-directed modifying polypeptides are recruited to the target sequence. A tetramer of integrase or a domain thereof, or a variant with altered target specificity is formed at the target sequence and activated by binding to its cognate sites, attP and attB, and mediates the recombination between the donor DNA and the target DNA. The attP site is located on a helper DNA or attB site is on a donor vector.

**[0017]** FIG.3 illustrates E. coli colony color assays of recombination used to test the capability of CasR to mediate recombination of C-sites. Strains containing a substrate plasmid are co-transformed with a CasR expression plasmid and a gRNA expression plasmid. Transformants are selected on MacConkey-galactose indicator plates (MacConkey agar base (Difco) supplemented with 1% galactose, kanamycin (to select for the substrate plasmid or its resolution product), and ampicillin and/or chloramphenicol (to select for pCasR expression plasmids). Pale coloured (galK<sup>-</sup>) colonies indicate recombination (resolution) proficiency, whereas red (galK<sup>+</sup>) colonies indicate lack of resolution.

**[0018]** FIG. 4 illustrates Tn3 res site I and the six 22 bp TATA-containing sequences similar to Tn3 res site I, which are aligned with the central TATA of site I. Bases identical to site I within the central 12 bp of the TATA-containing sequences are underlined.

**[0019]** FIG 5 illustrates sequences of site I and C-sites.

#### DETAILED DESCRIPTION OF THE INVENTION

**[0020]** In the Summary of the Invention above and in the Detailed Description of the Invention, and the claims below, and in the accompanying drawings, reference is made to particular features (including method steps) of the invention. It is to be understood that the disclosure of the invention in this specification includes all possible combinations of such particular features. For example, where a particular feature is disclosed in the context of a particular aspect or embodiment of the invention, or particular claim, that feature can also be

used, to the extent possible, in combination with and/or in the context of other particular aspects and embodiments of the invention, and in the invention generally.

[0021] The term “comprises” and grammatical equivalents thereof are used herein to mean that other components, ingredients, steps, etc. are optionally present. For example, an article “comprising” (or “which comprises”) components A, B, and C can consist of (i.e., contain only) components A, B, and C, or can contain not only components A, B, and C but also one or more other components.

[0022] Where reference is made herein to a method comprising two or more defined steps, the defined steps can be carried out in any order or simultaneously (except where the context excludes that possibility), and the method can include one or more other steps which are carried out before any of the defined steps, between two of the defined steps, or after all the defined steps (except where the context excludes that possibility).

[0023] Where a range of value is provided, it is understood that each intervening value, to the tenth of the unit of the lower limit unless the context clearly dictate otherwise, between the upper and lower limit of that range and any other stated or intervening value in that stated range, is encompassed within the disclosure, subject to any specifically excluded limit in the stated range. Where the stated range includes one or both of the limits, ranges excluding either or both of those included limits are also included in the disclosure.

[0024] It will be appreciated that for simplicity and clarity of illustration, where appropriate, reference numerals have been repeated among the different figures to indicate corresponding or analogous elements. In addition, numerous specific details are set forth in order to provide a thorough understanding of the embodiments described herein. However, the embodiments described herein can be practiced without their specific details. In other instances, methods, procedures and components have not been described in detail so as not to obscure the related relevant function being described. Also, the description is not to be considered as limiting the scope of the implementations described herein. It will be understood that descriptions and characterizations of the embodiments set forth in this disclosure are not to be considered as mutually exclusive, unless otherwise noted.

[0025] The following definitions are used in the disclosure:

[0026] The terms "nucleic acid molecule" and "polynucleotide" are used interchangeably and refer to a polymeric form of nucleotides of any length, either deoxyribonucleotides or ribonucleotides, or analogs thereof. Polynucleotides may have any three-dimensional structure, and may perform any function, known or unknown. Non-limiting examples of polynucleotides include a gene, a gene fragment, exons, introns,

messenger RNA (mRNA), transfer RNA, ribosomal RNA, ribozymes, cDNA, recombinant polynucleotides, branched polynucleotides, plasmids, vectors, isolated DNA of any sequence, isolated RNA of any sequence, nucleic acid probes, and primers.

[0027] As used herein, the term "vector" refers to a nucleic acid molecule capable of transporting between different genetic environments another nucleic acid to which it has been operatively linked. Preferred vectors are those capable of autonomous replication and expression of structural gene products present in the DNA segments to which they are operatively linked. Vectors, therefore, preferably contain the replicons and selectable markers described earlier. Vectors include, but are not necessarily limited to, expression vectors.

[0028] As used herein, the term "expression vector" refers to a plasmid, virus, phagemid, or other vehicle known in the art that has been manipulated by insertion or incorporation of heterologous DNA, such as nucleic acid encoding the fusion proteins herein or expression cassettes provided herein. Such expression vectors typically contain a promoter sequence for efficient transcription of the inserted nucleic acid in a cell. The expression vector typically contains an origin of replication, a promoter, as well as specific genes that permit phenotypic selection of transformed cells.

[0029] In one aspect, the present disclosure provides a CRISPR-Cas recombinase (CasR), which can be used to recombine DNA in bacterial and mammalian cell. The CasR includes a CRISPR-Cas domain and a recombinase domain. The CRISPR-Cas domain mediates the recognition of the CasR to a specific target sequence via a guide RNA, and the recombinase domain carries out the recombination.

[0030] As used herein, the term "recombinase" or "site-specific recombinase" refers to a family of highly specialized enzymes that promote DNA rearrangement between specific target sites (Greindley et al., 2006; Esposito, D., and Scocca, J. J., *Nucleic Acids Research* 25, 3605-3614 (1997); Nunes-Duby, S. E., et al, *Nucleic Acids Research* 26, 391-406 (1998); Stark, W. M., et al, *Trends in Genetics* 8, 432-439 (1992)). Virtually all site-specific recombinases can be categorized within one of two structurally and mechanistically distinct groups: the tyrosine (e.g., Cre, Flp, and the lambda integrase) or serine (e.g., phiC31 integrase, gamma-delta resolvase, Tn3 resolvase and Gin invertase) recombinases. Both recombinase families recognize target sites composed of two inversely repeated binding elements that flank a spacer sequence where DNA breakage and religation occur. The recombination process requires concomitant binding of two recombinase monomers to each target site: two DNA-bound dimers (a tetramer) then join to form a synaptic complex, leading to crossover

and strand exchange. "Hyperactive" forms of Tn3 resolvase containing activating mutations in Tn3 resolvase can catalyze strand exchange at a core site of 28bp without accessory sites, presumably through reconfiguration of the tertiary/quaternary structure of the tetramer.

[0031] "Site-directed recombinase activity" or "site-directed recombination" as used  
5 herein refers to the sequence specific recombination of a first nucleic acid sequence with a second nucleic acid sequence, typically mediated by a site-specific recombinase. In general, site-specific recombination occurs at particular defined sequences recognized by the recombinase. In contrast to random integration, site-directed recombinase occurs at a particular sequence (*e.g.*, a recombinase attachment site) at a higher efficiency.

10 [0032] In certain embodiments, the site-specific recombinase is Tn3 resolvase. In certain embodiments, the site-specific recombinase is Gin Invertase. In certain embodiments, the site-specific recombinase is an integrase selected from Bxb1, wBeta, BL3 phiR4, A118, TG1, MR11, phi370, SPBc, TP901-1, phiRV, FC1, K38, phiBT1 and phiC31.

[0033] In some embodiments, the recombinase domain comprises a variant (mutant)  
15 of a site-specific recombinase or a fragment thereof, wherein the variant has altered target specificity. Methods have been established to change the target specificity of a recombinase. For example, Buchholz and Stewart utilized random mutagenesis by error-prone PCR, DNA shuffling, and a lacZ-based blue/white recombination screen to generate Flp variants with enhanced thermostability (Buchholz et al., 1998). In another example, accessory-factor  
20 independent variants of the lambda-Integrase have been identified by selection in the presence of defective integration host factor activity (Miller et al., 1980) and reversion analysis (Wu et al., 1997). In yet another example, Voziyanov et al. (2002) developed a dual-reporter screen that enabled direct readout of the effects of individual mutations on Flp specificity and identified a number of Flp variants with altered target specificity. In another  
25 example, Buchholze and Stewart developed a technique termed Substrate-Linked Protein Evolution (SLiPE) to physically link individual recombinase variants to their DNA substrate (Buchholz and Stewart, 2001). In some embodiments, the variant of the site-specific recombinase has a reduced target specificity.

