



(11)

EP 3 486 318 B1

(12)

EUROPEAN PATENT SPECIFICATION

(45) Date of publication and mention
of the grant of the patent:
06.10.2021 Bulletin 2021/40

(51) Int Cl.:
C12N 15/113 (2010.01) **C12N 9/22** (2006.01)
C12Q 1/6869 (2018.01) **C12Q 1/6816** (2018.01)

(21) Application number: **18199195.1**(22) Date of filing: **08.08.2014**

(54) NUCLEASE PROFILING SYSTEM

NUKLEASEPROFILIERUNGSSYSTEM
SYSTÈME DE PROFILAGE DE NUCLÉASE

(84) Designated Contracting States:
**AL AT BE BG CH CY CZ DE DK EE ES FI FR GB
GR HR HU IE IS IT LI LT LU LV MC MK MT NL NO
PL PT RO RS SE SI SK SM TR**

(30) Priority: **09.08.2013 US 201361864289 P
30.06.2014 US 201414320370
30.06.2014 US 201414320413**

(43) Date of publication of application:
22.05.2019 Bulletin 2019/21

(62) Document number(s) of the earlier application(s) in
accordance with Art. 76 EPC:
14755947.0 / 3 030 650

(73) Proprietor: **President and Fellows of Harvard
College
Cambridge, MA 02138 (US)**

(72) Inventors:

- **LIU, David, R.
Lexington, MA 02420 (US)**
- **PATTANAYAK, Vikram
Cambridge, MA 02138 (US)**

(74) Representative: **Atkins, James Gordon John et al
Kilburn & Strode LLP
Lacon London
84 Theobalds Road
London WC1X 8NL (GB)**

(56) References cited:
WO-A2-2013/066438

- **PRASHANT MALI ET AL: "CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering", NATURE BIOTECHNOLOGY, vol. 31, no. 9, 1 August 2013 (2013-08-01), pages 833-838, XP055149275, ISSN: 1087-0156, DOI: 10.1038/nbt.2675**
- **PATRICK D HSU ET AL: "DNA targeting specificity of RNA-guided Cas9 nucleases", NATURE BIOTECHNOLOGY, NATURE PUBLISHING GROUP, UNITED STATES, vol. 31, no. 9, 21 July 2013 (2013-07-21), pages 827-832, XP002718604, ISSN: 1546-1696, DOI: 10.1038/NBT.2647 [retrieved on 2013-07-21]**
- **VIKRAM PATTANAYAK ET AL: "High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity", NATURE BIOTECHNOLOGY, vol. 31, no. 9, 11 August 2013 (2013-08-11), pages 839-843, XP055148795, ISSN: 1087-0156, DOI: 10.1038/nbt.2673**
- **JOHN P GUILINGER ET AL: "Broad specificity profiling of TALENs results in engineered nucleases with improved DNA-cleavage specificity", NATURE METHODS, vol. 11, no. 4, 16 February 2014 (2014-02-16), pages 429-435, XP055148794, ISSN: 1548-7091, DOI: 10.1038/nmeth.2845**

Note: Within nine months of the publication of the mention of the grant of the European patent in the European Patent Bulletin, any person may give notice to the European Patent Office of opposition to that patent, in accordance with the Implementing Regulations. Notice of opposition shall not be deemed to have been filed until the opposition fee has been paid. (Art. 99(1) European Patent Convention).

Description**BACKGROUND OF THE INVENTION**

5 [0001] Site-specific endonucleases theoretically allow for the targeted manipulation of a single site within a genome and are useful in the context of gene targeting as well as for therapeutic applications. In a variety of organisms, including mammals, site-specific endonucleases have been used for genome engineering by stimulating either non-homologous end joining or homologous recombination. In addition to providing powerful research tools, site-specific nucleases also have potential as gene therapy agents, and two site-specific endonucleases have recently entered clinical trials: one, CCR5-2246, targeting a human CCR-5 allele as part of an anti-HIV therapeutic approach (NCT00842634, NCT01044654, NCT01252641), and the other one, VF24684, targeting the human VEGF-A promoter as part of an anti-cancer therapeutic approach (NCT01082926).

10 [0002] Specific cleavage of the intended nuclease target site without or with only minimal off-target activity is a pre-requisite for clinical applications of site-specific endonuclease, and also for high-efficiency genomic manipulations in basic research applications, as imperfect specificity of engineered site-specific binding domains has been linked to cellular toxicity and undesired alterations of genomic loci other than the intended target. Most nucleases available today, however, exhibit significant off-target activity, and thus may not be suitable for clinical applications. WO2013/066438A2 relates to evaluation and improvement of nuclease cleavage specificity.

15 [0003] Technology for evaluating nuclease specificity and for engineering nucleases with improved specificity are therefore needed.

SUMMARY OF THE INVENTION

20 [0004] The invention is defined in the appended claims. Some aspects of this disclosure are based on the recognition that the reported toxicity of some engineered site-specific endonucleases is based on off-target DNA cleavage, rather than on off-target binding alone. Some aspects of this disclosure provide strategies, compositions, systems, and methods to evaluate and characterize the sequence specificity of site-specific nucleases, for example, RNA-programmable endonucleases, such as Cas9 endonucleases, zinc finger nucleases (ZNFs), homing endonucleases, or transcriptional activator-like element nucleases (TALENs).

25 [0005] The strategies, methods, and reagents of the present disclosure represent, in some aspects, an improvement over previous methods for assaying nuclease specificity. For example, some previously reported methods for determining nuclease target site specificity profiles by screening libraries of nucleic acid molecules comprising candidate target sites relied on a "two-cut" *in vitro* selection method which requires indirect reconstruction of target sites from sequences of two half-sites resulting from two adjacent cuts of the nuclease of a library member nucleic acid (see e.g., PCT Application 30 WO 2013/066438; and Pattanayak, V., Ramirez, C.L., Joung, J.K. & Liu, D.R. Revealing off-target cleavage specificities of zinc-finger nucleases by *in vitro* selection. *Nature methods* 8, 765-770 (2011).). In contrast to such "two-cut" strategies, the methods of the present disclosure utilize an optimized "one cut" screening strategy, which allows for the identification 35 of library members that have been cut at least once by the nuclease. As explained in more detail elsewhere herein, the "one-cut" selection strategies provided herein are compatible with single end high-throughput sequencing methods and do not require computational reconstruction of cleaved target sites from cut half-sites, thus streamlining the nuclease 40 profiling process.

[0006] Some aspects of this disclosure provide *in vitro* selection methods for evaluating the cleavage specificity of endonucleases and for selecting nucleases with a desired level of specificity. Such methods are useful, for example, for characterizing an endonuclease of interest and for identifying a nuclease exhibiting a desired level of specificity, for 45 example, for identifying a highly specific endonuclease for clinical applications.

[0007] Some aspects of this disclosure provide methods of identifying suitable nuclease target sites that are sufficiently different from any other site within a genome to achieve specific cleavage by a given nuclease without any or at least minimal off-target cleavage. Such methods are useful for identifying candidate nuclease target sites that can be cleaved with high specificity on a genomic background, for example, when choosing a target site for genomic manipulation *in vitro* or *in vivo*.

[0008] Some aspects of this disclosure provide methods of evaluating, selecting, and/or designing site-specific nucleases with enhanced specificity as compared to current nucleases. For example, provided herein are methods that are useful for selecting and/or designing site-specific nucleases with minimal off-target cleavage activity, for example, by designing variant nucleases with binding domains having decreased binding affinity, by lowering the final concentration 55 of the nuclease, by choosing target sites that differ by at least three base pairs from their closest sequence relatives in the genome, and, in the case of RNA-programmable nucleases, by selecting a guide RNA that results in the fewest off-target sites being bound and/or cut.

[0009] Compositions and kits useful in the practice of the methods described herein are also provided.

[0010] In a first aspect, the invention provides a method for identifying a target site of a nuclease the method comprising (a) providing a nuclease that cuts a double-stranded nucleic acid target site, wherein cutting of the target site results in cut nucleic acid strands comprising a 5' phosphate moiety; (b) contacting the nuclease of (a) with a library of candidate nucleic acid molecules, wherein each nucleic acid molecule comprises a concatemer of a sequence comprising a candidate nuclease target site and a constant insert sequence of at least 10 and not more than 80 nucleotides, under conditions suitable for the nuclease to cut a candidate nucleic acid molecule comprising a target site of the nuclease; and (c) identifying nuclease target sites cut by the nuclease in (b) by determining the sequence of an uncut nuclease target site on the nucleic acid strand that was cut by the nuclease in step (b), wherein the determining of step (c) comprises (i) ligating a first nucleic acid adapter to the 5' end of a nucleic acid strand that was cut by the nuclease in step (b) via 5'-phosphate-dependent ligation, and (ii) amplifying a fragment of the concatemer cut by the nuclease that comprises an uncut target site via a PCR reaction using a PCR primer that hybridizes with the adapter and a PCR primer that hybridizes with the constant insert sequence. In some embodiments, the nuclease creates blunt ends. In some embodiments, the nuclease creates a 5' overhang. In some embodiments, the nucleic acid adapter is provided in double-stranded form. In some embodiments, the 5'-phosphate-dependent ligation is a blunt end ligation. In some embodiments, 10 the method comprises filling in the 5'-overhang before ligating the first nucleic acid adapter to the nucleic acid strand that was cut by the nuclease. In some embodiments, the method further comprises enriching the amplified nucleic acid molecules for molecules comprising a single uncut target sequence. In some embodiments, the step of enriching comprises a size fractionation. In some embodiments, the determining of step (c) comprises sequencing the nucleic acid strand that was cut by the nuclease in step (b), or a copy thereof obtained via PCR. In some embodiments, the library 15 of candidate nucleic acid molecules comprises at least 10^8 , at least 10^9 , at least 10^{10} , at least 10^{11} , or at least 10^{12} different candidate nuclease cleavage sites. In some embodiments, the nuclease is a therapeutic nuclease which cuts a specific nuclease target site in a gene associated with a disease. In some embodiments, the method further comprises determining a maximum concentration of the therapeutic nuclease at which the therapeutic nuclease cuts the specific nuclease target site, and does not cut more than 10, more than 5, more than 4, more than 3, more than 2, more than 1, 20 or no additional nuclease target sites. The therapeutic nuclease can be administered to a subject in an amount effective to generate a final concentration equal or lower than the maximum concentration. In some embodiments, the nuclease is an RNA-programmable nuclease that forms a complex with an RNA molecule, and wherein the nuclease:RNA complex specifically binds a nucleic acid sequence complementary to the sequence of the RNA molecule. In some embodiments, the RNA molecule is a single-guide RNA (sgRNA). In some embodiments, the sgRNA comprises 5-50 nucleotides, 10-30 25 nucleotides, 15-25 nucleotides, 18-22 nucleotides, 19-21 nucleotides, e.g., 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, or 30 nucleotides. In some embodiments, the nuclease is a Cas9 nuclease. In some 30 embodiments, the nuclease target site comprises a [sgRNA-complementary sequence]-[protospacer adjacent motif (PAM)] structure, and the nuclease cuts the target site within the sgRNA-complementary sequence. In some embodiments, the sgRNA-complementary sequence comprises 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 35 26, 27, 28, 29, or 30 nucleotides. In some embodiments, the nuclease comprises an unspecific nucleic acid cleavage domain. In some embodiments, the nuclease comprises a FokI cleavage domain. In some embodiments, the nuclease comprises a nucleic acid cleavage domain that cleaves a target sequence upon cleavage domain dimerization. In some 40 embodiments, the nuclease comprises a binding domain that specifically binds a nucleic acid sequence. In some embodiments, the binding domain comprises a zinc finger. In some embodiments, the binding domain comprises at least 2, at least 3, at least 4, or at least 5 zinc fingers. In some embodiments, the nuclease is a Zinc Finger Nuclease. In some 45 embodiments, the binding domain comprises a Transcriptional Activator-Like Element. In some embodiments, the nuclease is a Transcriptional Activator-Like Element Nuclease (TALEN). In some embodiments, the nuclease is an organic compound. In some embodiments, the nuclease comprises an enediyne functional group. In some embodiments, the nuclease is an antibiotic. In some embodiments, the compound is dynemicin, neocarzinostatin, calicheamicin, esperamycin, 50 bleomycin, or a derivative thereof. In some embodiments, the nuclease is a homing endonuclease.

[0011] Some aspects of this disclosure provide libraries of nucleic acid molecules, in which each nucleic acid molecule comprises a concatemer of a sequence comprising a candidate nuclease target site and a constant insert sequence of 10-100 nucleotides. In some embodiments, the constant insert sequence is at least 15, at least 20, at least 25, at least 30, at least 35, at least 40, at least 45, at least 50, at least 55, at least 60, at least 65, at least 70, at least 75, at least 80, or at least 95 nucleotides long. In some embodiments, the constant insert sequence is not more than 15, not more than 20, not more than 25, not more than 30, not more than 35, not more than 40, not more than 45, not more than 50, not more than 55, not more than 60, not more than 65, not more than 70, not more than 75, not more than 80, or not more than 95 nucleotides long. In some embodiments, the candidate nuclease target sites are sites that can be cleaved by an RNA-programmable nuclease, a Zinc Finger Nuclease (ZFN), a Transcription Activator-Like Effector Nuclease (TALEN), a homing endonuclease, an organic compound nuclease, or an enediyne antibiotic (e.g., dynemicin, neocarzinostatin, calicheamicin, esperamycin, bleomycin). In some embodiments, the candidate nuclease target site can be cleaved by a Cas9 nuclease. In some embodiments, the library comprises at least 10^5 , at least 10^6 , at least 10^7 , at least 10^8 , at least 10^9 , at least 10^{10} , at least 10^{11} , or at least 10^{12} different candidate nuclease target sites. In some 55 embodiments,

iments, the library comprises nucleic acid molecules of a molecular weight of at least 0.5kDa, at least 1 kDa, at least 2 kDa, at least 3 kDa, at least 4 kDa, at least 5 kDa, at least 6 kDa, at least 7 kDa, at least 8 kDa, at least 9 kDa, at least 10 kDa, at least 12 kDa, or at least 15kDa. In some embodiments, the library comprises candidate nuclease target sites that are variations of a known target site of a nuclease of interest. In some embodiments, the variations of a known nuclease target site comprise 10 or fewer, 9 or fewer, 8 or fewer, 7 or fewer, 6 or fewer, 5 or fewer, 4 or fewer, 3 or fewer, or 2 or fewer mutations as compared to a known nuclease target site. In some embodiments, the variations differ from the known target site of the nuclease of interest by more than 5%, more than 10%, more than 15%, more than 20%, more than 25%, or more than 30% on average, distributed binomially. In some embodiments, the variations differ from the known target site by no more than 10%, no more than 15%, no more than 20%, no more than 25%, nor more than 30%, no more than 40%, or no more than 50% on average, distributed binomially. In some embodiments, the nuclease of interest is a Cas9 nuclease, a zinc finger nuclease, a TALEN, a homing endonuclease, an organic compound nuclease, or an enediyne antibiotic (e.g., dynemicin, neocarzinostatin, calicheamicin, esperamicin, bleomycin). In some embodiments, the candidate nuclease target sites are Cas9 nuclease target sites that comprise a [sgRNA-complementary sequence]-[PAM] structure, wherein the sgRNA-complementary sequence comprises 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, or 30 nucleotides.