[0034] More recently, mutants of several serine recombinases, *e.g.*, Tn3 have been  
30 identified that do not require accessory factors for recombination (Proudfoot et al., 2001, Zinc Finger Recombinases with Adaptable DNA Sequence Specificity. PloS ONE 6(4): e19537). It was shown that the native DNA binding domains (DBDs) of serine recombinases can be replaced with custom-designed ZFPs to generate chimeric zinc-finger recombinases (ZFRs). ZFRs are composed of an activated catalytic domain derived from the resolvase/invertase



family of serine recombinases and a zinc- finger DNA-binding domain that can be custom-  
designed to recognize almost any DNA sequence. ZFRs catalyze recombination between  
specific ZFR target sites that consist of two-inverted zinc- finger binding sites (ZFBS)  
flanking a central 20-bp core sequence recognized by the recombinase catalytic domain. In  
5 contrast to zinc- finger nucleases (ZFNs) and TAL effector nucleases (TALENs), ZFRs  
function autonomously and can excise and integrate transgenes in human and mouse cells  
without activating the cellular DNA damage response pathway. In principle, ZFRs capable  
of recognizing an extended number of sequences could be generated, however, the lack of  
zinc-finger domains capable of recognizing all possible DNA triplets limits the potential  
10 modular targeting capacity of these enzymes.

**[0035]** In a peptide or protein, suitable conservative substitutions of amino acids are  
known to those of skill in this art and may be made generally without altering the biological  
activity of the resulting molecule. Those of skill in this art recognize that, in general, single  
amino acid substitutions in non-essential regions of a polypeptide do not substantially alter  
15 biological activity (see, e.g. Watson et al. *Molecular Biology of the Gene*, 4th Edition, 1987,  
Benjamin/Cummings, p. 224). In particular, such a conservative variant has a modified amino  
acid sequence, such that the change(s) do not substantially alter the protein's (the conservative  
variant's) structure and/or activity, e.g., antibody activity, enzymatic activity, or receptor  
activity. These include conservatively modified variations of an amino acid sequence, i.e.,  
20 amino acid substitutions, additions or deletions of those residues that are not critical for  
protein activity, or substitution of amino acids with residues having similar properties (e.g.,  
acidic, basic, positively or negatively charged, polar or non-polar, etc.) such that the  
substitutions of even critical amino acids does not substantially alter structure and/or activity.  
Conservative substitution tables providing functionally similar amino acids are well known in  
25 the art. For example, one exemplary guideline to select conservative substitutions includes  
(original residue followed by exemplary substitution): Ala/Gly or Ser; Arg/Lys; Asn/Gln or  
His; Asp/Glu; Cys/Ser; Gln/Asn; Gly/Asp; Gly/Ala or Pro; His/Asn or Gln; Ile/Leu or Val;  
Leu/Ile or Val; Lys/Arg or Gln or Glu; Met/Leu or Tyr or He; Phe/Met or Leu or Tyr; Ser/Thr;  
Thr/Ser; Trp/Tyr; Tyr/Trp or Phe; Val/Ile or Leu. An alternative exemplary guideline uses the  
30 following six groups, each containing amino acids that are conservative substitutions for one  
another: (1) alanine (A or Ala), serine (S or Ser), threonine (T or Thr); (2) aspartic acid (D or  
Asp), glutamic acid (E or Glu); (3) asparagine (N or Asn), glutamine (Q or Gln); (4) arginine  
(R or Arg), lysine (K or Lys); (5) isoleucine (I or He), leucine (L or Leu), methionine (M or  
Met), valine (V or Val); and (6) phenylalanine (F or Phe), tyrosine (Y or Tyr), tryptophan (W

or Trp); (see also, e.g., Creighton (1984) *Proteins*, W. H. Freeman and Company; Schulz and Schirmer (1979) *Principles of Protein Structure*, Springer-Verlag). One of skill in the art will appreciate that the above-identified substitutions are not the only possible conservative substitutions. For example, for some purposes, one may regard all charged amino acids as conservative substitutions for each other whether they are positive or negative. In addition, individual substitutions, deletions or additions that alter, add or delete a single amino acid or a small percentage of amino acids in an encoded sequence can also be considered "conservatively modified variations" when the three-dimensional structure and the function of the protein to be delivered are conserved by such a variation.

10 **[0036]** In general, "CRISPR-Cas system" or "CRISPR system" refers collectively to transcripts and other elements involved in the expression of or directing the activity of CRISPR-associated ("Cas") genes, including sequences encoding a Cas gene, a *tracr* (transactivating CRISPR) sequence (e.g. *tracr*RNA or an active partial *tracr*RNA), a *tracr*-mate sequence (encompassing a "direct repeat" and a *tracr*RNA-processed partial direct repeat in the context of an endogenous CRISPR system), a guide sequence (also referred to as a "spacer" in the context of an endogenous CRISPR system), or other sequences and transcripts from a CRISPR locus. In some embodiments, one or more elements of a CRISPR system is derived from a type I, type II, or type III CRISPR system. In some embodiments, one or more elements of a CRISPR system are derived from a particular organism comprising an endogenous CRISPR system, such as *Streptococcus pyogenes*. In general, a CRISPR system is characterized by elements that promote the formation of a CRISPR complex at the site of a target sequence (also referred to as a protospacer in the context of an endogenous CRISPR system).

25 **[0037]** In general, a "CRISPR-Cas guide RNA" or "guide RNA" refers to an RNA that directs sequence-specific binding of a CRISPR complex to the target sequence. Typically, a guide RNA comprises (i) a guide sequence that has sufficient complementarity with a target polynucleotide sequence to hybridize with the target sequence and (ii) a transactivating *cr* (*tracr*) mate sequence. In some embodiments, the degree of complementarity between a guide sequence and its corresponding target sequence, when optimally aligned using a suitable alignment algorithm, is about or more than about 50%, 60%, 75%, 80%, 85%, 90%, 95%, 97.5%, 99%, or more. Optimal alignment may be determined with the use of any suitable algorithm for aligning sequences, non-limiting example of which include the Smith-Waterman algorithm, the Needleman-Wunsch algorithm, algorithms based on the Burrows-Wheeler Transform (e.g. the Burrows Wheeler Aligner), ClustalW, Clustal X, BLAT,

Novoalign (Novocraft Technologies, ELAND (Illumina, San Diego, Calif.), SOAP (available at soap.genomics.org.cn), and Maq (available at maq.sourceforge.net). In some embodiments, a guide sequence is about or more than about 5, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 35, 40, 45, 50, 75, or more nucleotides in length. In some  
5 embodiments, a guide sequence is less than about 75, 50, 45, 40, 35, 30, 25, 20, 15, 12, or fewer nucleotides in length. The ability of a guide sequence to direct sequence-specific binding of a CRISPR complex to a target sequence may be assessed by any suitable assay. For example, the components of a CRISPR system sufficient to form a CRISPR complex, including the guide sequence to be tested, may be provided to a host cell having the  
10 corresponding target sequence, such as by transfection with vectors encoding the components of the CRISPR sequence, followed by an assessment of preferential cleavage within the target sequence, such as by Surveyor assay as described herein. Similarly, cleavage of a target polynucleotide sequence may be evaluated in a test tube by providing the target sequence, components of a CRISPR complex, including the guide sequence to be tested and a control  
15 guide sequence different from the test guide sequence, and comparing binding or rate of cleavage at the target sequence between the test and control guide sequence reactions. Other assays are possible, and will occur to those skilled in the art.

**[0038]** In the context of formation of a CRISPR complex, a “target sequence” or “a sequence of a target DNA” refers to a sequence to which a guide sequence is designed to  
20 have complementarity, where hybridization between a target sequence and a guide sequence promotes the formation of a CRISPR complex. Full complementarity is not necessarily required, provided there is sufficient complementarity to cause hybridization and promote formation of a CRISPR complex. A target sequence may comprise any polynucleotide, such as DNA or RNA polynucleotides or DNA/RNA hybrid polynucleotides. In some  
25 embodiments, a target sequence is located in the nucleus or cytoplasm of a cell. In some embodiments, the target sequence may be within an organelle of a eukaryotic cell, for example, mitochondrion or chloroplast.

**[0039]** A guide sequence may be selected to target any target sequence. In some  
embodiments, the target sequence is a sequence within a genome of a cell. Exemplary target  
30 sequences include those that are unique in the target genome. For example, for the *S. pyogenes* Cas9, a unique target sequence in a genome may include a Cas9 target site of the form MMMMMMMNNNNNNNNNNXGG where NNNNNNNNNNNXGG (N is A, G, T, or C; and X can be any nucleotide) has a single occurrence in the genome. For the *S. thermophilus* CRISPR1Cas9, a unique target sequence in a genome may include a Cas9 target

site of the form MMMMMMMMNNNNNNNNNNNNNNXXAGAAW where  
NNNNNNNNNNNNXXAGAAW (SEQ ID NO: 2) (N is A, G, T, or C; X can be any  
nucleotide; and W is A or T) has a single occurrence in the genome. In each of these  
sequences “M” may be A, G, T, or C, and need not be considered in identifying a sequence as  
5 unique.

**[0040]** In some embodiments, a guide sequence is selected to reduce the degree of  
secondary structure within the guide sequence. Secondary structure may be determined by  
any suitable polynucleotide folding algorithm. Some programs are based on calculating the  
minimal Gibbs free energy. An example of one such algorithm is mFold, as described by  
10 Zuker and Stiegler (Nucleic Acids Res. 9 (1981), 133-148). Another example folding  
algorithm is the online webserver RNAfold, developed at Institute for Theoretical Chemistry  
at the University of Vienna, using the centroid structure prediction algorithm (see e.g. A. R.  
Gruber et al., 2008, *Cell* 106(1): 23-24; and PA Carr and GM Church, 2009, *Nature*  
*Biotechnology* 27(12): 1151-62).

**[0041]** As used herein, a tracr mate sequence includes any sequence that has  
sufficient complementarity with a tracr sequence to promote one or more of: (1) excision of a  
guide sequence flanked by tracr mate sequences in a cell containing the corresponding tracr  
sequence; and (2) formation of a CRISPR complex at a target sequence, wherein the CRISPR  
complex comprises the tracr mate sequence hybridized to the tracr sequence. In general,  
20 degree of complementarity is with reference to the optimal alignment of the tracr mate  
sequence and tracr sequence, along the length of the shorter of the two sequences. Optimal  
alignment may be determined by any suitable alignment algorithm, and may further account  
for secondary structures, such as self-complementarity within either the tracr sequence or  
tracr mate sequence. In some embodiments, the degree of complementarity between the tracr  
25 sequence and tracr mate sequence along the length of the shorter of the two when optimally  
aligned is about or more than about 25%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95%,  
97.5%, 99%, or higher.

**[0042]** In some embodiments, the guide RNA comprises a guide sequence fused to a  
tracr sequence, i.e., the tracr sequence and tracr mate sequence are contained within a single  
30 transcript, such that hybridization between the two produces a transcript having a secondary  
structure, such as a hairpin. In some embodiments, the tracr sequence is about or more than  
about 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 25, 30, 40, 50, or more  
nucleotides in length. Preferred loop forming sequences for use in hairpin structures are four  
nucleotides in length, and most preferably have the sequence GAAA. However, longer or

shorter loop sequences may be used, as may alternative sequences. The sequences preferably include a nucleotide triplet (for example, AAA), and an additional nucleotide (for example C or G). Examples of loop forming sequences include CAAA and AAAG. In an embodiment of the present application, the guide RNA has at least two or more hairpins. In preferred  
5       embodiments, the guide RNA has two, three, four or five hairpins. In a further embodiment of the invention, the guide RNA has at most five hairpins. In some embodiments, the guide RNA further includes a transcription termination sequence, preferably a polyT sequence, for example six T nucleotides. In some embodiments, the tracr sequence is a separate transcript from a transcript comprising the tracr mate sequence.

10       **[0043]**       In some embodiments, the CRISPR-Cas domain comprises a Cas protein. Non-limiting examples of Cas proteins include Cas1, Cas1B, Cas2, Cas3, Cas4, Cas5, Cas6, Cas7, Cas8, Cas9 (also known as Csn1 and Csx12), Cas10, Csy1, Csy2, Csy3, Cse1, Cse2, Csc1, Csc2, Csa5, Csn2, Csm2, Csm3, Csm4, Csm5, Csm6, Cmr1, Cmr3, Cmr4, Cmr5, Cmr6, Csb1, Csb2, Csb3, Csx17, Csx14, Csx10, Csx16, CsaX, Csx3, Csx1, Csx15, Csf1, Csf2, Csf3,  
15       Csf4, homologs thereof, or modified versions thereof. These enzymes are known; for example, the amino acid sequence of *S. pyogenes* Cas9 protein (SEQ ID NO. 1) may be found in the SwissProt database under accession number Q99ZW2. In some embodiments, the unmodified Cas protein has DNA cleavage activity. In some embodiments, the Cas protein directs cleavage of one or both strands at the location of a target sequence, such as  
20       within the target sequence and/or within the complement of the target sequence. In some embodiments, the Cas protein directs cleavage of one or both strands within about 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 50, 100, 200, 500, or more base pairs from the first or last nucleotide of a target sequence. In some embodiments, the Cas protein is mutated such that the mutated Cas protein lacks the ability to cleave one or both strands of a target polynucleotide  
25       containing a target sequence. For example, an aspartate-to-alanine substitution (D10A) in the RuvC I catalytic domain of Cas9 from *S. pyogenes* converts Cas9 from a nuclease that cleaves both strands to a nickase (cleaves a single strand). Other examples of mutations that render Cas9 a nickase include, without limitation, H840A, N854A, and N863A. In some  
30       embodiments, the fragment of the Cas protein lacks DNA cleavage activity, e.g., the fragment does not contain the catalytic domain of the Cas protein (e.g., RuvC I, RuvC II, and RuvC III domain of Cas9). In some embodiments, two or more catalytic domains of Cas9 (RuvC I, RuvC II, and RuvC III) may be mutated to produce a mutated Cas9 substantially lacking all DNA cleavage activity. In some embodiments, a D10A mutation is combined with one or more of H840A, N854A, or N863A mutations to produce a Cas9 enzyme substantially

lacking all DNA cleavage activity. In some embodiments, a Cas protein is considered to substantially lack all DNA cleavage activity when the DNA cleavage activity of the mutated enzyme is less than about 25%, 10%, 5%, 1%, 0.1%, 0.01%, or lower with respect to its non-mutated form. Other mutations may be useful; where the Cas9 or other Cas protein is from a species other than *S. pyogenes*, mutations in corresponding amino acids may be made to achieve similar effects.

**[0044]** In some embodiments, the site of recombinase activity is determined by the CRISPR-Cas system guide RNA. Upon binding to the guide RNA, the site-directed modifying polypeptide is directed to the target sequence and exerts a recombinase activity in the presence of a recombination site of the recombinase such as integrase attachment sites, attP and attB. For example, a Cas9 guide RNA can direct a polypeptide comprising a Cas9 and a phiC31 to form a complex at a target sequence. In the presence of its cognate attP and attB sites, phiC31 integrase is activated to mediate a site-specific recombination between a donor DNA and the target sequence, resulting in an integration of a DNA sequence into the target sequence.

**[0045]** In some embodiments, the donor DNA comprises a transgene to be inserted at the target sequence and a wild-type recombination attachment site of the integrase.

**[0046]** A "recombination site" as used herein means a recombination site recognized and used by a recombinase. For example, lambda is a temperate bacteriophage that infects *E. coli*. The phage has one attachment site for recombination (attP) and the *E. coli* bacterial genome has an attachment site for recombination (attB). Both of these sites are wild-type recombination sites for lambda integrase. In the context of the present invention, wild-type recombination sites occur in the homologous phage/bacteria system. Accordingly, wild-type recombination sites can be derived from the homologous system and associated with heterologous sequences, for example, the attB site can be placed in other systems to act as a substrate for the integrase.

**[0047]** In some embodiments, the integrase in the site-directed modifying polypeptide is bacteria phage phiC31. The wild-type attB and attP recognition sites of phage phiC31 are generally about 34 to 40 nucleotides in length (Groth et al. *Proc Natl Acad Sci USA* 97:5995-6000 (2000)). These sites are typically arranged as follows: AttB comprises a first DNA sequence attB5', a core region, and a second DNA sequence attB3' in the relative order from 5' to 3' attB5'-core region-attB3'. AttP comprises a first DNA sequence attP5', a core region, and a second DNA sequence attP3', in the relative order from 5' to 3' attP5' –core region-attP3'. The core region of attP and attB of phiC31 has the sequence 5'-TTG-3'.