[0012] According to a second aspect, the invention provides a method of selecting a nuclease that specifically cuts a consensus target site from a plurality of nucleases, the method comprising (a) providing a plurality of candidate nucleases that cut the same consensus sequence; (b) for each of the candidate nucleases of step (a), identifying a nuclease target site cleaved by the candidate nuclease that differ from the consensus target site using a method provided herein; (c) selecting a nuclease based on the nuclease target site(s) identified in step (b), which cleaves the consensus target site with the highest specificity. In some embodiments, the nuclease that cleaves the consensus target site with the highest specificity is the candidate nuclease that cleaves the lowest number of target sites that differ from the consensus site. In some embodiments, the candidate nuclease that cleaves the consensus target site with the highest specificity is the candidate nuclease that cleaves the lowest number of target sites that are different from the consensus site in the context of a target genome. In some embodiments, the candidate nuclease selected in step (c) is a nuclease that does not cleave any target site other than the consensus target site. In some embodiments, the candidate nuclease selected in step (c) is a nuclease that does not cleave any target site other than the consensus target site within the genome of a subject at a therapeutically effective concentration of the nuclease. In some embodiments, the nuclease selected in step (c) cleaves a genome *in vitro*. Also disclosed is a further step of contacting a genome with the nuclease selected in step (c). In some embodiments, the genome is a vertebrate, mammalian, human, non-human primate, rodent, mouse, rat, hamster, goat, sheep, cattle, dog, cat, reptile, amphibian, fish, nematode, insect, or fly genome. In some embodiments, the genome is within a living cell. In some embodiments disclosed herein, the genome is within a subject. In some embodiments, the consensus target site is within an allele that is associated with a disease or disorder. In some embodiments disclosed herein, cleavage of the consensus target site results in treatment or prevention of a disease or disorder, e.g., amelioration or prevention of at least one sign and/or symptom of the disease or disorder. In some embodiments disclosed herein, cleavage of the consensus target site results in the alleviation of a sign and/or symptom of the disease or disorder. In some embodiments disclosed herein, cleavage of the consensus target site results in the prevention of the disease or disorder. In some embodiments, the disease is HIV/AIDS. In some embodiments, the allele is a CCR5 allele. In some embodiments, the disease is a proliferative disease. In some embodiments, the disease is cancer. In some embodiments, the allele is a VEGFA allele.

[0013] Some aspects of this disclosure provide isolated nucleases that have been selected according to a method provided herein. In some embodiments, the nuclease has been engineered to cleave a target site within a genome. In some embodiments, the nuclease is a Cas9 nuclease comprising an sgRNA that is complementary to the target site within the genome. In some embodiments, the nuclease is a Zinc Finger Nuclease (ZFN) or a Transcription Activator-Like Effector Nuclease (TALEN), a homing endonuclease, or an organic compound nuclease (e.g., an enediyne, an antibiotic nuclease, dynemicin, neocarzinostatin, calicheamicin, esperamicin, bleomycin, or a derivative thereof). In some embodiments, the nuclease has been selected based on cutting no other candidate target site, not more than one candidate target site, not more than two candidate target sites, not more than three candidate target sites, not more than four candidate target sites, not more than five candidate target sites, not more than six candidate target sites, not more than seven candidate target sites, not more than eight candidate target sites, not more than nine candidate target sites, or not more than ten candidate target sites in addition to its known nuclease target site.

[0014] Some aspects of this disclosure provide kits comprising a library of nucleic acid molecules comprising candidate nuclease target sites as provided herein. Some aspects of this disclosure provide kits comprising an isolated nuclease as provided herein. In some embodiments, the nuclease is a Cas9 nuclease. In some embodiments, the kit further comprises a nucleic acid molecule comprising a target site of the isolated nuclease. In some embodiments, the kit comprises an excipient and instructions for contacting the nuclease with the excipient to generate a composition suitable for contacting a nucleic acid with the nuclease. In some embodiments, the composition is suitable for contacting a nucleic

acid within a genome. In some embodiments, the composition is suitable for contacting a nucleic acid within a cell. In some embodiments, the composition is suitable for contacting a nucleic acid within a subject. In some embodiments, the excipient is a pharmaceutically acceptable excipient.

[0015] Some aspects of this disclosure provide pharmaceutical compositions that are suitable for administration to a subject. In some embodiments, the composition comprises an isolated nuclease as provided herein. In some embodiments, the composition comprises a nucleic acid encoding such a nuclease. In some embodiments, the composition comprises a pharmaceutically acceptable excipient.

[0016] Other advantages, features, and uses of the invention will be apparent from the detailed description of certain non-limiting embodiments of the invention; the drawings, which are schematic and not intended to be drawn to scale; and the claims.

BRIEF DESCRIPTION OF THE DRAWINGS

[0017]

Figure 1. *In vitro* selection overview. (A) Cas9 complexed with a short guide RNA (sgRNA) recognizes ~20 bases of a target DNA substrate that is complementary to the sgRNA sequence and cleaves both DNA strands. The white triangles represent cleavage locations. (B) A modified version of our previously described *in vitro* selection was used to comprehensively profile Cas9 specificity. A concatemeric pre-selection DNA library in which each molecule contains one of 10^{12} distinct variants of a target DNA sequence (white rectangles) was generated from synthetic DNA oligonucleotides by ligation and rolling-circle amplification. This library was incubated with a Cas9:sgRNA complex of interest. Cleaved library members contain 5' phosphate groups (circles with "P") and therefore are substrates for adapter ligation and PCR. The resulting amplicons were subjected to high-throughput DNA sequencing and computational analysis.

Figure 2. *In vitro* selection results for Cas9:CLTA1 sgRNA. Heat maps²¹ show the specificity profiles of Cas9:CLTA1 sgRNA v2.1 under enzyme-limiting conditions (A, B), Cas9:CLTA1 sgRNA v1.0 under enzyme-saturating conditions (C, D), and Cas9:CLTA1 sgRNA v2.1 under enzyme-saturating conditions (E, F). Heat maps show all post-selection sequences (A, C, E) or only those sequences containing a single mutation in the 20-base pair sgRNA-specified target site and two-base pair PAM (B, D, F). Specificity scores of 1.0 and -1.0 corresponds to 100% enrichment for and against, respectively, a particular base pair at a particular position. Black boxes denote the intended target nucleotides. (G) Effect of Cas9:sgRNA concentration on specificity. Positional specificity changes between enzyme-limiting (200 nM DNA, 100 nM Cas9:sgRNA v2.1) and enzyme-saturating (200 nM DNA, 1000 nM Cas9:sgRNA v2.1) conditions, normalized to the maximum possible change in positional specificity, are shown for CLTA1. (H) Effect of sgRNA architecture on specificity. Positional specificity changes between sgRNA v1.0 and sgRNA v2.1 under enzyme-saturating conditions, normalized to the maximum possible change in positional specificity, are shown for CLTA1. See Figures 6-8, 25, and 26 for corresponding data for CLTA2, CLTA3, and CLTA4. Sequence Identifiers: The sgRNA sequences shown in (A-F) correspond to SEQ ID NO:1.

Figure 3. Target sites profiled in this study. (A) The 5' end of the sgRNA has 20 nucleotides that are complementary to the target site. The target site contains an NGG motif (PAM) adjacent to the region of RNA:DNA complementarity. (B) Four human clathrin gene (CLTA) target sites are shown. (C, D) Four human clathrin gene (CLTA) target sites are shown with sgRNAs. sgRNA v1.0 is shorter than sgRNA v2.1. The PAM is shown for each site. The non-PAM end of the target site corresponds to the 5' end of the sgRNA. Sequence Identifiers: The sequences shown in (B), from top to bottom, are SEQ ID NO:2; SEQ ID NO:3; SEQ ID NO:4; SEQ ID NO:5; SEQ ID NO:6; and SEQ ID NO:7. The sequences shown in (C), from top to bottom, are SEQ ID NO:8; SEQ ID NO:9; SEQ ID NO:10; SEQ ID NO:11; SEQ ID NO:12; SEQ ID NO:13; SEQ ID NO:14; SEQ ID NO:15; SEQ ID NO:16; SEQ ID NO:17; SEQ ID NO:18; and SEQ ID NO:19. The sequences shown in (D), from top to bottom, are SEQ ID NO:20; SEQ ID NO:21; SEQ ID NO:22; SEQ ID NO:23; SEQ ID NO:24; SEQ ID NO:25; SEQ ID NO:26; SEQ ID NO:27; SEQ ID NO:28; SEQ ID NO:29; SEQ ID NO:30; and SEQ ID NO:31.

Figure 4. Cas9:guide RNA cleavage of on-target DNA sequences *in vitro*. Discrete DNA cleavage assays on an approximately 1-kb linear substrate were performed with 200 nM on-target site and 100 nM Cas9:v1.0 sgRNA, 100 nM Cas9:v2.1 sgRNA, 1000 nM Cas9:v1.0 sgRNA, and 1000 nM Cas9:v2.1 sgRNA for each of four CLTA target sites. For CLTA1, CLTA2, and CLTA4, Cas9:v2.1 sgRNA shows higher activity than Cas9:v1.0 sgRNA. For CLTA3, the activities of the Cas9:v1.0 sgRNA and Cas9:v2.1 sgRNA were comparable.

Figure 5. *In vitro* selection results for four target sites. *In vitro* selections were performed on 200 nM pre-selection library with 100 nM Cas9:sgRNA v2.1, 1000 nM Cas9:sgRNA v1.0, or 1000 nM Cas9:sgRNA v2.1. (A) Post-selection PCR products are shown for the 12 selections performed. DNA containing 1.5 repeats were quantified for each selection and pooled in equimolar amounts before gel purification and sequencing. (B-E) Distributions of mutations are shown for pre-selection (black) and post-selection libraries (colored). The post-selection libraries are enriched

for sequences with fewer mutations than the pre-selection libraries. Mutations are counted from among the 20 base pairs specified by the sgRNA and the two-base pair PAM. P-values are < 0.01 for all pairwise comparisons between distributions in each panel. P-values were calculated using t-tests, assuming unequal size and unequal variance.

Figure 6. In vitro selection results for Cas9:CLTA2 sgRNA. Heat maps²⁴ show the specificity profiles of Cas9:CLTA2 sgRNA v2.1 under enzyme-limiting conditions (A, B), Cas9:CLTA2 sgRNA v1.0 under enzyme-excess conditions (C, D), and Cas9:CLTA2 sgRNA v2.1 under enzyme-excess conditions (E, F). Heat maps show all post-selection sequences (A, C, E) or only those sequences containing a single mutation in the 20-base pair sgRNA-specified target site and two-base pair PAM (B, D, F). Specificity scores of 1.0 and -1.0 corresponds to 100% enrichment for and against, respectively, a particular base pair at a particular position. Black boxes denote the intended target nucleotides. Sequence Identifiers: The sgRNA sequences shown in (A-F) correspond to SEQ ID NO:32.

Figure 7. In vitro selection results for Cas9:CLTA3 sgRNA. Heat maps²⁴ show the specificity profiles of Cas9:CLTA3 sgRNA v2.1 under enzyme-limiting conditions (A, B), Cas9:CLTA3 sgRNA v1.0 under enzyme-excess conditions (C, D), and Cas9:CLTA3 sgRNA v2.1 under enzyme- saturating conditions (E, F). Heat maps show all post-selection sequences (A, C, E) or only those sequences containing a single mutation in the 20-base pair sgRNA-specified target site and two-base pair PAM (B, D, F). Specificity scores of 1.0 and -1.0 corresponds to 100% enrichment for and against, respectively, a particular base pair at a particular position. Black boxes denote the intended target nucleotides. Sequence Identifiers: The sgRNA sequences shown in (A-F) correspond to SEQ ID NO:32.

Figure 8. In vitro selection results for Cas9:CLTA4 sgRNA. Heat maps²⁴ show the specificity profiles of Cas9:CLTA4 sgRNA v2.1 under enzyme-limiting conditions (A, B), Cas9:CLTA4 sgRNA v1.0 under enzyme-excess conditions (C, D), and Cas9:CLTA4 sgRNA v2.1 under enzyme- saturating conditions (E, F). Heat maps show all post-selection sequences (A, C, E) or only those sequences containing a single mutation in the 20-base pair sgRNA-specified target site and two-base pair PAM (B, D, F). Specificity scores of 1.0 and -1.0 corresponds to 100% enrichment for and against, respectively, a particular base pair at a particular position. Black boxes denote the intended target nucleotides. Sequence Identifiers: The sgRNA sequences shown in (A-F) correspond to SEQ ID NO:33.

Figure 9. In vitro selection results as sequence logos. Information content is plotted²⁵ for each target site position (1-20) specified by CLTA1 (A), CLTA2 (B), CLTA3 (C), and CLTA4 (D) sgRNA v2.1 under enzyme-limiting conditions. Positions in the PAM are labelled "P1," "P2," and "P3." Information content is plotted in bits. 2.0 bits indicates absolute specificity and 0 bits indicates no specificity.

Figure 10. Tolerance of mutations distal to the PAM for CLTA1. The maximum specificity scores at each position are shown for the Cas9:CLTA1 v2.1 sgRNA selections when considering only those sequences with on-target base pairs in gray, while allowing mutations in the first 1-12 base pairs (A-L). The positions that are not constrained to on-target base pairs are indicated by dark bars. Higher specificity score values indicate higher specificity at a given position. The positions that were not allowed to contain any mutations (gray) were plotted with a specificity score of +1. For all panels, specificity scores were calculated from pre-selection library sequences and post-selection library sequences with an n ≥ 5,130 and n ≥ 74,538, respectively.

Figure 11. Tolerance of mutations distal to the PAM for CLTA2. The maximum specificity scores at each position are shown for the Cas9:CLTA2 v2.1 sgRNA selections when considering only those sequences with on-target base pairs in gray, while allowing mutations in the first 1-12 base pairs (A-L). The positions that are not constrained to on-target base pairs are indicated by dark bars. Higher specificity score values indicate higher specificity at a given position. The positions that were not allowed to contain any mutations (gray) were plotted with a specificity score of +1. For all panels, specificity scores were calculated from pre-selection library sequences and post-selection library sequences with an n ≥ 3,190 and n ≥ 25,365, respectively.

Figure 12. Tolerance of mutations distal to the PAM for CLTA3. The maximum specificity scores at each position are shown for the Cas9:CLTA3 v2.1 sgRNA selections when considering only those sequences with on-target base pairs in gray, while allowing mutations in the first 1-12 base pairs (A-L). The positions that are not constrained to on-target base pairs are indicated by dark bars. Higher specificity score values indicate higher specificity at a given position. The positions that were not allowed to contain any mutations (gray) were plotted with a specificity score of +1. For all panels, specificity scores were calculated from pre-selection library sequences and post-selection library sequences with an n ≥ 5,604 and n ≥ 158,424, respectively.

Figure 13. Tolerance of mutations distal to the PAM for CLTA4. The maximum specificity scores at each position are shown for the Cas9:CLTA4 v2.1 sgRNA selections when considering only those sequences with on-target base pairs in gray, while allowing mutations in the first 1-12 base pairs (A-L). The positions that are not constrained to on-target base pairs are indicated by dark bars. Higher specificity score values indicate higher specificity at a given position. The positions that were not allowed to contain any mutations (gray) were plotted with a specificity score of +1. For all panels, specificity scores were calculated from pre-selection library sequences and post-selection

library sequences with an $n \geq 2,323$ and $n \geq 21,819$, respectively.

Figure 14. Tolerance of mutations distal to the PAM in CLTA1 target sites. Distributions of mutations are shown for *in vitro* selection on 200 nM pre-selection library with 1000 nM Cas9:CLTA1 sgRNA v2.1. The number of mutations shown are in a 1-12 base pair target site subsequence farthest from the PAM (A-L) when the rest of the target site, including the PAM, contains only on-target base pairs. The pre-selection and post-selection distributions are similar for up to three base pairs, demonstrating tolerance for target sites with mutations in the three base pairs farthest from the PAM when the rest of the target sites have optimal interactions with the Cas9:sgRNA. For all panels, graphs were generated from pre-selection library sequences and post-selection library sequences with an $n \geq 5,130$ and $n \geq 74,538$, respectively.

Figure 15. Tolerance of mutations distal to the PAM in CLTA2 target sites. Distributions of mutations are shown for *in vitro* selection on 200 nM pre-selection library with 1000 nM Cas9:CLTA2 sgRNA v2.1. The number of mutations shown are in a 1-12 base pair target site subsequence farthest from the PAM (A-L) when the rest of the target site, including the PAM, contains only on-target base pairs. The pre-selection and post-selection distributions are similar for up to three base pairs, demonstrating tolerance for target sites with mutations in the three base pairs farthest from the PAM when the rest of the target sites have optimal interactions with the Cas9:sgRNA. For all panels, graphs were generated from pre-selection library sequences and post-selection library sequences with an $n \geq 3,190$ and $n \geq 21,265$, respectively.