[0048] In some embodiments, the composition further comprises a helper DNA comprising a second nucleotide sequence recognized by the site-specific recombinase. The helper DNA is contemplated to facilitate the formation of a complex comprising the guide RNA, one or more site-directed modifying polypeptide at the target DNA sequence, wherein  
5 the complex is functional in mediating a recombination between the donor DNA and the target DNA, resulting in an insertion of a DNA sequence in the donor DNA into the target DNA. In some embodiments, the helper DNA hybridizes with a sequence in the guide RNA. In this scenario, in the presence of the helper DNA, at least two recombinase domains that bind to the second nucleotide sequence at the donor DNA interact with at least two  
10 recombinase domain that bind to the first nucleotide sequence at the help DNA, thus forming a tetramer that functions to facilitate a crossover and strand exchange between the donor DNA and the target DNA.

[0049] In some embodiments, the composition further comprises a recombinase or a fragment thereof. Upon binding to the guide RNA through the guide-RNA-binding domain,  
15 the site-directed modifying polypeptides are recruited to the target sequence. A tetramer of recombinase or recombinase domain is formed at the target sequence and activated by binding to its cognate sites, and mediates the recombination between the donor DNA and the target DNA.

[0050] In some embodiments, the composition further comprises a second CRISPR-Cas system guide RNA that is capable of hybridizing with a second sequence in the target  
20 DNA. The second sequence in the target DNA is separated from the first sequence in the target DNA by about 5, 6, 7, 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, 45, 50 nucleotide.

[0051] In another aspect, the present application discloses a composition comprising  
25 one or more vectors comprising (i) a first nucleotide sequence encoding a polypeptide as disclosed above; and (ii) a second nucleotide sequence encoding a guide RNA.

[0052] As used herein, the term “vector” refers to a polynucleotide molecule that comprises a gene or a nucleic acid sequence of particular interest. Typically, the construct also includes appropriate regulatory sequences. For example, the polynucleotide molecule  
30 can include regulatory sequences located in the 5'-flanking region of the nucleotide sequence encoding the guide RNA and/or the nucleotide sequence encoding a site-directed modifying polypeptide, operably linked to the coding sequences in a manner capable of expressing the desired transcript/gene in a host cell.

[0053] The vector comprising the nucleotide sequence encoding the guide RNA and/or the nucleotide sequence encoding a site-directed modifying polypeptide can be prepared using methods well known in the art. For example, the polynucleotide molecule can be prepared as part of a larger plasmid. Such preparation allows the cloning and isolation of the correct constructs in an efficient manner as is known in the art. The various methods employed in the preparation of the plasmids and transformation of host organisms are known in the art (see, e.g., *Molecular Cloning A Laboratory Manual*, 2nd Ed., ed. By Sambrook, Fritsch and Maniatis, Cold Spring Harbor Laboratory Press, 1989).

[0054] In some embodiments, the one or more vectors further comprise a nucleotide sequence encoding a recombinase. Upon binding to the guide RNA through the guide-RNA-binding domain, the site-directed modifying polypeptides are recruited to the target sequence. A tetramer of recombinase or recombinase domain or recombinase variant activated by binding to its cognate sites, or a tetramer of activated recombinase is formed at the target sequence and activated by binding to its cognate binding sites, and mediates the recombination between the donor DNA and the target DNA.

[0055] In yet another aspect, the present application discloses a method of integrating a DNA sequence into the genome of a cell comprising introducing into the cell a composition comprising one or more vectors comprising: (i) a first nucleotide sequence encoding a first CRISPR-Cas system guide RNA, wherein the first guide RNA is capable of hybridizing with a first sequence in a target DNA; and (ii) a second nucleotide sequence encoding a site-directed modifying polypeptide, wherein the site-directed modifying polypeptide comprising: (a) a guide-RNA-binding portion that is capable of binding to the guide RNA; and (b) a recombinase domain; (iii) a donor DNA comprising: (a) a third nucleotide sequence recognized by the site-specific recombinase; and (b) the transgene.

[0056] In some embodiments, one or more vectors comprising the nucleotide sequence encoding the guide RNA and/or the nucleotide sequence encoding a site-directed modifying polypeptide are introduced into a host cell such that expression of the elements of the composition direct formation of a complex at the target sites. For example, the complex comprises a first CRISPR-Cas system guide RNA that hybridizes with a first sequence in the target DNA, a second CRISPR-Cas system guide RNA that hybridize with a second sequence in the target DNA and at least two site-directed modifying polypeptides comprising a guide-RNA-binding domain and a recombinase domain. Upon binding to the guide RNA through the guide-RNA-binding domain, the site-directed modifying polypeptides are recruited to the target sequence. A tetramer of recombinase domain is formed at the target sequence and



activated by binding to its cognate sites, and mediates the recombination between the donor DNA and the target DNA.

[0057] Conventional viral and non-viral based gene transfer methods can be used to introduce the vectors in mammalian cells, target tissues or one-cell embryos. Such methods can be used to administer nucleic acids encoding components of the composition to cells in culture, or in a host organism. Non-viral vector delivery systems include DNA plasmids, RNA (e.g. a transcript of a vector described herein), naked nucleic acid, and nucleic acid complexed with a delivery vehicle, such as a liposome, protein complexed with a delivery vehicle, such as a liposome. Viral vector delivery systems include DNA and RNA viruses, which have either episomal or integrated genomes after delivery to the cell. For a review of gene therapy procedures, see Anderson, *Science* 256:808-813 (1992); Nabel & Felgner, *TIBTECH* 11:211-217 (1993); Mitani & Caskey, *TIBTECH* 11:162-166 (1993); Dillon, *TIBTECH* 11:167-175 (1993); Miller, *Nature* 357:455-460 (1992); Van Brunt, *Biotechnology* 6(10):1149-1154 (1988); Vigne, *Restorative Neurology and Neuroscience* 8:35-36 (1995); Kremer & Perricaudet, *British Medical Bulletin* 51(1):31-44 (1995); Haddada et al., in *Current Topics in Microbiology and Immunology* Doerfler and Bihm (eds) (1995); and Yu et al., *Gene Therapy* 1:13-26 (1994).

[0058] Methods of non-viral delivery of nucleic acids include lipofection, nucleofection, electroporation, microinjection, biolistics, virosomes, liposomes, immunoliposomes, polycation or lipid:nucleic acid conjugates, naked DNA, artificial virions, and agent-enhanced uptake of DNA. Lipofection is described in e.g., U.S. Pat. Nos. 5,049,386, 4,946,787; and 4,897,355) and lipofection reagents are sold commercially (e.g., Transfectam™ and Lipofectin™). Cationic and neutral lipids that are suitable for efficient receptor-recognition lipofection of polynucleotides include those of Feigner, WO 91/17424; WO 91/16024. Delivery can be to cells (e.g. in vitro or ex vivo administration) or target tissues (e.g. in vivo administration).

[0059] The preparation of lipid:nucleic acid complexes, including targeted liposomes such as immunolipid complexes, is well known to one of skill in the art (see, e.g., Crystal, *Science* 270:404-410 (1995); Blaese et al., *Cancer Gene Ther.* 2:291-297 (1995); Behr et al., *Bioconjugate Chem.* 5:382-389 (1994); Remy et al., *Bioconjugate Chem.* 5:647-654 (1994); Gao et al., *Gene Therapy* 2:710-722 (1995); Ahmad et al., *Cancer Res.* 52:4817-4820 (1992); U.S. Pat. Nos. 4,186,183, 4,217,344, 4,235,871, 4,261,975, 4,485,054, 4,501,728, 4,774,085, 4,837,028, and 4,946,787).

[0060] Microinjection is used to deliver DNA, RNA or peptides into a nucleus and cytoplasm of a one-cell embryo. It is well known to one of skill in the art (see *Manipulating the mouse embryo; A laboratory manual*, fourth edition, 2014).

[0061] The use of RNA or DNA viral based systems for the delivery of nucleic acids take advantage of highly evolved processes for targeting a virus to specific cells in the body and trafficking the viral payload to the nucleus. Viral vectors can be administered directly to patients (in vivo) or they can be used to treat cells in vitro, and the modified cells may optionally be administered to patients (in vivo). Conventional viral based systems could include retroviral, lentiviral, adenoviral, adeno-associated and herpes simplex viral vectors for gene transfer. Integration in the host genome is possible with the retrovirus, lentivirus, and adeno-associated virus gene transfer methods, often resulting in long term expression of the inserted transgene. Additionally, high transduction efficiencies have been observed in many different cell types and target tissues.

[0062] The tropism of a retrovirus can be altered by incorporating foreign envelope proteins, expanding the potential target population of target cells. Lentiviral vectors are retroviral vectors that are able to transduce or infect non-dividing cells and typically produce high viral titers. Selection of a retroviral gene transfer system would therefore depend on the target tissue. Retroviral vectors are comprised of cis-acting long terminal repeats with packaging capacity for up to 6-10 kb of foreign sequence. The minimum cis-acting LTRs are sufficient for replication and packaging of the vectors, which are then used to integrate the therapeutic gene into the target cell to provide permanent transgene expression. Widely used retroviral vectors include those based upon murine leukemia virus (MuLV), gibbon ape leukemia virus (GaLV), Simian Immuno deficiency virus (SIV), human immuno deficiency virus (HIV), and combinations thereof (see, e.g., Buchscher et al., *J. Virol.* 66:2731-2739 (1992); Johann et al., *J. Virol.* 66:1635-1640 (1992); Sommerfelt et al., *Virol.* 176:58-59 (1990); Wilson et al., *J. Virol.* 63:2374-2378 (1989); Miller et al., *J. Virol.* 65:2220-2224 (1991); PCT/US94/05700). In applications where transient expression is preferred, adenoviral based systems may be used. Adenoviral based vectors are capable of very high transduction efficiency in many cell types and do not require cell division. With such vectors, high titer and levels of expression have been obtained. This vector can be produced in large quantities in a relatively simple system. Adeno-associated virus (“AAV”) vectors may also be used to transduce cells with target nucleic acids, e.g., in the in vitro production of nucleic acids and peptides, and for in vivo and ex vivo gene therapy procedures (see, e.g., West et al., *Virology* 160:38-47 (1987); U.S. Pat. No. 4,797,368; WO 93/24641; Kotin, *Human Gene Therapy*

5:793-801 (1994); Muzyczka, J. Clin. Invest. 94:1351 (1994). Construction of recombinant AAV vectors are described in a number of publications, including U.S. Pat. No. 5,173,414; Tratschin et al., Mol. Cell. Biol. 5:3251-3260 (1985); Tratschin, et al., Mol. Cell. Biol. 4:2072-2081 (1984); Hermonat & Muzyczka, PNAS 81:6466-6470 (1984); and Samulski et al., J. Virol. 63:03822-3828 (1989).

**[0063]** Packaging cells are typically used to form virus particles that are capable of infecting a host cell. Such cells include 293 cells, which package adenovirus, and  $\psi$ 2 cells or PA317 cells, which package retrovirus. Viral vectors used in gene therapy are usually generated by producing a cell line that packages a nucleic acid vector into a viral particle.

10 The vectors typically contain the minimal viral sequences required for packaging and subsequent integration into a host, other viral sequences being replaced by an expression cassette for the polynucleotide(s) to be expressed. The missing viral functions are typically supplied in trans by the packaging cell line. For example, AAV vectors used in gene therapy typically only possess ITR sequences from the AAV genome which are required for  
15 packaging and integration into the host genome. Viral DNA is packaged in a cell line, which contains a helper plasmid encoding the other AAV genes, namely rep and cap, but lacking ITR sequences. The cell line may also be infected with adenovirus as a helper. The helper virus promotes replication of the AAV vector and expression of AAV genes from the helper plasmid. The helper plasmid is not packaged in significant amounts due to a lack of ITR  
20 sequences. Contamination with adenovirus can be reduced by, e.g., heat treatment to which adenovirus is more sensitive than AAV. Additional methods for the delivery of nucleic acids to cells are known to those skilled in the art. See, for example, US20030087817, incorporated herein by reference.

**[0064]** In some embodiments, a host cell is transiently or non-transiently transfected  
25 with one or more vectors described herein. In some embodiments, a cell is transfected as it naturally occurs in a subject. In some embodiments, a cell that is transfected is taken from a subject such as a primary cell or a stem cell. In some embodiments, the cell is derived from cells taken from a subject, such as a cell line. A wide variety of cell lines for tissue culture are known in the art. Examples of cell lines include, but are not limited to, C8161, CCRF-CEM,  
30 MOLT, mIMCD-3, NHDF, HeLa-S3, Huh1, Huh4, Huh7, HUVEC, HASMC, HEK<sub>n</sub>, HEK<sub>a</sub>, MiaPaCell, Panel, PC-3, TF1, CTLL-2, C1R, Rat6, CV1, RPTE, A10, T24, J82, A375, ARH-77, Calu1, SW480, SW620, SKOV3, SK-UT, CaCo2, P388D1, SEM-K2, WEHI-231, HB56, TIB55, Jurkat, J45.01, LRMB, Bcl-1, BC-3, IC21, DLD2, Raw264.7, NRK, NRK-52E, MRC5, MEF, Hep G2, HeLa B, HeLa T4, COS, COS-1, COS-6, COS-M6A, BS-C-1 monkey

kidney epithelial, BALB/3T3 mouse embryo fibroblast, 3T3 Swiss, 3T3-L1, 132-d5 human fetal fibroblasts; 10.1 mouse fibroblasts, 293-T, 3T3, 721, 9L, A2780, A2780ADR, A2780cis, A 172, A20, A253, A431, A-549, ALC, B16, B35, BCP-1 cells, BEAS-2B, bEnd.3, BHK-21, BR 293. BxPC3. C3H-10T1/2, C6/36, Cal-27, CHO, CHO-7, CHO-IR, CHO-K1, CHO-K2, 5 CHO-T, CHO Dhfr  $-/-$ , COR-L23, COR-L23/CPR, COR-L23/5010, COR-L23/R23, COS-7, COV-434, CML T1, CMT, CT26, D17, DH82, DU145, DuCaP, EL4, EM2, EM3, EMT6/AR1, EMT6/AR10.0, FM3, H1299, H69, HB54, HB55, HCA2, HEK-293, HeLa, Hepa1c1c7, HL-60, HMEC, HT-29, Jurkat, JY cells, K562 cells, Ku812, KCL22, KG1, KYO1, LNCap, Ma-Mel 1-48, MC-38, MCF-7, MCF-10A, MDA-MB-231, MDA-MB-468, 10 MDA-MB-435, MDCK II, MDCK 11, MOR/0.2R, MONO-MAC 6, MTD-1A, MyEnd, NCI-H69/CPR, NCI-H69/LX10, NCI-H69/LX20, NCI-H69/LX4, NIH-3T3, NALM-1, NW-145, OPCN/OPCT cell lines, Peer, PNT-1A/PNT 2, RenCa, RIN-5F, RMA/RMAS, Saos-2 cells, Sf-9, SkBr3, T2, T-47D, T84, THP1 cell line, U373, U87, U937, VCaP, Vero cells, WM39, WT-49, X63, YAC-1, YAR, and transgenic varieties thereof. Cell lines are available from a 15 variety of sources known to those with skill in the art (see, e.g., the American Type Culture Collection (ATCC) (Manassus, Va.)). In some embodiments, a cell transfected with one or more vectors described herein is used to establish a new cell line comprising one or more vector-derived sequences. In some embodiments, a cell transiently transfected with the components of the composition as described herein (such as by transient transfection of one 20 or more vectors, or transfection with RNA, or transfection with protein), and modified through the activity of the complex, is used to establish a new cell line comprising cells containing the modification but lacking any other exogenous sequence. In some embodiments, cells transiently or non-transiently transfected with one or more vectors described herein, or cell lines derived from such cells are used in assessing one or more test compounds.

25

### Example 1

[0065] This example illustrates an active CasR mutant including variant of chimeric Cas9 protein and a catalytic domain of active mutant of Tn3 resolvase (CasTn3). CasTn3's capability of mediating recombination is tested in a recombination assay using a substrate plasmid containing a sequence, Tn3 res site I, recognized and bound by CasTn3. The 30 potential target sequence (C-site) comprises 22 bp of Tn3 res site I flanked by 17 bp motifs recognized by dCas9 domain via suitable gRNA. The substrate plasmid is then constructed in which two identical C-sites are separated by a galK marker gene. If CasTn3-mediated recombination in E. coli cells deletes galK, the colonies on MacConkey-galactose indicator plates are pale, whereas galK<sup>+</sup> colonies are red. .

[0066] Expression plasmids for ‘progenitor’ CasTn3 are generated as follows. A resolvase expression plasmid with a p15a origin of replication (pEK76) is derived from pACYC184 (Chang AC, Cohen SN (1978) Construction and characterization of amplifiable multicopy DNA cloning vehicles derived from the P15A cryptic miniplasmid. *J Bacteriol* 134: 1141–1156.), by insertion of an SspI-EcoRV resolvase-encoding fragment from pAT5 (Arnold PH, Blake DG, Grindley NDF, Boocock MR, Stark WM (1999) Mutants of Tn3 resolvase which do not require accessory binding sites for recombination activity. *EMBO J* 18: 1407–1414.) into the pACYC184 AvaI site. CasTn3 expression plasmids (pCasTn3) are then created by insertion of ORF-containing fragments of dCas9 between unique NdeI and Asp718 sites in pEK76.