Figure 16. Tolerance of mutations distal to PAM in CLTA3 target sites. Distributions of mutations are shown for *in vitro* selection on 200 nM pre-selection library with 1000 nM Cas9:CLTA3 sgRNA v2.1. The number of mutations shown are in a 1-12 base pair target site subsequence farthest from the PAM (A-L) when the rest of the target site, including the PAM, contains only on-target base pairs. The pre-selection and post-selection distributions are similar for up to three base pairs, demonstrating tolerance for target sites with mutations in the three base pairs farthest from the PAM when the rest of the target sites have optimal interactions with the Cas9:sgRNA. For all panels, graphs were generated from pre-selection library sequences and post-selection library sequences with an $n \geq 5,604$ and $n \geq 158,424$, respectively.

Figure 17. Tolerance of mutations distal to PAM in CLTA4 target sites. Distributions of mutations are shown for *in vitro* selection on 200 nM pre-selection library with 1000 nM Cas9:CLTA4 sgRNA v2.1. The number of mutations shown are in a 1-12 base pair target site subsequence farthest from the PAM (A-L) when the rest of the target site, including the PAM, contains only on-target base pairs. The pre-selection and post-selection distributions are similar for up to three base pairs, demonstrating tolerance for target sites with mutations in the three base pairs farthest from the PAM when the rest of the target sites have optimal interactions with the Cas9:sgRNA. For all panels, graphs were generated from pre-selection library sequences and post-selection library sequences with an $n \geq 2,323$ and $n \geq 21,819$, respectively.

Figure 18. Positional specificity patterns for 100 nM Cas9:sgRNA v2.1. Positional specificity, defined as the sum of the magnitude of the specificity score for each of the four possible base pairs recognized at a certain position in the target site, is plotted for each target site under enzyme-limiting conditions for sgRNA v2.1. The positional specificity is shown as a value normalized to the maximum positional specificity value of the target site. Positional specificity is highest at the end of the target site proximal to the PAM and is lowest in the middle of the target site and in the several nucleotides most distal to the PAM

Figure 19. Positional specificity patterns for 1000 nM Cas9:sgRNA v1.0. Positional specificity, defined as the sum of the magnitude of the specificity score for each of the four possible base pairs recognized at a certain position in the target site, is plotted for each target site under enzyme-excess conditions with sgRNA v1.0. The positional specificity is shown as a value normalized to the maximum positional specificity value of the target site. Positional specificity is relatively constant across the target site but is lowest in the middle of the target site and in the several nucleotides most distal to the PAM

Figure 20. Positional specificity patterns for 1000 nM Cas9:sgRNA v2.1. Positional specificity, defined as the sum of the magnitude of the specificity score for each of the four possible base pairs recognized at a certain position in the target site, is plotted for each target site under enzyme-excess conditions with sgRNA v2.1. The positional specificity is shown as a value normalized to the maximum positional specificity value of the target site. Positional specificity is relatively constant across the target site but is lowest in the middle of the target site and in the several nucleotides most distal to the PAM

Figure 21. PAM nucleotide preferences. The abundance in the pre-selection library and post-selection libraries under enzyme-limiting or enzyme-excess conditions are shown for all 16 possible PAM dinucleotides for selections with CLTA1 (A), CLTA2 (B), CLTA3 (C), and CLTA4 (D) sgRNA v2.1. GG dinucleotides increased in abundance in the post-selection libraries, while the other possible PAM dinucleotides decreased in abundance after the selection.

Figure 22. PAM nucleotide preferences for on-target sites. Only post-selection library members containing no mutations in the 20 base pairs specified by the guide RNAs were included in this analysis. The abundance in the pre-selection library and post-selection libraries under enzyme-limiting and enzyme-excess conditions are shown

for all 16 possible PAM dinucleotides for selections with CLTA1 (A), CLTA2 (B), CLTA3 (C), and CLTA4 (D) sgRNA v2.1. GG dinucleotides increased in abundance in the post-selection libraries, while the other possible PAM dinucleotides generally decreased in abundance after the selection, although this effect for the enzyme-excess concentrations of Cas9:sgRNA was modest or non-existent for many dinucleotides.

Figure 23. PAM dinucleotide specificity scores. The specificity scores under enzyme-limiting and enzyme-excess conditions are shown for all 16 possible PAM dinucleotides (positions 2 and 3 of the three-nucleotide NGG PAM) for selections with CLTA1 (A), CLTA2 (B), CLTA3 (C), and CLTA4 (D) sgRNA v2.1. The specificity score indicates the enrichment of the PAM dinucleotide in the post-selection library relative to the pre-selection library, normalized to the maximum possible enrichment of that dinucleotide. A specificity score of +1.0 indicates that a dinucleotide is 100% enriched in the post-selection library, and a specificity score of -1.0 indicates that a dinucleotide is 100% de-enriched. GG dinucleotides were the most enriched in the post-selection libraries, and AG, GA, GC, GT, and TG show less relative de-enrichment compared to the other possible PAM dinucleotides.

Figure 24. PAM dinucleotide specificity scores for on-target sites. Only post-selection library members containing no mutations in the 20 base pairs specified by the guide RNAs were included in this analysis. The specificity scores under enzyme-limiting and enzyme-excess conditions are shown for all 16 possible PAM dinucleotides (positions 2 and 3 of the three-nucleotide NGG PAM) for selections with CLTA1 (A), CLTA2 (B), CLTA3 (C), and CLTA4 (D) sgRNA v2.1. The specificity score indicates the enrichment of the PAM dinucleotide in the post-selection library relative to the pre-selection library, normalized to the maximum possible enrichment of that dinucleotide. A specificity score of +1.0 indicates that a dinucleotide is 100% enriched in the post-selection library, and a specificity score of -1.0 indicates that a dinucleotide is 100% de-enriched. GG dinucleotides were the most enriched in the post-selection libraries, AG and GA nucleotides were neither enriched or de-enriched in at least one selection condition, and GC, GT, and TG show less relative de-enrichment compared to the other possible PAM dinucleotides.

Figure 25. Effects of Cas9:sgRNA concentration on specificity. Positional specificity changes between enzyme-limiting (200 nM DNA, 100 nM Cas9:sgRNA v2.1) and enzyme-excess (200 nM DNA, 1000 nM Cas9:sgRNA v2.1) conditions are shown for selections with sgRNAs targeting CLTA1 (A), CLTA2 (B), CLTA3 (C), and CLTA4 (D) target sites. Lines indicate the maximum possible change in positional specificity for a given position. The highest changes in specificity occur proximal to the PAM as enzyme concentration is increased.

Figure 26. Effects of sgRNA architecture on specificity. Positional specificity changes between Cas9:sgRNA v1.0 and Cas9:sgRNA v2.1 under enzyme-excess (200 nM DNA, 1000 nM Cas9:sgRNA v2.1) conditions are shown for selections with sgRNAs targeting CLTA1 (A), CLTA2 (B), CLTA3 (C), and CLTA4 (D) target sites. Lines indicate the maximum possible change in positional specificity for a given position.

Figure 27. Cas9:guide RNA cleavage of off-target DNA sequences *in vitro*. Discrete DNA cleavage assays on a 96-bp linear substrate were performed with 200 nM DNA and 1000 nM Cas9:CLTA4 v2.1 sgRNA for the on-target CLTA4 site (CLTA4-0) and five CLTA4 off-target sites identified by *in vitro* selection. Enrichment values shown are from the *in vitro* selection with 1000 nM Cas9:CLTA4 v2.1 sgRNA. CLTA4-1 and CLTA4-3 were the most highly enriched sequences under these conditions. CLTA4-2a, CLTA4-2b, and CLTA4-2c are two-mutation sequences that represent a range of enrichment values from high enrichment to no enrichment to high de-enrichment. Lowercase letters indicate mutations relative to the on-target CLTA4 site. The enrichment values are qualitatively consistent with the observed amount of cleavage *in vitro*. Sequence Identifiers: The sequences shown from top to bottom, are SEQ ID NO:34; SEQ ID NO:35; SEQ ID NO:36; SEQ ID NO:37; SEQ ID NO:38; and SEQ ID NO:39.

Figure 28. Effect of guide RNA architecture and Cas9:sgRNA concentration on *in vitro* cleavage of an off-target site. Discrete DNA cleavage assays on a 96-bp linear substrate were performed with 200 nM DNA and 100 nM Cas9:v1.0 sgRNA, 100 nM Cas9:v2.1 sgRNA, 1000 nM Cas9:v1.0 sgRNA, or 1000 nM Cas9:v2.1 sgRNA for the CLTA4-3 off-target site (5' GggGATGTAGTGTTCACtGGG 3' (SEQ ID NO:39)-mutations are shown in lowercase letters). DNA cleavage is observed under all four conditions tested, and cleavage rates are higher under enzyme-excess conditions, or with v2.1 sgRNA compared with v1.0 sgRNA.

DEFINITIONS

[0018] As used herein and in the claims, the singular forms "a," "an," and "the" include the singular and the plural reference unless the context clearly indicates otherwise. Thus, for example, a reference to "an agent" includes a single agent and a plurality of such agents.

[0019] The term "Cas9" or "Cas9 nuclease" refers to an RNA-guided nuclease comprising a Cas9 protein, or a fragment thereof. A Cas9 nuclease is also referred to sometimes as a cas9n nuclease or a CRISPR (clustered regularly interspaced short palindromic repeat)-associated nuclease. CRISPR is an adaptive immune system that provides protection against mobile genetic elements (e.g., viruses, transposable elements and conjugative plasmids). CRISPR clusters contain spacers, sequences complementary to antecedent mobile elements, and target invading nucleic acids. CRISPR clusters are transcribed and processed into CRISPR RNA (crRNA). In type II CRISPR systems correct processing of pre-crRNA

requires a trans-encoded small RNA (tracrRNA), endogenous ribonuclease 3 (rnc) and a Cas9 protein. The tracrRNA serves as a guide for ribonuclease 3-aided processing of pre-crRNA. Subsequently, Cas9/crRNA/tracrRNA endonucleolytically cleaves linear or circular dsDNA target complementary to the spacer. The target strand not complementary to crRNA is first cut endonucleolytically, then trimmed 3'-5' exonucleolytically. In nature, DNA-binding and cleavage typically requires protein and both RNA species. However, single guide RNAs ("sgRNA", or simply "gRNA") can be engineered so as to incorporate aspects of both the crRNA and tracrRNA into a single RNA molecule. See, e.g., Jinek M., Chylinski K., Fonfara I., Hauer M., Doudna J.A., Charpentier E. Science 337:816-821(2012). Cas9 recognizes a short motif in the CRISPR repeat sequences (the PAM or protospacer adjacent motif) to help distinguish self versus non-self. Cas9 nuclease sequences and structures are well known to those of skill in the art (see, e.g., "Complete genome sequence of an M1 strain of *Streptococcus pyogenes*." Ferretti J.J., McShan W.M., Ajdic D.J., Savic D.J., Savic G., Lyon K., Primeaux C., Sezate S., Suvorov A.N., Kenton S., Lai H.S., Lin S.P., Qian Y., Jia H.G., Najar F.Z., Ren Q., Zhu H., Song L. expand/collapse author list McLaughlin R.E., Proc. Natl. Acad. Sci. U.S.A. 98:4658-4663(2001); "CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III." Deltcheva E., Chylinski K., Sharma C.M., Gonzales K., Chao Y., Pirzada Z.A., Eckert M.R., Vogel J., Charpentier E., Nature 471:602-607(2011); and "A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity." Jinek M., Chylinski K., Fonfara I., Hauer M., Doudna J.A., Charpentier E. Science 337:816-821(2012)). Cas9 orthologs have been described in various species, including, but not limited to, *S. pyogenes* and *S. thermophilus*. Additional suitable Cas9 nucleases and sequences will be apparent to those of skill in the art based on this disclosure, and such Cas9 nucleases and sequences include Cas9 sequences from the organisms and loci disclosed in Chylinski, Rhun, and Charpentier, "The tracrRNA and Cas9 families of type II CRISPR-Cas immunity systems" (2013) RNA Biology 10:5, 726-737. In some embodiments, proteins comprising Cas9 or fragments thereof proteins are referred to as "Cas9 variants." A Cas9 variant shares homology to Cas9, or a fragment thereof. For example a Cas9 variant is at least about 70% identical, at least about 80% identical, at least about 90% identical, at least about 95% identical, at least about 98% identical, at least about 99% identical, at least about 99.5% identical, or at least about 99.9% to wild type Cas9. In some embodiments, the Cas9 variant comprises a fragment of Cas9 (e.g., a gRNA binding domain or a DNA-cleavage domain), such that the fragment is at least about 70% identical, at least about 80% identical, at least about 90% identical, at least about 95% identical, at least about 98% identical, at least about 99% identical, at least about 99.5% identical, or at least about 99.9% to the corresponding fragment of wild type Cas9. In some embodiments, wild type Cas9 corresponds to Cas9 from *Streptococcus pyogenes* (NCBI Reference Sequence: NC_017053.1, SEQ ID NO:40 (nucleotide); SEQ ID NO:41 (amino acid)).

30

35

40

45

50

55

ATGGATAAGAAACTCAATAGGCTAGATATCGGCACAAATAGCGTCGGATGGCGGTGATCACTGATG
 ATTATAAGGTTCCGTCTAAAAGTTCAAGGTTCTGGAAATACAGACGCCACAGTATCAAAAAAATCT
 TATAGGGCTCTTTATTTGGCAGTGGAGAGACAGCGGAAGCGACTCGTCTCAAACGGACAGCTGTAGA
 5 AGGTATACACGTCGGAAGAATCGTATTGTTATCTACAGGAGATTTTCAAATGAGATGGCAGAGTAG
 ATGATAGTTCTTCATCGACTTGAAAGAGTCTTTGGTGAAGACAAGAACATGAACGTACATCC
 TATTTTGAAATATAGTAGATGAAGTTGCTTATCATGAGAAATATCCAACATCTATCATCTGCAGAA
 10 AAATTGGCAGATTCTACTGATAAAGCGGATTCGCTTAATCTATTGGCCTAGCGCATATGATTAAGT
 TTCGGTGCATTTTGATTGAGGGAGTTAAATCCTGATAATAGTGATGTGGACAAACTATTTATCCA
 GTTGGTACAAATCTACAATCAATTATTGAAGAAAACCCTATTAACGCAAGTAGAGTAGATGCTAAAGCG
 ATTCTTCTGCACGATTGAGTAAATCAAGACGATTAGAAAATCTCATTGCTCAGCTCCCCGGTGAGAAGA
 15 GAAATGGCTTGGATCTCATTGCTTGTCAATTGGGATTGACCCCTAATTAAATCAAATTGAA
 TTTGGCAGAAGATGCTAAATTACAGCTTCAAAGATACTACGATGATGATTAGATAATTATTGGCG
 CAAATTGGAGATCAATATGCTGATTGTTGGCAGCTAAGAATTATCAGATGCTATTACTTCAG
 ATATCCTAACAGAGTAAATAGTGAAATAACTAAGGCTCCCCTACGCTCAATGATTAAGCGTACGATGA
 20 ACATCATCAAGACTTGAECTTTAAAAGCTTGTGACAACAACCTCCAGAAAAGTATAAAGAAATC
 TTTTGATCAATCAAAAACGGATATGCAGGTTATATTGATGGGGAGCTAGCCAAGAAGAATTTATA
 AATTATCAAACCAATTAGAAAAATGGATGGTACTGAGGAATTATTGGTAAACTAAATCGTGAAGA
 TTTGCTGCGCAAGCAACGGACCTTGACAACGGCTCTATTCCCCATCAAATTCACTGGGTGAGCTGCAT
 25 GCTATTGAGAAGACAAGAAGACTTTATCCATTAAAGACAATCGTGAGAAGATTGAAAAATCT
 TGACTTTGCAATTCTTATTATGTTGGTCCATTGGCGCTGGCAATAGTCGTTTGATGGTGA
 GAAGTCTGAAGAACATTACCCATGGAATTGAGAAGATTGTCGATAAAGGTGCTCAGCTCAATCA
 TTTATTGACGCAACTTGAACAAACTTGTGAAATTGACAAAGGTCAAATATGTTACTGAGGGATGCGAAAACC
 30 AGCATTCTTCAGGTGAACAGAAGAAAGCCATTGTTGATTTACTCTCAAAACAAATCGAAAAGTAACC
 GTTAAGCAATTAAAAGAAGATTATTCAAAAAAATAGAATGTTGATAGTGTGAAATTTCAGGAGTTG
 AAGATAGATTAAATGCTCATTAGGCGCTACCATGATTGCTAAAATTATTAAAGATAAAGATTTTT
 GGATAATGAAGAAAATGAAGATATCTTAGAGGATTGTTAACATTGACCTTATTGAAGATAGGGGG
 35 ATGATTGAGGAAAGACTAAACATATGCTCACCTCTTGATGATAAGGTGATGAAACAGCTTAAACGTC
 GCCGTATACTGGTGGGACGTTGTCGAAAATTGATTAATGGTATTAGGATAAGCAATCTGGCAA
 AACAAATTAGATTGAAATCAGATGGTTGCCAATCGCAATTATGAGCTGATCCATGATGAT
 40 AGTTTGACATTAAAGAAGATATTCAAAAGCACAGGTGCTGGACAAGGCCATAGTTACATGAACAGA
 TTGCTAACCTAGCTGGCAGTCCTGCTATTAAAAAGGTATTACAGACTGTAAGGTTGATGA
 GGTCAAAGTAATGGGCATAAGCCAGAAAATACGTTATTGAAATGGCACGTGAAAATCAGACAACTCAA
 45 AAGGCCAGAAAATCGCAGAGCGTATGAAACGAATCGAAGAAGGTCAAAGAATTAGGAAGTCAGA
 TTCTAAAGAGCATCTGTTGAAAATACTCAATTGCAAATGAAAAGCTCTATCTATTATCTACAAA
 TGGAAGAGACATGTATGTGGACCAAGAATTAGATATTACGTTAAGTGATTATGATGTCGATCACATT
 GTTCCACAAAGTTCTTAAAGACGATTCAATAGACAATAAGGTACTAACGCGTTCTGATAAAAATCGT
 50 GTAAATCGATAACGTTCAAGTGAAAGAAGTAGTCAAAAGATGAAAAGCTATTGGAGACAACCTCTAAA
 CGCCAAGTTAACACTCAACGTAAGTTGATAATTAAACGAAAGCTGAACTGGAGGTTGAGTGAACCTT
 GATAAGCTGGTTTATCAAACGCCAATTGGTGAAACTCGCCAAATCACTAAGCATGTGGCACAAATT
 55