[0067] The ‘progenitor’ CasTn3 used here consists of the Tn3 resolvase catalytic domain (residues 1–148) containing the seven mutations R2A E56K G101S D102Y M103I Q105L V107F, followed by the 2-amino acid linker TS, then the dCas9. The codons for the TS linker introduce a unique SpeI restriction site.

[0068] The C-site, which is designed to be recognized by CasTn3, consists of 22bp of Tn3 res site I flanked by 20 bp motifs recognized by dCas9. The sequences of the C-sites used here (C1tn3C1) are shown in FIG. 5. Recombination substrate plasmid, with two C-sites flanking a galK marker gene, is shown in FIG. 3.

[0069] *E. coli* colony color assays of recombination are then used to test the capability of CasTn3 to mediate recombination of C-sites. The *E. coli* colony color assay of recombination has been described in Akopian A, He J, Boocock MR, Stark WM (2003) Chimeric recombinases with designed DNA sequence recognition. *Proc Natl Acad Sci USA* 100: 8688–8691. Briefly, strains containing a substrate plasmid are co-transformed with a CasTn3 expression plasmid and a gRNA expression plasmid. Transformants are selected on MacConkey-galactose indicator plates (MacConkey agar base (Difco) supplemented with 1% galactose, kanamycin (to select for the substrate plasmid or its resolution product), and ampicillin and/or chloramphenicol (to select for pCasTn3 expression plasmids). Pale coloured (galK<sup>-</sup>) colonies indicate recombination (resolution) proficiency, whereas red (galK<sup>+</sup>) colonies indicate lack of resolution.

[0070] To analyze the DNA products of recombination, strains containing a substrate plasmid are co-transformed with a CasTn3 expression plasmid and a gRNA expression plasmid. Transformants are grown with appropriate antibiotic selection on L-agar plates. Following incubation at 37°C, cells are washed from the plates with L-broth, and plasmid DNA is purified from the cells using a Qiagen miniprep kit. The DNA is visualized by

ethidium staining after 1.2% agarose gel electrophoresis. To quantify the extent of recombination (leading to deletion of the galK gene) more accurately, the recovered plasmid DNA is used to transform the strain DS941, test plasmid-containing transformants are selected on MacConkey agar plates with kanamycin, and the percentage of pale (galK<sup>-</sup>) colonies (from a total of >100) is determined.

[0071] The results show that C1Tn3C1 is resolved by CasTn3.

### Example 2

[0072] This example demonstrates that CasTn3 can be used to target asymmetric C-sites. The substrate plasmid used here is similar to that in Example 1 except that it contains a C-site consisting of 22bp of Tn3 res site I flanked by two different 17 bp motifs recognized by dCas9. The sequence of the C-site used here (C1tn3C2) is shown in FIG. 4.

[0073] To test CasTn3's capability of recombination targeting asymmetric C-site, strains containing the substrate plasmid are co-transformed with a CasTn3 expression plasmid and two gRNA expression plasmids that express gRNA recognizing C1 and C2 respectively. Transformants are selected on MacConkey-galactose indicator plates (MacConkey agar base (Difco) supplemented with 1% galactose, kanamycin (to select for the substrate plasmid or its resolution product), and ampicillin and/or chloramphenicol (to select for pCasTn3 expression plasmids and gRNA expression plasmids, respectively). Pale coloured (galK<sup>-</sup>) colonies indicate recombination (resolution) proficiency, whereas red (galK<sup>+</sup>) colonies indicate lack of resolution.

[0074] The DNA products of recombination are analyzed as in Example 1.

[0075] The result shows that CasTn3 can mediate recombination at asymmetric C-sites.

### Example 3

[0076] This example illustrates identification of CasTn3 variants that would promote recombination at potential target sequence similar but not identical to Tn3 res site I.

[0077] In general, a series of potential target sequences that are similar to Tn3 res site I, and which might be optimal for resolvase activity are used (shown in FIG. 4). These potential target sequences are over 12 bp centered on the TATA motif and include at least 8 out of 12 base pairs identical to one orientation. "C1resC1" site are made, each comprising 22 bp of a potential Tn3 res site flanked by 17 bp motifs recognized by dCas9 domain via corresponding gRNA.

[0078] To find CasTn3 variants that can promote recombination at the potential C1resC1 sites, the entire catalytic domain of Tn3 is subject to random mutagenesis using two

error-prone PCR protocols, and made libraries of plasmids expressing mutant CasTn3s are made (see Burke ME, Arnold PH, He J, Wenwieser SVCT, Rowland SJ, et al. (2004) Activating mutations of Tn3 resolvase marking interfaces important in recombination catalysis and its regulation. Mol Microbiol 51: 937–948.). To screen for recombination-  
 5 proficient CasTn3 mutants, DS941 cells containing a recombination substrate plasmid are transformed with a library of mutant expression plasmids and gRNA expression plasmids.

Aliquots of the transformants are selected on MacConkey plates (aiming for ~1000 colonies

per plate; 30–60 plates). Pale-colored colonies are picked and are streaked on MacConkey plates to confirm the colony color. Plasmid DNA purified from positive isolates is used to  
 10 transform the same DS941-substrate plasmid strain. If the transformant colonies on MacConkey plates are all pale-colored, the CasTn3 is deemed to be a recombination-proficient mutant, so the expression plasmid is isolated and the CasTn3 reading frame is sequenced.

**[0079]** Screens of the libraries using C1resC1 substrate plasmids yields several active  
 15 mutants as shown in Table 1 below. Recovery of DNA from the cells show that one mutant with the single conservative change I77L has significant recombination activity on C1resC1 substrates.

**[0080]** The mutagenesis-screening procedure can be repeated, using newly identified CasTn3 mutant as the template for mutagenesis, and other C1resC1 test substrate. Several  
 20 active mutants are isolated which are shown to promote efficient recombination of the C1resC1 sites. The mutations thus broaden rather than switch specificity.

**[0081]** Table 1. Active mutants that promote recombination at potential target sequence.

Mutant	C-site						
	C1tn3C1	C1res1C1	C1res2C1	C1res3C1	C1res4C1	C1res5C1	C1res6C1
CasTn3							
I77L	W	W	R	R	R	R	R
F83L F107L	W	W	W	R	W	W	R
I3T D95E	W	W	R	R	W	R	R
G70S F107L	W	W	W	R	W	W	R
E57G A89V	W	W	W	R	W	W	R

K37R D44N	W	W	W	R	W	W	R
I97V I103T F107L	W	W	W	R	W	W	W
G70S	W	W	R	R	W	R	R
F107L	W	W	R	R	W	R	R
E57G	W	W	R	R	R	R	R
A89V	W	W	R	R	W	R	R
I77L I3P V108A	W	W	W	W	W	W	W
I77L L135R	W	W	W	W	W	W	W
I77L S12R I103V	W	W	W	W	W	W	W
I77L E132A	W	W	W	W	W	W	W
I77L I3L L135R	W	W	W	W	W	W	W
I77L I3S E132A	W	W	W	W	W	W	W
I77L N127H E132A	W	W	W	W	W	W	W

**[0082]** The Table shows mutants isolated from libraries of mutagenized CasTn3. The left-hand column gives the mutants, and the other columns show the phenotype of each mutant in the MacConkey agar colony assay, using substrates with two identical CresC sites as indicated. W, 'white' (pale-colored) colonies; R, red colonies.

5

#### Example 4

**[0083]** This example illustrates CasR-mediated recombination between non-identical sites.

**[0084]** For transgene integration at any chosen genomic locus, only one of the recombination sites (the genomic one) must be targeted by 'designer' CasRs. The other site (on the transgenic DNA) could be acted on by any recombinase with a compatible catalytic domain, and could be optimized for integration efficiency and specificity. To model this scenario, substrates with non-identical pairs of sites are generated. The C1tn3C1 site contains at its center 22 bp of the natural target sequence for CasTn3 catalytic domains (Tn3 res site I), and a substrate with two of these sites is therefore recombined efficiently. To analyze recombination between C1tn3C1 and other non-identical sites, a set of six C1tn3C1x C1res1C1 are made, in which res1 represents variants of the 22bp center of the natural target sequence for CasTn3. These plasmids are recombined efficiently.

**[0085]** In order to show that the partner site need not be a C-site, substrates which contained a C1res1C1 site paired with Tn3 res site I is made. Recombination of C1resC1×site I substrates is efficient when both an activated Tn3 resolvase variant (NM resolvase; Olorunniyi FJ, He J, Wenwieser SVCT, Boocock MR, Stark WM (2008) Synapsis



and catalysis by activated Tn3 resolvase mutants. Nucleic Acids Res 36: 7181–7191.) and a CasTn3-1 are co-expressed, whereas NM resolvase alone or CasTn3-1 alone do not promote recombination.

5 [0086] Finally, a substrate is designed to model transgene integration as outlined above, where one "genomic" recombination site is targeted by the combined action of two gRNA, and the other "transgene" site is recognized by an activated recombinase. For the "genomic" site we chose to use C1res1C2. For the "transgene" site Tn3 res site I is used, as in the experiments described above. The results show that C1res1C2×site I recombination requires the expression of NM resolvase, CasTn3 and two gRNA.

10

## WHAT IS CLAIMED IS:

- 1) A polypeptide comprising:
  - (i) a CRISPR-Cas domain that is capable of binding to a guide RNA, wherein the guide RNA is capable of hybridizing to a target DNA sequence; and
  - 5 (ii) a recombinase domain comprising a site-specific recombinase, a mutant thereof or a fragment thereof.
- 2) The polypeptide of claim 1, wherein the CRISPR-Cas domain comprises a Type-II Cas9 protein or a fragment thereof.
- 3) The polypeptide of claim 1, wherein the CRISPR-Cas domain does not have  
10 endonuclease activity.
- 4) The polypeptide of claim 1, wherein the CRISPR-Cas domain is a dead Cas9 (dCas9).
- 5) The polypeptide of claim 1, wherein the CRISPR-Cas domain comprises an amino acid sequence of SEQ ID 1.
- 6) The polypeptide of claim 1, wherein the site-specific recombinase is a tyrosine  
15 recombinase.
- 7) The polypeptide of claim 6, wherein the tyrosine recombinase is selected from the group consisting of Cre, Flp and the lambda integrase.
- 8) The polypeptide of claim 1, wherein the site-specific recombinase is a serine recombinase.
- 20 9) The polypeptide of claim 8, wherein the serine recombinase is selected from the group consisting of gamma-delta resolvase, Tn3 resolvase, Sin resolvase, Gin invertase, Hin invertase, Tn5044 resolvase, IS607 transposase and integrase
- 10) The polypeptide of claim 9, wherein the integrase is selected from Bxb1, wBeta, BL3 phiR4, A118, TG1, MR11, phi370, SPBc, TP901-1, phiRV, FC1, K38, phiBT1 and phiC31.
- 25 11) The polypeptide of claim 1, wherein the recombinase domain is a catalytic domain of a recombinase or a mutant thereof.
- 12) The polypeptide of claim 1, wherein the recombinase domain is a catalytic domain of an activated mutant of a recombinase.
- 13) The polypeptide of claim 12, wherein the recombinase is Tn3 (SEQ ID NO.: 3).
- 30 14) The polypeptide of claim 13, wherein the recombinase domain has an amino acid sequence selected from the group consisting of SEQ ID NO: 5 and the variants thereof listed in Table 1.

- 15) The polypeptide of claim 1, further comprising a linker that links the CRISPR-Cas domain and the recombinase domain.
- 16) The polypeptide of claim 16, wherein the linker has an amino acid sequence of The-Ser.
- 5 17) A composition comprising the polypeptide of claim 1, and the guide RNA.
- 18) The composition of claim 17, further comprising a donor DNA comprising a first nucleotide sequence recognized by a site-specific recombinase and a transgene.
- 19) A composition comprising:
- 10 (i) a first nucleotide sequence encoding the polypeptide of claim 1; and
- (ii) a second nucleotide sequence encoding the guide RNA;
- wherein the first and second nucleotide are located on same or different vectors.
- 20) The composition of claim 19, further comprising a donor DNA comprising a first nucleotide sequence recognized by a site-specific recombinase and a transgene.
- 21) The composition of claim 19, further comprising a third nucleotide sequence encoding
- 15 the site-specific recombinase.
- 22) A method of integrating a transgene into the genome of a cell comprising introducing into the cell the composition of claim 19.
- 23) The method of claim 22, wherein the composition further comprises a third nucleotide sequence encoding the site-specific recombinase
- 20 24) The method of claim 22, wherein the cell is a eukaryotic cell.
- 25) The method of claim 22, wherein the cell is a mammalian or human cell.
- 26) The method of claim 22, wherein the cell is a one-cell embryo.