TGGATAGTCGCATGAATACTAAATACGATGAAAATGATAAACTTATTGAGAGGTTAAAGTGATTACCTT
 AAAATCTAAATTAGTTCTGACTTCCGAAAAGATTCCAATTCTATAAAGTACGTGAGATTAACAATTAC
 5 CATCATGCCCATGATGCGTATCTAAATGCCGTCGGACTGCTTGATTAAGAAATATCCAAAACCTG
 AATCGGAGTTGTCTATGGTATTAAAGTTATGATGTCGTAAGTCTGAGCAAGA
 10 AATAGGCAAAGCAACCGAAAATATTCTTTACTCTAATATCATGAACCTCTCAAAACAGAAATTACA
 CTTGCAAATGGAGAGATTGCAAACGCCCTTAATCGAAACTAATGGGAAACTGGAGAAATTGTCTGGG
 ATAAAGGGCGAGATTGCCCACAGTGCAGTAAAGTATTGTCATGCCCAAGTCAATATTGTCAGAAAAC
 15 AGAAGTACAGACAGGCAGATTCTCAAGGAGTCATTTACCAAAAAAGAAATTGGACAAGCTTATTGCT
 CGTAAAAAGACTGGGATCBBBBBAAATGGGGTTTGATAGTCCACGGTAGCTTATTGTCAGCTAG
 TGGTTGCTAAGGTGGAAAAGGGAAATCGAAGAAGTTAAATCGTTAAAGAGTTACTAGGGATCACAAT
 20 TATGGAAAGAAGTCCATTGAAAAAAATCCGATTGACTTTAGAAGCTAAAGGATATAAGGAAGTTAAA
 AAAGACTTAATCATTAAACTACCTAAATATAGTCTTTGAGTTAGAAAACGGTCGAAACGGATGCTGG
 CTAGTGCAGGAGATTACAAAAGGAAATGAGCTGGCTGCAAGCAAATATGTGAATTTTATATT
 25 AGCTAGTCATTATGAAAAGTTGAAGGGTAGTCCAGAAGATAACGAACAAAACAATTGTTGTGGAGCAG
 CATAAGCATTATTAGATGAGATTATTGAGCAAATCAGTGAATTCTAAGCGTGTATTAGCAGATG
 CCAATTAGATAAAAGTTCTAGTCATATAACAAACATAGAGACAAACCAATACGTGAACAAGCAGAAA
 TATTATTCAATTACGTTGACGAATCTGGAGCTCCGCTGCTTTAAATATTTGATACAACAATT
 GATCGTAAACGATATACGCTACAAAAGAAGTTAGATGCCACTCTTATCCATCAATCCATCACTGGTC
 30 TTTATGAAACACGCATTGAGTCAGCTAGGAGGTGACTGA (SEQ ID NO:1)

MDKKYSIGLDIGTNSVGAVITDDYKVPSSKKVVLGNTDRHSIKKNLIGALLFGSGETAERTRLKRTARR
 RYTRRKRNRCYLQEIFSNEAKVDDSFHRLEESFLVEEDKKHERHPIFGNIVDEVAYHEKYPTIYHLRK
 KLADSTDKADLRLLIYLALAHMIKFRGHFLIEGDLNPDNSDVKLFITQLVQIYNQLFEENPINASRVDAKA
 ILSARLSKSRRLENLIAQLPGEKRNGLFGNLIALSGLTPNFKSNFDLAEDAKLQLSKDTYDDLDNLIA
 QIGDQYADLFLAAKNLSDAILLSDILRVNSEITKAPLSASMIKRYDEHHQDLTLLKALVRQQLPKEKYKEI
 FFDQSNGYAGYIDGGASQEEFYKFIKPILEKMDGTEELLVKLNRDNGSIPHQIHLGELH
 AILRQEDFYPFLKDNRKIEKILFRIPYYVGPLARGNSRFAMTRKSEETITPWNFEVVVDKGASAQS
 35 FIERMTNFDKNLPNEKVLPHSLLYEYFTVYNELTKVVKYVTEGMRKPAAFLSGEQKKAIVDLLFKTNRKVT
 VKQLKEDYFKKIECFDSVEISGVEDRFNASLGAYHDLLKIKDQFLDNEENEDILEDIVLTLFEDRG
 MIEERLKYTAHLFDDKVMQQLKRRRTGWRRLSRKLINGIRDQSGKTIIDFLKSDGFANRNFQMLIHDD
 SLTFKEDIQKAQVGQGHSLHEQIANLAGSPAIIKKGILQTVKIVDELVKVMGHKPENIVIEMARENQTTQ
 40 KGQKNSRERMKRIEEGIKELGSQILKEHPVENTQLQNEKLYLYLQNGRDMYVDQELDINRLSDYDVHD
 VPQSFIKDDSIDNKVLTRSDKNRGSDNVPSEEVVKKMKNYWRQLLNAKLITQRKFDNLTKAERGGLSEL
 DKAGFIKRQLVETRQITKVAQILDSSRMNTKYDENDKLIREVKVITLKSCLVSDFRKDFQFYKVREINNY
 HHAHDAYLNAAVVGTLAKKYPKLESEFVYGDYKVDVRKMIAKSEQEIGKATAKYFFYSNIMNFFKTEIT
 45 LANGEIRKRPLIETNGETGEIVWDKGRDFATVRKVLSPMQVNIVKKTEVQTGGFSKESILPKRNSDKLIA
 RKKDWDPKKYGGFDSPTVAYSVLVAKVEKGSKKLKSVKELLGITIMERSSFEKNPIDFLEAKGYKEVK
 KDLIIKLPKYSIFELENGRKRMILASAGEIQLKGNELALPSKYVNFLASHYEKLKGSPEDNEQKQLFVEQ
 HKHYLDEIIEQISEFSKRVILADANLDKVL SAYNKHRDKPIREQAENIIHLFTLTNLGAPAAFKYFDTTI
 DRKRYTSTKEVLDATLIHQ SITGLYETRIDSQLGGD (SEQ ID NO:2)

[0020] The term "concatemer," as used herein in the context of nucleic acid molecules, refers to a nucleic acid molecule that contains multiple copies of the same DNA sequences linked in a series. For example, a concatemer comprising ten copies of a specific sequence of nucleotides (e.g., [XYZ]₁₀), would comprise ten copies of the same specific sequence linked to each other in series, e.g., 5'-XYZXYZXYZXYZXYZXYZXYZXYZ -3'. A concatemer may comprise any number of copies of the repeat unit or sequence, e.g., at least 2 copies, at least 3 copies, at least 4 copies, at least 5 copies, at least 10 copies, at least 100 copies, at least 1000 copies, etc. An example of a concatemer of a nucleic acid sequence comprising a nuclease target site and a constant insert sequence would be [(target site)-(constant insert sequence)]₃₀₀. A concatemer may be a linear nucleic acid molecule, or may be circular.

[0021] The terms "conjugating," "conjugated," and "conjugation" refer to an association of two entities, for example, of two molecules such as two proteins, two domains (e.g., a binding domain and a cleavage domain), or a protein and

an agent, e.g., a protein binding domain and a small molecule. In some aspects, the association is between a protein (e.g., RNA-programmable nuclease) and a nucleic acid (e.g., a guide RNA). The association can be, for example, via a direct or indirect (e.g., via a linker) covalent linkage or via noncovalent interactions. In some embodiments, the association is covalent. In some embodiments, two molecules are conjugated via a linker connecting both molecules. For example, in some embodiments where two proteins are conjugated to each other, e.g., a binding domain and a cleavage domain of an engineered nuclease, to form a protein fusion, the two proteins may be conjugated via a polypeptide linker, e.g., an amino acid sequence connecting the C-terminus of one protein to the N-terminus of the other protein.

[0022] The term "consensus sequence," as used herein in the context of nucleic acid sequences, refers to a calculated sequence representing the most frequent nucleotide residues found at each position in a plurality of similar sequences.

Typically, a consensus sequence is determined by sequence alignment in which similar sequences are compared to each other and similar sequence motifs are calculated. In the context of nuclease target site sequences, a consensus sequence of a nuclease target site may, in some embodiments, be the sequence most frequently bound, or bound with the highest affinity, by a given nuclease. With respect to RNA-programmable nuclease (e.g., Cas9) target site sequences, the consensus sequence may, in some embodiments, be the sequence or region to which a gRNA, or a plurality of gRNAs, is expected or designed to bind, e.g., based on complementary base pairing.

[0023] The term "effective amount," as used herein, refers to an amount of a biologically active agent that is sufficient to elicit a desired biological response. For example, in some embodiments, an effective amount of a nuclease may refer to the amount of the nuclease that is sufficient to induce cleavage of a target site specifically bound and cleaved by the nuclease. As will be appreciated by the skilled artisan, the effective amount of an agent, e.g., a nuclease, a hybrid protein, or a polynucleotide, may vary depending on various factors as, for example, on the desired biological response, the specific allele, genome, target site, cell, or tissue being targeted, and the agent being used.

[0024] The term "enediyne," as used herein, refers to a class of bacterial natural products characterized by either nine- and ten-membered rings containing two triple bonds separated by a double bond (see, e.g., K. C. Nicolaou; A. L. Smith; E. W. Yue (1993). "Chemistry and biology of natural and designed enediynes". PNAS 90 (13): 5881-5888). Some enediynes are capable of undergoing Bergman cyclization, and the resulting diradical, a 1,4-dehydrobenzene derivative, is capable of abstracting hydrogen atoms from the sugar backbone of DNA which results in DNA strand cleavage (see, e.g., S. Walker; R. Landovitz; W.D. Ding; G.A. Ellestad; D. Kahne (1992). "Cleavage behavior of calicheamicin gamma 1 and calicheamicin T". Proc Natl Acad Sci U.S.A. 89 (10): 4608-12). Their reactivity with DNA confers an antibiotic character to many enediynes, and some enediynes are clinically investigated as anticancer antibiotics. Nonlimiting examples of enediynes are dynemicin, neocarzinostatin, calicheamicin, esperamicin (see, e.g., Adrian L. Smith and K. C. Bicolaou, "The Enediyne Antibiotics" J. Med. Chem., 1996, 39 (11), pp 2103-2117; and Donald Borders, "Enediyne antibiotics as antitumor agents," Informa Healthcare; 1st edition (November 23, 1994, ISBN-10: 0824789385).

[0025] The term "homing endonuclease," as used herein, refers to a type of restriction enzymes typically encoded by introns or inteins Edgell DR (February 2009). "Selfish DNA: homing endonucleases find a home". Curr Biol 19 (3): R115-R117; Jasin M (Jun 1996). "Genetic manipulation of genomes with rare-cutting endonucleases". Trends Genet 12 (6): 224-8; Burt A, Koufopanou V (December 2004). "Homing endonuclease genes: the rise and fall and rise again of a selfish element". Curr Opin Genet Dev 14 (6): 609-15. Homing endonuclease recognition sequences are long enough to occur randomly only with a very low probability (approximately once every 7×10^{10} bp), and are normally found in only one instance per genome.

[0026] The term "library," as used herein in the context of nucleic acids or proteins, refers to a population of two or more different nucleic acids or proteins, respectively. For example, a library of nuclease target sites comprises at least two nucleic acid molecules comprising different nuclease target sites. In some embodiments, a library comprises at least 10^1 , at least 10^2 , at least 10^3 , at least 10^4 , at least 10^5 , at least 10^6 , at least 10^7 , at least 10^8 , at least 10^9 , at least 10^{10} , at least 10^{11} , at least 10^{12} , at least 10^{13} , at least 10^{14} , or at least 10^{15} different nucleic acids or proteins. In some embodiments, the members of the library may comprise randomized sequences, for example, fully or partially randomized sequences. In some embodiments, the library comprises nucleic acid molecules that are unrelated to each other, e.g., nucleic acids comprising fully randomized sequences. In other embodiments, at least some members of the library may be related, for example, they may be variants or derivatives of a particular sequence, such as a consensus target site sequence.

[0027] The term "linker," as used herein, refers to a chemical group or a molecule linking two adjacent molecules or moieties, e.g., a binding domain and a cleavage domain of a nuclease. Typically, the linker is positioned between, or flanked by, two groups, molecules, or other moieties and connected to each one via a covalent bond, thus connecting the two. In some embodiments, the linker is an amino acid or a plurality of amino acids (e.g., a peptide or protein). In some embodiments, the linker is an organic molecule, group, polymer, or chemical moiety.