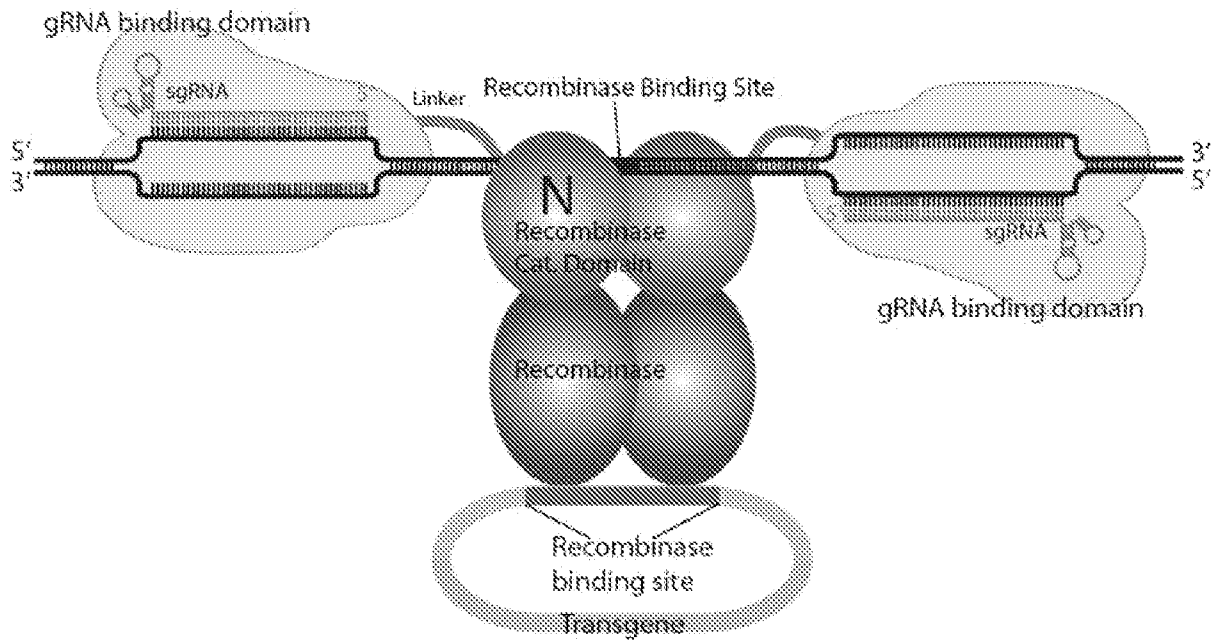


FIG. 1

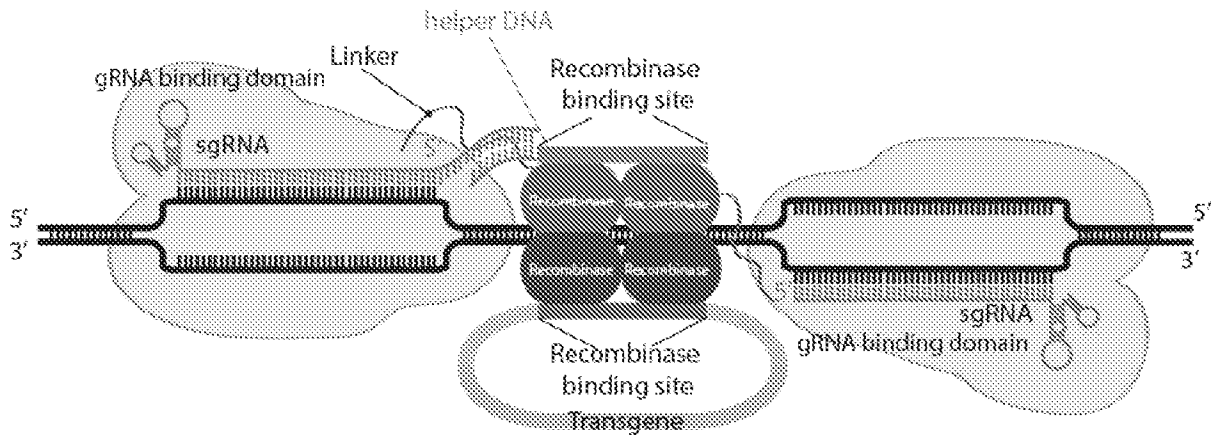


FIG. 2A

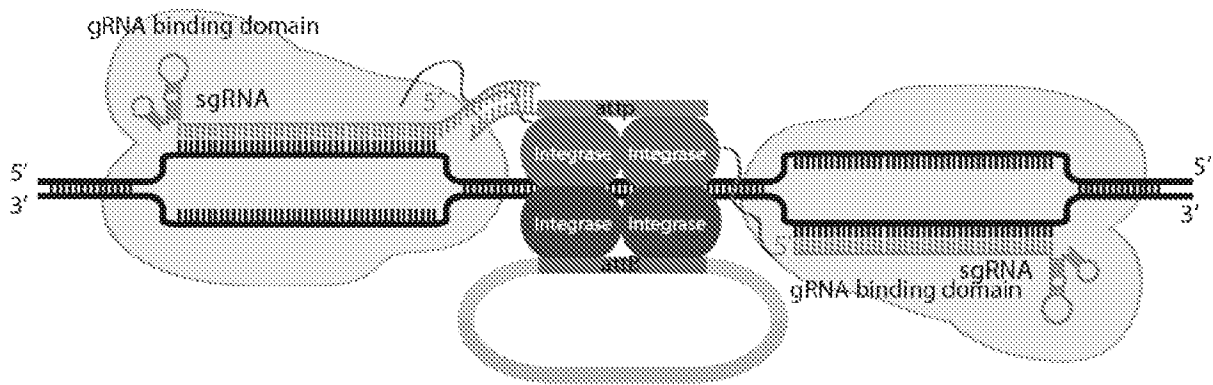


FIG. 2B

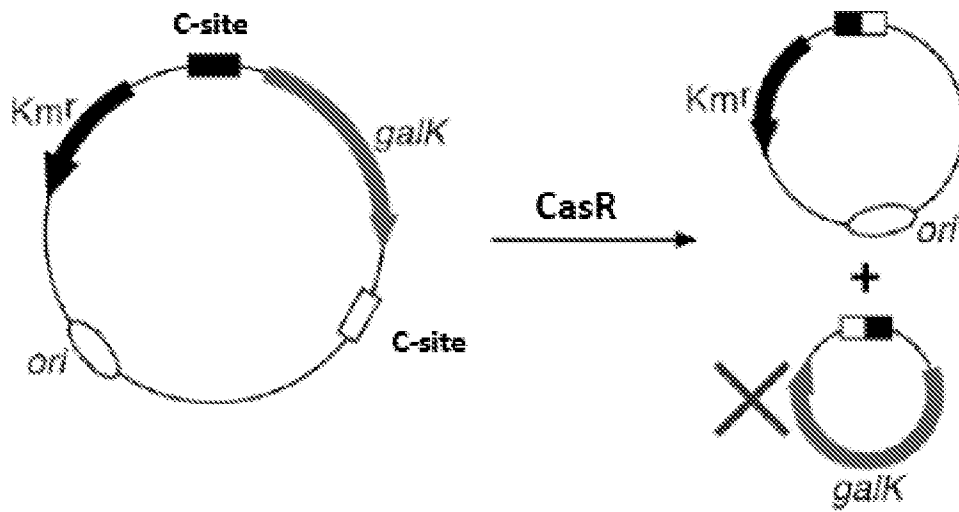


FIG. 3

Tn3 site I -----CGTTCGAAATATTA | TAAATTATCAGACA-----  
 res1 -----AATGCTAAATTA | TAAAATAAACA-----  
 res2 -----TTTTCAGACTA | TAAAATGCACA-----  
 res3 -----TATTTGTATTA | TACACTATCTT-----  
 res4 -----CATTTAAATTTA | TACTATGGGAA-----  
 res5 -----AAAGGTATTTTA | TAAAAGAAGAG-----  
 res6 -----AGCTCAAAAGTA | TAGTAGTCAGT-----

**FIG. 4**

C1tn3C1  
 atagccttgtggctaataaccaggTCGAAATATTA | TAAATTATCAGcctggattagccacaaggctat  
 C1res1C1  
 atagccttgtggctaataaccaggAATGCTAAATTA | TAAAATAAACAcctggattagccacaaggctat  
 C1res2C1  
 atagccttgtggctaataaccaggTTTTCAGACTA | TAAAATGCACAcctggattagccacaaggctat  
 C1res3C1  
 atagccttgtggctaataaccaggTATTTGTATTA | TACACTATCTTcctggattagccacaaggctat  
 C1res4C1  
 atagccttgtggctaataaccaggCATTTAAATTTA | TACTATGGGAAcctggattagccacaaggctat  
 C1res5C1  
 atagccttgtggctaataaccaggAAAGGTATTTTA | TAAAAGAAGAGcctggattagccacaaggctat  
 C1res6C1  
 atagccttgtggctaataaccaggAGCTCAAAAGTA | TAGTAGTCAGTcctggattagccacaaggctat  
 C1tn3C2  
 atagccttgtggctaataaccaggTCGAAATATTA | TAAATTATCAGccccaacaggtcagtttatactg  
 C1res1C2  
 atagccttgtggctaataaccaggAATGCTAAATTA | TAAAATAAACAccccaacaggtcagtttatactg

**FIG. 5**

**A. CLASSIFICATION OF SUBJECT MATTER**

C12N 9/22(2006.01)i, C12N 15/113(2010.01)i, C12N 15/52(2006.01)i, C12N 15/85(2006.01)i

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)

C12N 9/22; C12N 15/63; C12N 15/90; C12N 15/10; C12N 9/12; C12N 15/113; C12N 15/52; C12N 15/85

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

Korean utility models and applications for utility models

Japanese utility models and applications for utility models

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

eKOMPASS(KIPO internal) &amp; keywords: CRISPR-Cas, guide RNA, recombinase, domain, dCas9, integrase, hybridizing

**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2014-0357530 A1 (THE BROAD INSTITUTE INC. et al.) 4 December 2014 See paragraphs [0016], [0023], [0024], [0302], [0496], [0512] and [0536]; table 1; claims 1-6 and 19-24; figures 1-2B.	1-4, 6-12, 15-26
Y		5, 13
A		14
Y	NCBI, PDB: 4CMP_A, `Chain A, Crystal Structure of S. Pyogenes Cas9`, 26 March 2014 See the whole document.	5
Y	NCBI, Reference Sequence: WP_013279315.1, `Tn3 resolvase [Escherichia coli]`, 18 May 2013 See the whole document.	13
A	US 2014-0295556 A1 (THE GENERAL HOSPITAL CORPORATION) 2 October 2014 See claims 1-21; figure 1.	1-26
PX	US 2015-0071898 A1 (PRESIDENT AND FELLOWS OF HARVARD COLLEGE) 12 March 2015 See paragraphs [0007], [0025], [0026], [0042] and [0061]; claims 1-27; figures 1A-6B.	1-4, 6-12, 15-26

 Further documents are listed in the continuation of Box C. 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

"E" 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

"&amp;" document member of the same patent family

Date of the actual completion of the international search

22 March 2016 (22.03.2016)

Date of mailing of the international search report

**22 March 2016 (22.03.2016)**

Name and mailing address of the ISA/KR

International Application Division  
Korean Intellectual Property Office  
189 Cheongsa-ro, Seo-gu, Daejeon, 35208, Republic of Korea

Facsimile No. +82-42-481-8578

Authorized officer

KIM, Seung Beom

Telephone No. +82-42-481-3371





**INTERNATIONAL SEARCH REPORT**

Information on patent family members

International application No.

**PCT/US2015/064352**

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2014-0357530 A1	04/12/2014	EP 2931899 A1 WO 2014-093701 A1	21/10/2015 19/06/2014
US 2014-0295556 A1	02/10/2014	AU 2014-227653 A1 AU 2014-228981 A1 AU 2014-239665 A1 CA 2906553 A1 CA 2906724 A1 CA 2907198 A1 US 2014-0295557 A1 WO 2014-144288 A1 WO 2014-144592 A2 WO 2014-144592 A3 WO 2014-144761 A2 WO 2014-144761 A3 WO 2014-152432 A2 WO 2014-152432 A3	01/10/2015 01/10/2015 01/10/2015 25/09/2014 18/09/2014 18/09/2014 02/10/2014 18/09/2014 18/09/2014 31/12/2014 18/09/2014 29/10/2015 25/09/2014 29/10/2015
US 2015-0071898 A1	12/03/2015	US 2015-0071899 A1 WO 2015-035162 A2 WO 2015-035162 A3	12/03/2015 12/03/2015 04/06/2015