[0028] The term "nuclease," as used herein, refers to an agent, for example a protein or a small molecule, capable of cleaving a phosphodiester bond connecting nucleotide residues in a nucleic acid molecule. In some embodiments, a nuclease is a protein, e.g., an enzyme that can bind a nucleic acid molecule and cleave a phosphodiester bond connecting nucleotide residues within the nucleic acid molecule. A nuclease may be an endonuclease, cleaving a phosphodiester

bonds within a polynucleotide chain, or an exonuclease, cleaving a phosphodiester bond at the end of the polynucleotide chain. In some embodiments, a nuclease is a site-specific nuclease, binding and/or cleaving a specific phosphodiester bond within a specific nucleotide sequence, which is also referred to herein as the "recognition sequence," the "nuclease target site," or the "target site." In some embodiments, a nuclease is a RNA-guided (*i.e.*, RNA-programmable) nuclease,

5 which complexes with (*e.g.*, binds with) an RNA having a sequence that complements a target site, thereby providing the sequence specificity of the nuclease. In some embodiments, a nuclease recognizes a single stranded target site, while in other embodiments, a nuclease recognizes a double-stranded target site, for example a double-stranded DNA target site. The target sites of many naturally occurring nucleases, for example, many naturally occurring DNA restriction nucleases, are well known to those of skill in the art. In many cases, a DNA nuclease, such as EcoRI, HindIII, or BamHI, 10 recognize a palindromic, double-stranded DNA target site of 4 to 10 base pairs in length, and cut each of the two DNA strands at a specific position within the target site. Some endonucleases cut a double-stranded nucleic acid target site symmetrically, *i.e.*, cutting both strands at the same position so that the ends comprise base-paired nucleotides, also referred to herein as blunt ends. Other endonucleases cut a double-stranded nucleic acid target sites asymmetrically, 15 *i.e.*, cutting each strand at a different position so that the ends comprise unpaired nucleotides. Unpaired nucleotides at the end of a double-stranded DNA molecule are also referred to as "overhangs," *e.g.*, as "5'-overhang" or as "3'-overhang," depending on whether the unpaired nucleotide(s) form(s) the 5' or the 3' end of the respective DNA strand. Double-stranded DNA molecule ends ending with unpaired nucleotide(s) are also referred to as sticky ends, as they can "stick to" other double-stranded DNA molecule ends comprising complementary unpaired nucleotide(s). A nuclease protein typically comprises a "binding domain" that mediates the interaction of the protein with the nucleic acid substrate, and 20 also, in some cases, specifically binds to a target site, and a "cleavage domain" that catalyzes the cleavage of the phosphodiester bond within the nucleic acid backbone. In some embodiments a nuclease protein can bind and cleave a nucleic acid molecule in a monomeric form, while, in other embodiments, a nuclease protein has to dimerize or multimerize in order to cleave a target nucleic acid molecule. Binding domains and cleavage domains of naturally occurring nucleases, as well as modular binding domains and cleavage domains that can be fused to create nucleases 25 binding specific target sites, are well known to those of skill in the art. For example, zinc fingers or transcriptional activator like elements can be used as binding domains to specifically bind a desired target site, and fused or conjugated to a cleavage domain, for example, the cleavage domain of FokI, to create an engineered nuclease cleaving the target site.

30 [0029] The terms "nucleic acid" and "nucleic acid molecule," as used herein, refer to a compound comprising a nucleobase and an acidic moiety, *e.g.*, a nucleoside, a nucleotide, or a polymer of nucleotides. Typically, polymeric nucleic acids, *e.g.*, nucleic acid molecules comprising three or more nucleotides are linear molecules, in which adjacent nucleotides are linked to each other via a phosphodiester linkage. In some embodiments, "nucleic acid" refers to individual nucleic acid residues (*e.g.* nucleotides and/or nucleosides). In some embodiments, "nucleic acid" refers to an oligonucleotide chain comprising three or more individual nucleotide residues. As used herein, the terms "oligonucleotide" and "polynucleotide" can be used interchangeably to refer to a polymer of nucleotides (*e.g.*, a string of at least three nucleotides). In some embodiments, "nucleic acid" encompasses RNA as well as single and/or double-stranded DNA. Nucleic acids may be naturally occurring, for example, in the context of a genome, a transcript, an mRNA, tRNA, rRNA, siRNA, snRNA, a plasmid, cosmid, chromosome, chromatid, or other naturally occurring nucleic acid molecule. On the other hand, a nucleic acid molecule may be a non-naturally occurring molecule, *e.g.*, a recombinant DNA or RNA, an artificial chromosome, an engineered genome, or fragment thereof, or a synthetic DNA, RNA, DNA/RNA hybrid, or including 35 non-naturally occurring nucleotides or nucleosides. Furthermore, the terms "nucleic acid," "DNA," "RNA," and/or similar terms include nucleic acid analogs, *i.e.* analogs having other than a phosphodiester backbone. Nucleic acids can be purified from natural sources, produced using recombinant expression systems and optionally purified, chemically synthesized, etc. Where appropriate, *e.g.*, in the case of chemically synthesized molecules, nucleic acids can comprise nucleoside analogs such as analogs having chemically modified bases or sugars, and backbone modifications. A nucleic 40 acid sequence is presented in the 5' to 3' direction unless otherwise indicated. In some embodiments, a nucleic acid is or comprises natural nucleosides (*e.g.* adenosine, thymidine, guanosine, cytidine, uridine, deoxyadenosine, deoxythymidine, deoxyguanosine, and deoxycytidine); nucleoside analogs (*e.g.*, 2-aminoadenosine, 2-thiothymidine, inosine, pyrrolo-pyrimidine, 3-methyl adenosine, 5-methylcytidine, 2-aminoadenosine, C5-bromouridine, C5-fluorouridine, C5-iodouridine, C5-propynyl-uridine, C5-propynyl-cytidine, C5-methylcytidine, 2-aminoadenosine, 7-deazaadenosine, 7-deazaguanosine, 8-oxoadenosine, 8-oxoguanosine, O(6)-methylguanine, and 2-thiocytidine); chemically modified bases; biologically modified bases (*e.g.*, methylated bases); intercalated bases; modified sugars (*e.g.*, 2'-fluororibose, ribose, 2'-deoxyribose, arabinose, and hexose); and/or modified phosphate groups (*e.g.*, phosphorothioates and 5'-N-phosphoramide linkages).

45 [0030] The term "pharmaceutical composition," as used herein, refers to a composition that can be administrated to a subject in the context of treatment of a disease or disorder. In some embodiments of the disclosure, a pharmaceutical composition comprises an active ingredient, *e.g.*, a nuclease or a nucleic acid encoding a nuclease, and a pharmaceutically acceptable excipient.

50 [0031] The term "proliferative disease," as used herein, refers to any disease in which cell or tissue homeostasis is

disturbed in that a cell or cell population exhibits an abnormally elevated proliferation rate. Proliferative diseases include hyperproliferative diseases, such as pre-neoplastic hyperplastic conditions and neoplastic diseases. Neoplastic diseases are characterized by an abnormal proliferation of cells and include both benign and malignant neoplasias. Malignant neoplasia is also referred to as cancer.

[0032] The terms "protein," "peptide," and "polypeptide" are used interchangeably herein, and refer to a polymer of amino acid residues linked together by peptide (amide) bonds. The terms refer to a protein, peptide, or polypeptide of any size, structure, or function. Typically, a protein, peptide, or polypeptide will be at least three amino acids long. A protein, peptide, or polypeptide may refer to an individual protein or a collection of proteins. One or more of the amino acids in a protein, peptide, or polypeptide may be modified, for example, by the addition of a chemical entity such as a carbohydrate group, a hydroxyl group, a phosphate group, a farnesyl group, an isofarnesyl group, a fatty acid group, a linker for conjugation, functionalization, or other modification, etc. A protein, peptide, or polypeptide may also be a single molecule or may be a multi-molecular complex. A protein, peptide, or polypeptide may be just a fragment of a naturally occurring protein or peptide. A protein, peptide, or polypeptide may be naturally occurring, recombinant, or synthetic, or any combination thereof. A protein may comprise different domains, for example, a nucleic acid binding domain and a nucleic acid cleavage domain. In some embodiments, a protein comprises a proteinaceous part, e.g., an amino acid sequence constituting a nucleic acid binding domain, and an organic compound, e.g., a compound that can act as a nucleic acid cleavage agent. In some embodiments, a protein is in a complex with, or is in association with, a nucleic acid, e.g., RNA.

[0033] The term "randomized," as used herein in the context of nucleic acid sequences, refers to a sequence or residue within a sequence that has been synthesized to incorporate a mixture of free nucleotides, for example, a mixture of all four nucleotides A, T, G, and C. Randomized residues are typically represented by the letter N within a nucleotide sequence. In some embodiments, a randomized sequence or residue is fully randomized, in which case the randomized residues are synthesized by adding equal amounts of the nucleotides to be incorporated (e.g., 25% T, 25% A, 25% G, and 25% C) during the synthesis step of the respective sequence residue. In some embodiments, a randomized sequence or residue is partially randomized, in which case the randomized residues are synthesized by adding non-equal amounts of the nucleotides to be incorporated (e.g., 79% T, 7% A, 7% G, and 7% C) during the synthesis step of the respective sequence residue. Partial randomization allows for the generation of sequences that are templated on a given sequence, but have incorporated mutations at a desired frequency. For example, if a known nuclease target site is used as a synthesis template, partial randomization in which at each step the nucleotide represented at the respective residue is added to the synthesis at 79%, and the other three nucleotides are added at 7% each, will result in a mixture of partially randomized target sites being synthesized, which still represent the consensus sequence of the original target site, but which differ from the original target site at each residue with a statistical frequency of 21% for each residue so synthesized (distributed binomially). In some embodiments, a partially randomized sequence differs from the consensus sequence by more than 5%, more than 10%, more than 15%, more than 20%, more than 25%, or more than 30% on average, distributed binomially. In some embodiments, a partially randomized sequence differs from the consensus site by no more than 10%, no more than 15%, no more than 20%, no more than 25%, nor more than 30%, no more than 40%, or no more than 50% on average, distributed binomially.

[0034] The term "RNA-programmable nuclease," and "RNA-guided nuclease" are used interchangeably herein and refer to a nuclease that forms a complex with (e.g., binds or associates with) one or more RNA that is not a target for cleavage. In some embodiments, an RNA-programmable nuclease, when in a complex with an RNA, may be referred to as a nuclease:RNA complex. Typically, the bound RNA(s) is referred to as a guide RNA (gRNA) or a single-guide RNA (sgRNA). The gRNA/sgRNA comprises a nucleotide sequence that complements a target site, which mediates binding of the nuclease/RNA complex to said target site and providing the sequence specificity of the nuclease:RNA complex. In some embodiments, the RNA-programmable nuclease is the (CRISPR-associated system) Cas9 endonuclease, for example Cas9 (Csn1) from *Streptococcus pyogenes* (see, e.g., "Complete genome sequence of an M1 strain of *Streptococcus pyogenes*." Ferretti J.J., McShan W.M., Ajdic D.J., Savic D.J., Savic G., Lyon K., Primeaux C., Sezate S., Suvorov A.N., Kenton S., Lai H.S., Lin S.P., Qian Y., Jia H.G., Najar F.Z., Ren Q., Zhu H., Song L. expand/collapse author list McLaughlin R.E., Proc. Natl. Acad. Sci. U.S.A. 98:4658-4663(2001); "CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III." Deltcheva E., Chylinski K., Sharma C.M., Gonzales K., Chao Y., Pirzada Z.A., Eckert M.R., Vogel J., Charpentier E., Nature 471:602-607(2011); and "A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity." Jinek M., Chylinski K., Fonfara I., Hauer M., Doudna J.A., Charpentier E. Science 337:816-821(2012).

[0035] Because RNA-programmable nucleases (e.g., Cas9) use RNA:DNA hybridization to determine target DNA cleavage sites, these proteins are able to cleave, in principle, any sequence specified by the guide RNA. Methods of using RNA-programmable nucleases, such as Cas9, for site-specific cleavage (e.g., to modify a genome) are known in the art (See e.g., Cong, L. et al. Multiplex genome engineering using CRISPR/Cas systems. Science 339, 819-823 (2013); Mali, P. et al. RNA-guided human genome engineering via Cas9. Science 339, 823-826 (2013); Hwang, W.Y. et al. Efficient genome editing in zebrafish using a CRISPR-Cas system. Nature biotechnology 31, 227-229 (2013);

Jinek, M. et al. RNA-programmed genome editing in human cells. *eLife* 2, e00471 (2013); Dicarlo, J.E. et al. Genome engineering in *Saccharomyces cerevisiae* using CRISPR-Cas systems. *Nucleic acids research* (2013); Jiang, W. et al. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nature biotechnology* 31, 233-239 (2013).

[0036] The terms "small molecule" and "organic compound" are used interchangeably herein and refer to molecules, whether naturally-occurring or artificially created (e.g., via chemical synthesis) that have a relatively low molecular weight.

Typically, an organic compound contains carbon. An organic compound may contain multiple carbon-carbon bonds, stereocenters, and other functional groups (e.g., amines, hydroxyl, carbonyls, or heterocyclic rings). In some embodiments, organic compounds are monomeric and have a molecular weight of less than about 1500 g/mol. In certain embodiments, the molecular weight of the small molecule is less than about 1000 g/mol or less than about 500 g/mol. In certain embodiments, the small molecule is a drug, for example, a drug that has already been deemed safe and effective for use in humans or animals by the appropriate governmental agency or regulatory body. In certain embodiments, the organic molecule is known to bind and/or cleave a nucleic acid. In some embodiments, the organic compound is an enediyne. In some embodiments, the organic compound is an antibiotic drug, for example, an anticancer antibiotic such as dynemicin, neocarzinostatin, calicheamicin, esperamicin, bleomycin, or a derivative thereof.

[0037] The term "subject," as used herein, refers to an individual organism, for example, an individual mammal. In some embodiments, the subject is a human. In some embodiments, the subject is a non-human mammal. In some embodiments, the subject is a non-human primate. In some embodiments, the subject is a rodent. In some embodiments, the subject is a sheep, a goat, a cattle, a cat, or a dog. In some embodiments, the subject is a vertebrate, an amphibian, a reptile, a fish, an insect, a fly, or a nematode.

[0038] The terms "target nucleic acid," and "target genome," as used herein in the context of nucleases, refer to a nucleic acid molecule or a genome, respectively, that comprises at least one target site of a given nuclease.

[0039] The term "target site," used herein interchangeably with the term "nuclease target site," refers to a sequence within a nucleic acid molecule that is bound and cleaved by a nuclease. A target site may be single-stranded or double-stranded. In the context of nucleases that dimerize, for example, nucleases comprising a FokI DNA cleavage domain, a target sites typically comprises a left-half site (bound by one monomer of the nuclease), a right-half site (bound by the second monomer of the nuclease), and a spacer sequence between the half sites in which the cut is made. This structure ([left-half site]-[spacer sequence]-[right-half site]) is referred to herein as an LSR structure. In some embodiments, the left-half site and/or the right-half site is between 10-18 nucleotides long. In some embodiments, either or both half-sites are shorter or longer. In some embodiments, the left and right half sites comprise different nucleic acid sequences. In the context of zinc finger nucleases, target sites may, in some embodiments comprise two half-sites that are each 6-18 bp long flanking a non-specified spacer region that is 4-8 bp long. In the context of TALENs, target sites may, in some embodiments, comprise two half-sites sites that are each 10-23 bp long flanking a non-specified spacer region that is 10-30 bp long. In the context of RNA-guided (e.g., RNA-programmable) nucleases, a target site typically comprises a nucleotide sequence that is complementary to the sgRNA of the RNA-programmable nuclease, and a protospacer adjacent motif (PAM) at the 3' end adjacent to the sgRNA-complementary sequence. For the RNA-guided nuclease Cas9, the target site may be, in some embodiments, 20 base pairs plus a 3 base pair PAM (e.g., NNN, wherein N represents any nucleotide). Typically, the first nucleotide of a PAM can be any nucleotide, while the two downstream nucleotides are specified depending on the specific RNA-guided nuclease. Exemplary target sites for RNA-guided nucleases, such as Cas9, are known to those of skill in the art and include, without limitation, NNG, NGN, NAG, and NGG, wherein N represents any nucleotide. In addition, Cas9 nucleases from different species (e.g., *S. thermophilus* instead of *S. pyogenes*) recognizes a PAM that comprises the sequence NGGNG. Additional PAM sequences are known, including, but not limited to NNAGAAW and NAAR (see, e.g., Esvelt and Wang, *Molecular Systems Biology*, 9:641 (2013)). For example, the target site of an RNA-guided nuclease, such as, e.g., Cas9, may comprise the structure [Nz]-[PAM], where each N is, independently, any nucleotide, and z is an integer between 1 and 50. In some embodiments, z is at least 2, at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, at least 15, at least 16, at least 17, at least 18, at least 19, at least 20, at least 25, at least 30, at least 35, at least 40, at least 45, or at least 50. In some embodiments, z is 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, 41, 42, 43, 44, 45, 46, 47, 48, 49, or 50. In some embodiments, Z is 20.

[0040] The term "Transcriptional Activator-Like Effector," (TALE) as used herein, refers to bacterial proteins comprising a DNA binding domain, which contains a highly conserved 33-34 amino acid sequence comprising a highly variable two-amino acid motif (Repeat Variable Diresidue, RVD). The RVD motif determines binding specificity to a nucleic acid sequence, and can be engineered according to methods well known to those of skill in the art to specifically bind a desired DNA sequence (see, e.g., Miller, Jeffrey; et.al. (February 2011). "A TALE nuclease architecture for efficient genome editing". *Nature Biotechnology* 29 (2): 143-8; Zhang, Feng; et.al. (February 2011). "Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription". *Nature Biotechnology* 29 (2): 149-53; Geißler, R.; Scholze, H.; Hahn, S.; Streubel, J.; Bonas, U.; Behrens, S. E.; Boch, J. (2011), Shiu, Shin-Han. ed. "Transcriptional Activators of Human Genes with Programmable DNA-Specificity". *PLoS ONE* 6 (5): e19509; Boch, Jens (February 2011).

"TALEs of genome targeting". *Nature Biotechnology* 29 (2): 135-6; Boch, Jens; et.al. (December 2009). "Breaking the Code of DNA Binding Specificity of TAL-Type III Effectors". *Science* 326 (5959): 1509-12; and Moscou, Matthew J.; Adam J. Bogdanove (December 2009). "A Simple Cipher Governs DNA Recognition by TAL Effectors". *Science* 326 (5959): 1501). The simple relationship between amino acid sequence and DNA recognition has allowed for the engineering of specific DNA binding domains by selecting a combination of repeat segments containing the appropriate RVDs.

[0041] The term "Transcriptional Activator-Like Element Nuclease," (TALEN) as used herein, refers to an artificial nuclease comprising a transcriptional activator like effector DNA binding domain to a DNA cleavage domain, for example, a FokI domain. A number of modular assembly schemes for generating engineered TALE constructs have been reported (see e.g., Zhang, Feng; et.al. (February 2011). "Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription". *Nature Biotechnology* 29 (2): 149-53; Geißler, R.; Scholze, H.; Hahn, S.; Streubel, J.; Bonas, U.; Behrens, S. E.; Boch, J. (2011), Shiu, Shin-Han. ed. "Transcriptional Activators of Human Genes with Programmable DNA-Specificity". *PLoS ONE* 6 (5): e19509; Cermak, T.; Doyle, E. L.; Christian, M.; Wang, L.; Zhang, Y.; Schmidt, C.; Baller, J. A.; Somia, N. V. et al. (2011). "Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting". *Nucleic Acids Research*; Morbitzer, R.; Elsaesser, J.; Hausner, J.; Lahaye, T. (2011). "Assembly of custom TALE-type DNA binding domains by modular cloning". *Nucleic Acids Research*; Li, T.; Huang, S.; Zhao, X.; Wright, D. A.; Carpenter, S.; Spalding, M. H.; Weeks, D. P.; Yang, B. (2011). "Modularly assembled designer TAL effector nucleases for targeted gene knockout and gene replacement in eukaryotes". *Nucleic Acids Research*; Weber, E.; Gruetzner, R.; Werner, S.; Engler, C.; Marillonnet, S. (2011). Bendahmane, Mohammed. ed. "Assembly of Designer TAL Effectors by Golden Gate Cloning". *PLoS ONE* 6 (5): e19722).

[0042] The terms "treatment," "treat," and "treating," refer to a clinical intervention aimed to reverse, alleviate, delay the onset of, or inhibit the progress of a disease or disorder, or one or more symptoms thereof, as described herein. As used herein, the terms "treatment," "treat," and "treating" refer to a clinical intervention aimed to reverse, alleviate, delay the onset of, or inhibit the progress of a disease or disorder, or one or more symptoms thereof, as described herein. In some embodiments, treatment may be administered after one or more symptoms have developed and/or after a disease has been diagnosed. In other embodiments, treatment may be administered in the absence of symptoms, e.g., to prevent or delay onset of a symptom or inhibit onset or progression of a disease. For example, treatment may be administered to a susceptible individual prior to the onset of symptoms (e.g., in light of a history of symptoms and/or in light of genetic or other susceptibility factors). Treatment may also be continued after symptoms have resolved, for example to prevent or delay their recurrence.

[0043] The term "zinc finger," as used herein, refers to a small nucleic acid-binding protein structural motif characterized by a fold and the coordination of one or more zinc ions that stabilize the fold. Zinc fingers encompass a wide variety of differing protein structures (see, e.g., Klug A, Rhodes D (1987). "Zinc fingers: a novel protein fold for nucleic acid recognition". *Cold Spring Harb. Symp. Quant. Biol.* 52: 473-82). Zinc fingers can be designed to bind a specific sequence of nucleotides, and zinc finger arrays comprising fusions of a series of zinc fingers, can be designed to bind virtually any desired target sequence. Such zinc finger arrays can form a binding domain of a protein, for example, of a nuclease, e.g., if conjugated to a nucleic acid cleavage domain. Different type of zinc finger motifs are known to those of skill in the art, including, but not limited to, Cys₂His₂, Gag knuckle, Treble clef, Zinc ribbon, Zn₂/Cys₆, and TAZ2 domain-like motifs (see, e.g., Krishna SS, Majumdar I, Grishin NV (January 2003). "Structural classification of zinc fingers: survey and summary". *Nucleic Acids Res.* 31 (2): 532-50). Typically, a single zinc finger motif binds 3 or 4 nucleotides of a nucleic acid molecule. Accordingly, a zinc finger domain comprising 2 zinc finger motifs may bind 6-8 nucleotides, a zinc finger domain comprising 3 zinc finger motifs may bind 9-12 nucleotides, a zinc finger domain comprising 4 zinc finger motifs may bind 12-16 nucleotides, and so forth. Any suitable protein engineering technique can be employed to alter the DNA-binding specificity of zinc fingers and/or design novel zinc finger fusions to bind virtually any desired target sequence from 3 - 30 nucleotides in length (see, e.g., Pabo CO, Peisach E, Grant RA (2001). "Design and selection of novel cys2His2 Zinc finger proteins". *Annual Review of Biochemistry* 70: 313-340; Jamieson AC, Miller JC, Pabo CO (2003). "Drug discovery with engineered zinc-finger proteins". *Nature Reviews Drug Discovery* 2 (5): 361-368; and Liu Q, Segal DJ, Ghiara JB, Barbas CF (May 1997). "Design of polydactyl zinc-finger proteins for unique addressing within complex genomes". *Proc. Natl. Acad. Sci. U.S.A.* 94 (11)). Fusions between engineered zinc finger arrays and protein domains that cleave a nucleic acid can be used to generate a "zinc finger nuclease." A zinc finger nuclease typically comprises a zinc finger domain that binds a specific target site within a nucleic acid molecule, and a nucleic acid cleavage domain that cuts the nucleic acid molecule within or in proximity to the target site bound by the binding domain. Typical engineered zinc finger nucleases comprise a binding domain having between 3 and 6 individual zinc finger motifs and binding target sites ranging from 9 base pairs to 18 base pairs in length. Longer target sites are particularly attractive in situations where it is desired to bind and cleave a target site that is unique in a given genome.

[0044] The term "zinc finger nuclease," as used herein, refers to a nuclease comprising a nucleic acid cleavage domain conjugated to a binding domain that comprises a zinc finger array. In some embodiments, the cleavage domain is the cleavage domain of the type II restriction endonuclease FokI. Zinc finger nucleases can be designed to target virtually any desired sequence in a given nucleic acid molecule for cleavage, and the possibility to the design zinc finger binding

domains to bind unique sites in the context of complex genomes allows for targeted cleavage of a single genomic site in living cells, for example, to achieve a targeted genomic alteration of therapeutic value. Targeting a double-strand break to a desired genomic locus can be used to introduce frame-shift mutations into the coding sequence of a gene due to the error-prone nature of the non-homologous DNA repair pathway. Zinc finger nucleases can be generated to target a site of interest by methods well known to those of skill in the art. For example, zinc finger binding domains with a desired specificity can be designed by combining individual zinc finger motifs of known specificity. The structure of the zinc finger protein Zif268 bound to DNA has informed much of the work in this field and the concept of obtaining zinc fingers for each of the 64 possible base pair triplets and then mixing and matching these modular zinc fingers to design proteins with any desired sequence specificity has been described (Pavletich NP, Pabo CO (May 1991). "Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A". *Science* 252 (5007): 809-17). In some embodiments, separate zinc fingers that each recognize a 3 base pair DNA sequence are combined to generate 3-, 4-, 5-, or 6-finger arrays that recognize target sites ranging from 9 base pairs to 18 base pairs in length. In some embodiments, longer arrays are contemplated. In other embodiments, 2-finger modules recognizing 6-8 nucleotides are combined to generate 4-, 6-, or 8- zinc finger arrays. In some embodiments, bacterial or phage display is employed to develop a zinc finger domain that recognizes a desired nucleic acid sequence, for example, a desired nuclease target site of 3-30 bp in length. Zinc finger nucleases, in some embodiments, comprise a zinc finger binding domain and a cleavage domain fused or otherwise conjugated to each other via a linker, for example, a polypeptide linker. The length of the linker determines the distance of the cut from the nucleic acid sequence bound by the zinc finger domain. If a shorter linker is used, the cleavage domain will cut the nucleic acid closer to the bound nucleic acid sequence, while a longer linker will result in a greater distance between the cut and the bound nucleic acid sequence. In some embodiments, the cleavage domain of a zinc finger nuclease has to dimerize in order to cut a bound nucleic acid. In some such embodiments, the dimer is a heterodimer of two monomers, each of which comprise a different zinc finger binding domain. For example, in some embodiments, the dimer may comprise one monomer comprising zinc finger domain A conjugated to a FokI cleavage domain, and one monomer comprising zinc finger domain B conjugated to a FokI cleavage domain. In this nonlimiting example, zinc finger domain A binds a nucleic acid sequence on one side of the target site, zinc finger domain B binds a nucleic acid sequence on the other side of the target site, and the dimerize FokI domain cuts the nucleic acid in between the zinc finger domain binding sites.

DETAILED DESCRIPTION

Introduction

[0045] Site-specific nucleases are powerful tools for targeted genome modification *in vitro* or *in vivo*. Some site specific nucleases can theoretically achieve a level of specificity for a target cleavage site that would allow one to target a single unique site in a genome for cleavage without affecting any other genomic site. It has been reported that nuclease cleavage in living cells triggers a DNA repair mechanism that frequently results in a modification of the cleaved, repaired genomic sequence, for example, via homologous recombination. Accordingly, the targeted cleavage of a specific unique sequence within a genome opens up new avenues for gene targeting and gene modification in living cells, including cells that are hard to manipulate with conventional gene targeting methods, such as many human somatic or embryonic stem cells. Nuclease-mediated modification of disease-related sequences, e.g., the CCR-5 allele in HIV/AIDS patients, or of genes necessary for tumor neovascularization, can be used in the clinical context, and two site specific nucleases are currently in clinical trials.

[0046] One important aspect in the field of site-specific nuclease-mediated modification are off-target nuclease effects, e.g., the cleavage of genomic sequences that differ from the intended target sequence by one or more nucleotides. Undesired side effects of off-target cleavage range from insertion into unwanted loci during a gene targeting event to severe complications in a clinical scenario. Off-target cleavage of sequences encoding essential gene functions or tumor suppressor genes by an endonuclease administered to a subject may result in disease or even death of the subject. Accordingly, it is desirable to characterize the cleavage preferences of a nuclease before using it in the laboratory or the clinic in order to determine its efficacy and safety. Further, the characterization of nuclease cleavage properties allows for the selection of the nuclease best suited for a specific task from a group of candidate nucleases, or for the selection of evolution products obtained from a plurality of nucleases. Such a characterization of nuclease cleavage properties may also inform the de-novo design of nucleases with enhanced properties, such as enhanced specificity or efficiency.

[0047] In many scenarios where a nuclease is employed for the targeted manipulation of a nucleic acid, cleavage specificity is a crucial feature. The imperfect specificity of some engineered nuclease binding domains can lead to off-target cleavage and undesired effects both *in vitro* and *in vivo*. Current methods of evaluating site-specific nuclease specificity, including ELISA assays, microarrays, one-hybrid systems, SELEX, and its variants, and Rosetta-based computational predictions, are all premised on the assumption that the binding specificity of the nuclease is equivalent

or proportionate to their cleavage specificity.

[0048] It was previously discovered that the prediction of nuclease off-target binding effects constitute an imperfect approximation of a nuclease's off-target cleavage effects that may result in undesired biological effects (see PCT Application WO 2013/066438; and Pattanayak, V., Ramirez, C.L., Joung, J.K. & Liu, D.R. Revealing off-target cleavage specificities of zinc-finger nucleases by *in vitro* selection. *Nature methods* 8, 765-770 (2011)). This finding was consistent with the notion that the reported toxicity of some site specific DNA nucleases results from off-target DNA cleavage, rather than off-target binding alone.

[0049] The methods and reagents of the present disclosure represent, in some aspects, an improvement over previous methods and allow for an accurate evaluation of a given nuclease's target site specificity and provide strategies for the selection of suitable unique target sites and the design or selection of highly specific nucleases for the targeted cleavage of a single site in the context of a complex genome. For example, some previously reported methods for determining nuclease target site specificity profiles by screening libraries of nucleic acid molecules comprising candidate target sites relied on a "two-cut" *in vitro* selection method which requires indirect reconstruction of target sites from sequences of two half-sites resulting from two adjacent cuts of the nuclease of a library member nucleic acid (see e.g., Pattanayak, V. et al., *Nature Methods* 8, 765-770 (2011)). In contrast to such "two-cut" strategies, the methods of the present disclosure utilize a "one cut" screening strategy, which allows for the identification of library members that have been cut at least once by the nuclease. The "one-cut" selection strategies provided herein are compatible with single end high-throughput sequencing methods and do not require computational reconstruction of cleaved target sites from cut half-sites because they feature, in some embodiments, direct sequencing of an intact target nuclease sequence in a cut library member nucleic acid.

[0050] Additionally, the presently disclosed "one-cut" screening methods utilize concatemers of a candidate nuclease target site and constant insert region that are about 10-fold shorter than previously reported constructs used for two-cut strategies (~50bp repeat sequence length versus ~500bp repeat sequence length in previous reports). This difference in repeat sequence length in the concatemers of the library allows for the generation of highly complex libraries of candidate nuclease target sites, e.g., of libraries comprising 10^{12} different candidate nuclease target sequences. As described herein, an exemplary library of such complexity has been generated, templated on a known Cas9 nuclease target site by varying the sequence of the known target site. The exemplary library demonstrated that a greater than 10-fold coverage of all sequences with eight or fewer mutations of the known target site can be achieved using the strategies provided herein. The use of a shorter repeat sequence also allows the use of single-end sequencing, since both a cut half-site and an adjacent uncut site of the same library member are contained within a 100 nucleotide sequencing read.

[0051] The strategies, methods, libraries, and reagents provided herein can be utilized to analyze the sequence preferences and specificity of any site-specific nuclease, for example, to Zinc Finger Nucleases (ZFNs), Transcription Activator-Like Effector Nucleases (TALENs), homing endonucleases, organic compound nucleases, and enediyne antibiotics (e.g., dynemicin, neocarzinostatin, calicheamicin, esperamicin, bleomycin). Suitable nucleases in addition to the ones described herein will be apparent to those of skill in the art based on this disclosure.

[0052] Further, the methods, reagents, and strategies provided herein allow those of skill in the art to identify, design, and/or select nucleases with enhanced specificity and minimize the off-target effects of any given nuclease (e.g., site-specific nucleases such as ZFNs, and TALENs which produce cleavage products with sticky ends, as well as RNA-programmable nucleases, for example Cas9, which produce cleavage products having blunt ends). While of particular relevance to DNA and DNA-cleaving nucleases, the inventive concepts, methods, strategies, and reagents provided herein are not limited in this respect, but can be applied to any nucleic acid:nuclease pair.

Identifying Nuclease Target Sites Cleaved by a Site-Specific Nuclease

[0053] Some aspects of this disclosure provide improved methods and reagents to determine the nucleic acid target sites cleaved by any site-specific nuclease. The methods provided herein can be used for the evaluation of target site preferences and specificity of both nucleases that create blunt ends and nucleases that create sticky ends. In general, such methods comprise contacting a given nuclease with a library of target sites under conditions suitable for the nuclease to bind and cut a target site, and determining which target sites the nuclease actually cuts. A determination of a nuclease's target site profile based on actual cutting has the advantage over methods that rely on binding in that it measures a parameter more relevant for mediating undesired off-target effects of site-specific nucleases. In general, the methods provided herein comprise ligating an adapter of a known sequence to nucleic acid molecules that have been cut by a nuclease of interest via 5'-phosphate-dependent ligation. Accordingly, the methods provided herein are particularly useful for identifying target sites cut by nucleases that leave a phosphate moiety at the 5'-end of the cut nucleic acid strand when cleaving their target site. After ligating an adapter to the 5'-end of a cut nucleic acid strand, the cut strand can directly be sequenced using the adapter as a sequencing linker, or a part of the cut library member concatemer comprising an intact target site identical to the cut target site can be amplified via PCR and the amplification product can then be sequenced.

[0054] The method of the first aspect of the invention comprises (a) providing a nuclease that cuts a double-stranded nucleic acid target site, wherein cutting of the target site results in cut nucleic acid strands comprising a 5'-phosphate moiety; (b) contacting the nuclease of (a) with a library of candidate nucleic acid molecules, wherein each nucleic acid molecule comprises a concatemer of a sequence comprising a candidate nuclease target site and a constant insert sequence of at least 10 and not more than 80 nucleotides, under conditions suitable for the nuclease to cut a candidate nucleic acid molecule comprising a target site of the nuclease; and (c) identifying nuclease target sites cut by the nuclease in (b) by determining the sequence of an uncut nuclease target site on the nucleic acid strand that was cut by the nuclease in step (b)), wherein the determining of step (c) comprises (i) ligating a first nucleic acid adapter to the 5' end of a nucleic acid strand that was cut by the nuclease in step (b) via 5'-phosphate-dependent ligation, and (ii) amplifying a fragment of the concatemer cut by the nuclease that comprises an uncut target site via a PCR reaction using a PCR primer that hybridizes with the adapter and a PCR primer that hybridizes with the constant insert sequence.

[0055] Disclosed herein is a method comprising providing a nuclease and contacting the nuclease with a library of candidate nucleic acid molecules comprising candidate target sites. In some embodiments, the candidate nucleic acid molecules are double-stranded nucleic acid molecules. In some embodiments, the candidate nucleic acid molecules are DNA molecules. In some embodiments, each nucleic acid molecule in the library comprises a concatemer of a sequence comprising a candidate nuclease target site and a constant insert sequence. For example, in some embodiments, the library comprises nucleic acid molecules that comprise the structure $R_1-[(candidate\ nuclease\ target\ site)-(constant\ insert\ sequence)]_n - R_2$, wherein R_1 and R_2 are, independently, nucleic acid sequences that may comprise a fragment of the [(candidate nuclease target site)-(constant insert sequence)] structure, and n is an integer between 2 and y . In some embodiments, y is at least 10^1 , at least 10^2 , at least 10^3 , at least 10^4 , at least 10^5 , at least 10^6 , at least 10^7 , at least 10^8 , at least 10^9 , at least 10^{10} , at least 10^{11} , at least 10^{12} , at least 10^{13} , at least 10^{14} , or at least 10^{15} . In some embodiments, y is less than 10^2 , less than 10^3 , less than 10^4 , less than 10^5 , less than 10^6 , less than 10^7 , less than 10^8 , less than 10^9 , less than 10^{10} , less than 10^{11} , less than 10^{12} , less than 10^{13} , less than 10^{14} , or less than 10^{15} .

[0056] For example, in some embodiments, the candidate nucleic acid molecules of the library comprise a candidate nuclease target site of the structure [(Nz)-(PAM)], and, thus, the nucleic acid molecules of the library comprise the structure $R_1-[(Nz)-(PAM)-(constant\ region)]_X-R_2$, wherein R_1 and R_2 are, independently, nucleic acid sequences that may comprise a fragment of the [(Nz)-(PAM)-(constant region)] repeat unit; each N represents, independently, any nucleotide; Z is an integer between 1 and 50; and X is an integer between 2 and y . In some embodiments, y is at least 10^1 , at least 10^2 , at least 10^3 , at least 10^4 , at least 10^5 , at least 10^6 , at least 10^7 , at least 10^8 , at least 10^9 , at least 10^{10} , at least 10^{11} , at least 10^{12} , at least 10^{13} , at least 10^{14} , or at least 10^{15} . In some embodiments, y is less than 10^2 , less than 10^3 , less than 10^4 , less than 10^5 , less than 10^6 , less than 10^7 , less than 10^8 , less than 10^9 , less than 10^{10} , less than 10^{11} , less than 10^{12} , less than 10^{13} , less than 10^{14} , or less than 10^{15} . In some embodiments, Z is at least 2, at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, at least 15, at least 16, at least 17, at least 18, at least 19, at least 20, at least 25, at least 30, at least 35, at least 40, at least 45, or at least 50. In some embodiments, Z is 20. Each N represents, independently, any nucleotide. Accordingly, a sequence provided as Nz with z=2 would be NN, with each N, independently, representing A, T, G, or C. Accordingly, Nz with z=2 can represent AA, AT, AG, AC, TA, TT, TG, TC, GA, GT, GG, GC, CA, CT, CG, and CC.

[0057] In other embodiments, the candidate nucleic acid molecules of the library comprise a candidate nuclease target site of the structure [left-half site]-[spacer sequence]-[right-half site] ("LSR"), and, thus, the nucleic acid molecules of the library comprise the structure $R_1-[(LSR)-(constant\ region)]_X-R_2$, wherein R_1 and R_2 are, independently, nucleic acid sequences that may comprise a fragment of the [(LSR)-(constant region)] repeat unit, and X is an integer between 2 and y . In some embodiments, y is at least 10^1 , at least 10^2 , at least 10^3 , at least 10^4 , at least 10^5 , at least 10^6 , at least 10^7 , at least 10^8 , at least 10^9 , at least 10^{10} , at least 10^{11} , at least 10^{12} , at least 10^{13} , at least 10^{14} , or at least 10^{15} . In some embodiments, y is less than 10^2 , less than 10^3 , less than 10^4 , less than 10^5 , less than 10^6 , less than 10^7 , less than 10^8 , less than 10^9 , less than 10^{10} , less than 10^{11} , less than 10^{12} , less than 10^{13} , less than 10^{14} , or less than 10^{15} . The constant region, in some embodiments, is of a length that allows for efficient self-ligation of a single repeat unit. Suitable lengths will be apparent to those of skill in the art. For example, in some embodiments, the constant region is between 5 and 100 base pairs long, for example, about 5 base pairs, about 10 base pairs, about 15 base pairs, about 20 base pairs, about 25 base pairs, about 30 base pairs, about 35 base pairs, about 40 base pairs, about 50 base pairs, about 60 base pairs, about 70 base pairs, about 80 base pairs, about 90 base pairs, or about 100 base pairs long. In some embodiments, the constant region is 16 base pairs long. In some embodiments, the nuclease cuts a double-stranded nucleic acid target site and creates blunt ends. In other embodiments, the nuclease creates a 5'-overhang. In some such embodiments, the target site comprises a [left-half site]-[spacer sequence]-[right-half site] (LSR) structure, and the nuclease cuts the target site within the spacer sequence.

[0058] In some embodiments, a nuclease cuts a double-stranded target site and creates blunt ends. In some embodiments, a nuclease cuts a double-stranded target site and creates an overhang, or sticky end, for example, a 5'-overhang. In some such embodiments, the method comprises filling in the 5'-overhangs of nucleic acid molecules produced from a nucleic acid molecule that has been cut once by the nuclease, wherein the nucleic acid molecules comprise a constant

insert sequence flanked by a left or right half-site and cut spacer sequence on one side, and an uncut target site sequence on the other side, thereby creating blunt ends.

[0059] According to the invention, the determining of step (c) comprises ligating a first nucleic acid adapter to the 5' end of a nucleic acid strand that was cut by the nuclease in step (b) via 5'-phosphate-dependent ligation. In some embodiments, the nuclease creates blunt ends. In such embodiments, an adapter can directly be ligated to the blunt ends resulting from the nuclease cut of the target site by contacting the cut library members with a double-stranded, blunt-ended adapter lacking 5' phosphorylation. In some embodiments, the nuclease creates an overhang (sticky end). In some such embodiments, an adapter may be ligated to the cut site by contacting the cut library member with an excess of adapter having a compatible sticky end. If a nuclease is used that cuts within a constant spacer sequence between variable half-sites, the sticky end can be designed to match the 5' overhang created from the spacer sequence. In embodiments, where the nuclease cuts within a variable sequence, a population of adapters having a variable overhang sequence and a constant annealed sequence (for use as a sequencing linker or PCR primer) may be used, or the 5' overhangs may be filled in to form blunt ends before adapter ligation.

[0060] According to the invention, the determining of step (c) further comprises amplifying a fragment of the concatemer cut by the nuclease that comprises an uncut target site via PCR using a PCR primer that hybridizes with the adapter and a PCR primer that hybridizes with the constant insert sequence. Typically, the amplification of concatemers via PCR will yield amplicons comprising at least one intact candidate target site identical to the cut target sites because the target sites in each concatemer are identical. For single-direction sequencing, an enrichment of amplicons that comprise one intact target site, no more than two intact target sites, no more than three intact target sites, no more than four intact target sites, or no more than five intact target sites may be desirable. In embodiments where PCR is used for amplification of cut nucleic acid molecules, the PCR parameters can be optimized to favor the amplification of short sequences and disfavor the amplification of longer sequences, e.g., by using a short elongation time in the PCR cycle. Another possibility for enrichment of short amplicons is size fractionation, e.g., via gel electrophoresis or size exclusion chromatography. Size fractionation can be performed before and/or after amplification. Other suitable methods for enrichment of short amplicons will be apparent to those of skill in the art and the disclosure is not limited in this respect.

[0061] In some embodiments, the determining of step (c) comprises sequencing the nucleic acid strand that was cut by the nuclease in step (b), or a copy thereof obtained via amplification, e.g., by PCR. Sequencing methods are well known to those of skill in the art. The disclosure is not limited in this respect.

[0062] In some embodiments, the nuclease being profiled using the disclosed system is an RNA-programmable nuclease that forms a complex with an RNA molecule, and wherein the nuclease:RNA complex specifically binds a nucleic acid sequence complementary to the sequence of the RNA molecule. In some embodiments, the RNA molecule is a single-guide RNA (sgRNA). In some embodiments, the sgRNA comprises 5-50 nucleotides, 10-30 nucleotides, 15-25 nucleotides, 18-22 nucleotides, 19-21 nucleotides, e.g., 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, or 30 nucleotides. In some embodiments, the sgRNA comprises 5-50 nucleotides, 10-30 nucleotides, 15-25 nucleotides, 18-22 nucleotides, 19-21 nucleotides, e.g., 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, or 30 nucleotides that are complementary to a sequence of the nuclease target site. In some embodiments, the sgRNA comprises 20 nucleotides that are complementary to the nuclease target site. In some embodiments, the nuclease is a Cas9 nuclease. In some embodiments, the nuclease target site comprises a [sgRNA-complementary sequence]-[protospacer adjacent motif (PAM)] structure, and the nuclease cuts the target site within the sgRNA-complementary sequence. In some embodiments, the sgRNA-complementary sequence comprises 5-50 nucleotides, 10-30 nucleotides, 15-25 nucleotides, 18-22 nucleotides, 19-21 nucleotides, e.g., 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, or 30 nucleotides.

[0063] In some embodiments, the RNA-programmable nuclease is a Cas9 nuclease. The RNA-programmable Cas9 endonuclease cleaves double-stranded DNA (dsDNA) at sites adjacent to a two-base-pair PAM motif and complementary to a guide RNA sequence (sgRNA). Typically, the sgRNA sequence that is complementary to the target site sequence is about 20 nucleotides long, but shorter and longer complementary sgRNA sequences can be used as well. For example, in some embodiments, the sgRNA comprises 5-50 nucleotides, 10-30 nucleotides, 15-25 nucleotides, 18-22 nucleotides, 19-21 nucleotides, e.g., 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, or 30 nucleotides. The Cas9 system has been used to modify genomes in multiple cell types, demonstrating its potential as a facile genome-engineering tool.

[0064] In some embodiments, the nuclease comprises an unspecific nucleic acid cleavage domain. In some embodiments, the nuclease comprises a FokI cleavage domain. In some embodiments, the nuclease comprises a nucleic acid cleavage domain that cleaves a target sequence upon cleavage domain dimerization. In some embodiments, the nuclease comprises a binding domain that specifically binds a nucleic acid sequence. In some embodiments, the binding domain comprises a zinc finger. In some embodiments, the binding domain comprises at least 2, at least 3, at least 4, or at least 5 zinc fingers. In some embodiments, the nuclease is a Zinc Finger Nuclease. In some embodiments, the binding domain comprises a Transcriptional Activator-Like Element. In some embodiments, the nuclease is a Transcriptional Activator-Like Element Nuclease (TALEN). In some embodiments, the nuclease is a homing endonuclease. In some embodiments,

the nuclease is an organic compound. In some embodiments, the nuclease comprises an enediyne functional group. In some embodiments, the nuclease is an antibiotic. In some embodiments, the compound is dynemicin, neocarzinostatin, calicheamicin, esperamicin, bleomycin, or a derivative thereof.

[0065] Incubation of the nuclease with the library nucleic acids will result in cleavage of those concatemers in the library that comprise target sites that can be bound and cleaved by the nuclease. If a given nuclease cleaves a specific target site with high efficiency, a concatemer comprising target sites will be cut, e.g., once or multiple times, resulting in the generation of fragments comprising a cut target site adjacent to one or more repeat units. Depending on the structure of the library members, an exemplary cut nucleic acid molecule released from a library member concatemer by a single nuclease cleavage may, for example, be of the structure (cut target site)-(constant region)-[(target site)-(constant region)]_X-R₂. For example, in the context of an RNA-guided nuclease, an exemplary cut nucleic acid molecule released from a library member concatemer by a single nuclease cleavage may, for example, be of the structure (PAM)-(constant region)-[(Nz)-(PAM)-(constant region)]_X-R₂. And in the context of a nuclease cutting an LSR structure within the spacer region, an exemplary cut nucleic acid molecule released from a library member concatemer by a single nuclease cleavage may, for example, be of the structure (cut spacer region)-(right half site)-(constant region)-[(LSR)-(constant region)]_X-R₂. Such cut fragments released from library candidate molecules can then be isolated and/or the sequence of the target site cleaved by the nuclease identified by sequencing an intact target site (e.g., an intact (Nz)-(PAM) site of released repeat units. See, e.g., Figure 1B for an illustration.

[0066] Suitable conditions for exposure of the library of nucleic acid molecules will be apparent to those of skill in the art. In some embodiments, suitable conditions do not result in denaturation of the library nucleic acids or the nuclease and allow for the nuclease to exhibit at least 50%, at least 60%, at least 70%, at least 80%, at least 90%, at least 95%, or at least 98% of its nuclease activity.

[0067] Additionally, if a given nuclease cleaves a specific target site, some cleavage products will comprise a cut half site and an intact, or uncut target site. As described herein, such products can be isolated by routine methods, and because the insert sequence, in some aspects, is less than 100 base pairs, such isolated cleavage products may be sequenced in a single read-through, allowing identification of the target site sequence without reconstructing the sequence, e.g., from cut half sites.

[0068] Any method suitable for isolation and sequencing of the repeat units can be employed to elucidate the LSR sequence cleaved by the nuclease. For example, since the length of the constant region is known, individual released repeat units can be separated based on their size from the larger uncut library nucleic acid molecules as well as from fragments of library nucleic acid molecules that comprise multiple repeat units (indicating non-efficient targeted cleavage by the nuclease). Suitable methods for separating and/or isolating nucleic acid molecules based on their size are well-known to those of skill in the art and include, for example, size fractionation methods, such as gel electrophoresis, density gradient centrifugation, and dialysis over a semi-permeable membrane with a suitable molecular cutoff value. The separated/isolated nucleic acid molecules can then be further characterized, for example, by ligating PCR and/or sequencing adapters to the cut ends and amplifying and/or sequencing the respective nucleic acids. Further, if the length of the constant region is selected to favor self-ligation of individual released repeat units, such individual released repeat units may be enriched by contacting the nuclease treated library molecules with a ligase and subsequent amplification and/or sequencing based on the circularized nature of the self-ligated individual repeat units.

[0069] In some embodiments, where a nuclease is used that generates 5'-overhangs as a result of cutting a target nucleic acid, the 5'-overhangs of the cut nucleic acid molecules are filled in. Methods for filling in 5'-overhangs are well known to those of skill in the art and include, for example, methods using DNA polymerase I Klenow fragment lacking exonuclease activity (Klenow (3'→5' exo-)). Filling in 5'-overhangs results in the overhang-templated extension of the recessed strand, which, in turn, results in blunt ends. In the case of single repeat units released from library concatemers, the resulting structure is a blunt-ended S₂'R-(constant region)-LSi', with S₁' and S₂' comprising blunt ends. PCR and/or sequencing adapters can then be added to the ends by blunt end ligation and the respective repeat units (including S₂'R and LS₁' regions) can be sequenced. From the sequence data, the original LSR region can be deduced. Blunting of the overhangs created during the nuclease cleavage process also allows for distinguishing between target sites that were properly cut by the respective nuclease and target sites that were non-specifically cut, e.g., based on non-nuclease effects such as physical shearing. Correctly cleaved nuclease target sites can be recognized by the existence of complementary S₂'R and LS₁' regions, which comprise a duplication of the overhang nucleotides as a result of the overhang fill in while target sites that were not cleaved by the respective nuclease are unlikely to comprise overhang nucleotide duplications. In some embodiments, the method comprises identifying the nuclease target site cut by the nuclease by determining the sequence of the left-half site, the right-half-site, and/or the spacer sequence of a released individual repeat unit. Any suitable method for amplifying and/or sequencing can be used to identify the LSR sequence of the target site cleaved by the respective nuclease. Methods for amplifying and/or sequencing nucleic acids are well known to those of skill in the art and the disclosure is not limited in this respect. In the case of nucleic acids released from library concatemers that comprise a cut half site and an uncut target site (e.g., comprises at least about 1.5 repeat sequences), filling in the 5'-overhangs also provides for assurance that the nucleic acid was cleaved by the nuclease. Because the

nucleic acid also comprises an intact, or uncut target site, the sequence of said site can be determined without having to reconstruct the sequence from a left-half site, right-half site, and/or spacer sequence.

[0070] Some of the methods and strategies provided herein allow for the simultaneous assessment of a plurality of candidate target sites as possible cleavage targets for any given nuclease. Accordingly, the data obtained from such methods can be used to compile a list of target sites cleaved by a given nuclease, which is also referred to herein as a target site profile. If a sequencing method is used that allows for the generation of quantitative sequencing data, it is also possible to record the relative abundance of any nuclease target site detected to be cleaved by the respective nuclease. Target sites that are cleaved more efficiently by the nuclease will be detected more frequently in the sequencing step, while target sites that are not cleaved efficiently will only rarely release an individual repeat unit from a candidate concatemer, and thus, will only generate few, if any, sequencing reads. Such quantitative sequencing data can be integrated into a target site profile to generate a ranked list of highly preferred and less preferred nuclease target sites.

[0071] The methods and strategies of nuclease target site profiling provided herein can be applied to any site-specific nuclease, including, for example, ZFNs, TALENs, homing endonucleases, and RNA-programmable nucleases, such as Cas9 nucleases. As described in more detail herein, nuclease specificity typically decreases with increasing nuclease concentration, and the methods described herein can be used to determine a concentration at which a given nuclease efficiently cuts its intended target site, but does not efficiently cut any off-target sequences. In some embodiments, a maximum concentration of a therapeutic nuclease is determined at which the therapeutic nuclease cuts its intended nuclease target site but does not cut more than 10, more than 5, more than 4, more than 3, more than 2, more than 1, or any additional sites. A therapeutic nuclease can be administered to a subject in an amount effective to generate a final concentration equal or lower than the maximum concentration determined as described above.

[0072] In some embodiments, the library of candidate nucleic acid molecules used in the methods provided herein comprises at least 10^8 , at least 10^9 , at least 10^{10} , at least 10^{11} , or at least 10^{12} different candidate nuclease target sites.

[0073] In some embodiments, the nuclease is a therapeutic nuclease which cuts a specific nuclease target site in a gene associated with a disease. In some embodiments, the method further comprises determining a maximum concentration of the therapeutic nuclease at which the therapeutic nuclease cuts the specific nuclease target site and does not cut more than 10, more than 5, more than 4, more than 3, more than 2, more than 1, or no additional sites. The therapeutic nuclease can be administered to a subject in an amount effective to generate a final concentration equal or lower than the maximum concentration.

30 *Nuclease Target Site Libraries*

[0074] Some embodiments of this disclosure provide libraries of nucleic acid molecules for nuclease target site profiling. In some embodiments, the candidate nucleic acid molecules of the library comprise the structure $R_1-[N_z]-(PAM)-(constant\ region)]_x-R_2$, wherein R_1 and R_2 are, independently, nucleic acid sequences that may comprise a fragment of the $[(N_z)-(PAM)-(constant\ region)]$ repeat unit; each N represents, independently, any nucleotide; Z is an integer between 1 and 50; and X is an integer between 2 and y . In some embodiments, y is at least 10^1 , at least 10^2 , at least 10^3 , at least 10^4 , at least 10^5 , at least 10^6 , at least 10^7 , at least 10^8 , at least 10^9 , at least 10^{10} , at least 10^{11} , at least 10^{12} , at least 10^{13} , at least 10^{14} , or at least 10^{15} . In some embodiments, y is less than 10^2 , less than 10^3 , less than 10^4 , less than 10^5 , less than 10^6 , less than 10^7 , less than 10^8 , less than 10^9 , less than 10^{10} , less than 10^{11} , less than 10^{12} , less than 10^{13} , less than 10^{14} , or less than 10^{15} . In some embodiments, Z is at least 2, at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, at least 15, at least 16, at least 17, at least 18, at least 19, at least 20, at least 25, at least 30, at least 35, at least 40, at least 45, or at least 50. In some embodiments, Z is 20. Each N represents, independently, any nucleotide. Accordingly, a sequence provided as N_z with $z=2$ would be NN, with each N , independently, representing A, T, G, or C. Accordingly, N_z with $z=2$ can represent AA, AT, AG, AC, TA, TT, TG, TC, GA, GT, GG, GC, CA, CT, CG, and CC.

[0075] In some embodiments of this disclosure, a library is provided comprising candidate nucleic acid molecules that comprise target sites with a partially randomized left-half site, a partially randomized right-half site, and/or a partially randomized spacer sequence. In some embodiments, the library is provided comprising candidate nucleic acid molecules that comprise target sites with a partially randomized left half site, a fully randomized spacer sequence, and a partially randomized right half site. In some embodiments, a library is provided comprising candidate nucleic acid molecules that comprise target sites with a partially or fully randomized sequence, wherein the target sites comprise the structure $[N_z-(PAM)]$, for example as described herein. In some embodiments, partially randomized sites differ from the consensus site by more than 5%, more than 10%, more than 15%, more than 20%, more than 25%, or more than 30% on average, distributed binomially.

[0076] In some embodiments of this disclosure such a library comprises a plurality of nucleic acid molecules, each comprising a concatemer of a candidate nuclease target site and a constant insert sequence, also referred to herein as a constant region. For example, in some embodiments, the candidate nucleic acid molecules of the library comprise the structure $R_1-(sgRNA-complementary\ sequence)-(PAM)-(constant\ region)]_x-R_2$, or the structure $R_1-[(LSR)-(constant$

region)]_X-R₂, wherein the structure in square brackets ("[...]"') is referred to as a repeat unit or repeat sequence; R₁ and R₂ are, independently, nucleic acid sequences that may comprise a fragment of the repeat unit, and X is an integer between 2 and y. In some embodiments, y is at least 10¹, at least 10², at least 10³, at least 10⁴, at least 10⁵, at least 10⁶, at least 10⁷, at least 10⁸, at least 10⁹, at least 10¹⁰, at least 10¹¹, at least 10¹², at least 10¹³, at least 10¹⁴, or at least 10¹⁵. In some embodiments, y is less than 10², less than 10³, less than 10⁴, less than 10⁵, less than 10⁶, less than 10⁷, less than 10⁸, less than 10⁹, less than 10¹⁰, less than 10¹¹, less than 10¹², less than 10¹³, less than 10¹⁴, or less than 10¹⁵. The constant region, in some embodiments, is of a length that allows for efficient self-ligation of a single repeat unit. In some embodiments, the constant region is of a length that allows for efficient separation of single repeat units from fragments comprising two or more repeat units. In some embodiments, the constant region is of a length that allows for efficient sequencing of a complete repeat unit in one sequencing read. Suitable lengths will be apparent to those of skill in the art. For example, in some embodiments, the constant region is between 5 and 100 base pairs long, for example, about 5 base pairs, about 10 base pairs, about 15 base pairs, about 20 base pairs, about 25 base pairs, about 30 base pairs, about 35 base pairs, about 40 base pairs, about 50 base pairs, about 60 base pairs, about 70 base pairs, about 80 base pairs, about 90 base pairs, or about 100 base pairs long. In some embodiments, the constant region is 1, 2, 3, 4, 5, 6, 7, 8, 9, 0, 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, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, or 80 base pairs long.

[0077] An LSR site typically comprises a [left-half site]-[spacer sequence]-[right-half site] structure. The lengths of the half-site and the spacer sequence will depend on the specific nuclease to be evaluated. In general, the half-sites will be 6-30 nucleotides long, and preferably 10-18 nucleotides long. For example, each half site individually may be 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, or 30 nucleotides long. In some embodiments, an LSR site may be longer than 30 nucleotides. In some embodiments, the left half site and the right half site of an LSR are of the same length. In some embodiments, the left half site and the right half site of an LSR are of different lengths. In some embodiments, the left half site and the right half site of an LSR are of different sequences. In some embodiments, a library is provided that comprises candidate nucleic acids which comprise LSRs that can be cleaved by a FokI cleavage domain, a Zinc Finger Nuclease (ZFN), a Transcription Activator-Like Effector Nuclease (TALEN), a homing endonuclease, or an organic compound (e.g., an enediyne antibiotic such as dynemicin, neocarzinostatin, calicheamicin, and esperamicin; and bleomycin).

[0078] In some embodiments of this disclosure, a library of candidate nucleic acid molecules is provided that comprises at least 10⁵, at least 10⁶, at least 10⁷, at least 10⁸, at least 10⁹, at least 10¹⁰, at least 10¹¹, at least 10¹², at least 10¹³, at least 10¹⁴, or at least 10¹⁵ different candidate nuclease target sites. In some embodiments, the candidate nucleic acid molecules of the library are concatemers produced from a secularized templates by rolling cycle amplification. In some embodiments, the library comprises nucleic acid molecules, e.g., concatemers, of a molecular weight of at least 5 kDa, at least 6 kDa, at least 7 kDa, at least 8 kDa, at least 9 kDa, at least 10 kDa, at least 12 kDa, or at least 15kDa. In some embodiments, the molecular weight of the nucleic acid molecules within the library may be larger than 15 kDa. In some embodiments, the library comprises nucleic acid molecules within a specific size range, for example, within a range of 5-7 kDa, 5-10 kDa, 8-12 kDa, 10-15 kDa, or 12-15 kDa, or 5-10 kDa or any possible subrange. While some methods suitable for generating nucleic acid concatemers according to some aspects of this disclosure result in the generation of nucleic acid molecules of greatly different molecular weights, such mixtures of nucleic acid molecules may be size fractionated to obtain a desired size distribution. Suitable methods for enriching nucleic acid molecules of a desired size or excluding nucleic acid molecules of a desired size are well known to those of skill in the art and the disclosure is not limited in this respect.

[0079] In some embodiments, partially randomized sites differ from the consensus site by no more than 10%, no more than 15%, no more than 20%, no more than 25%, nor more than 30%, no more than 40%, or no more than 50% on average, distributed binomially. For example, in some embodiments partially randomized sites differ from the consensus site by more than 5%, but by no more than 10%; by more than 10%, but by no more than 20%; by more than 20%, but by no more than 25%; by more than 5%, but by no more than 20%, and so on. Using partially randomized nuclease target sites in the library is useful to increase the concentration of library members comprising target sites that are closely related to the consensus site, for example, that differ from the consensus sites in only one, only two, only three, only four, or only five residues. The rationale behind this is that a given nuclease, for example a given ZFN or RNA-programmable nuclease, is likely to cut its intended target site and any closely related target sites, but unlikely to cut a target sites that is vastly different from or completely unrelated to the intended target site. Accordingly, using a library comprising partially randomized target sites can be more efficient than using libraries comprising fully randomized target sites without compromising the sensitivity in detecting any off-target cleavage events for any given nuclease. Thus, the use of partially randomized libraries significantly reduces the cost and effort required to produce a library having a high likelihood of covering virtually all off-target sites of a given nuclease. In some embodiments however it may be desirable to use a fully randomized library of target sites, for example, in embodiments, where the specificity of a given nuclease is to be evaluated in the context of any possible site in a given genome.

Selection and Design of Site-Specific Nucleases

[0080] Some aspects of this disclosure provide methods and strategies for selecting and designing site-specific nucleases that allow the targeted cleavage of a single, unique sites in the context of a complex genome. In some embodiments, a method is provided that comprises providing a plurality of candidate nucleases that are designed or known to cut the same consensus sequence; profiling the target sites actually cleaved by each candidate nuclease, thus detecting any cleaved off-target sites (target sites that differ from the consensus target site); and selecting a candidate nuclease based on the off-target site(s) so identified. In some embodiments, this method is used to select the most specific nuclease from a group of candidate nucleases, for example, the nuclease that cleaves the consensus target site with the highest specificity, the nuclease that cleaves the lowest number of off-target sites, the nuclease that cleaves the lowest number of off-target sites in the context of a target genome, or a nuclease that does not cleave any target site other than the consensus target site. In some embodiments, this method is used to select a nuclease that does not cleave any off-target site in the context of the genome of a subject at concentration that is equal to or higher than a therapeutically effective concentration of the nuclease.

[0081] The methods and reagents provided herein can be used, for example, to evaluate a plurality of different nucleases targeting the same intended targets site, for example, a plurality of variations of a given site-specific nuclease, for example a given zinc finger nuclease. Accordingly, such methods may be used as the selection step in evolving or designing a novel site-specific nucleases with improved specificity.

20 *Identifying Unique Nuclease Target Sites within a Genome*

[0082] Some embodiments of this disclosure provide a method for selecting a nuclease target site within a genome. As described in more detail elsewhere herein, it was surprisingly discovered that off target sites cleaved by a given nuclease are typically highly similar to the consensus target site, e.g., differing from the consensus target site in only one, only two, only three, only four, or only five nucleotide residues. Based on this discovery, a nuclease target sites within the genome can be selected to increase the likelihood of a nuclease targeting this site not cleaving any off target sites within the genome. For example, in some embodiments, a method is provided that comprises identifying a candidate nuclease target site; and comparing the candidate nuclease target site to other sequences within the genome. Methods for comparing candidate nuclease target sites to other sequences within the genome are well known to those of skill in the art and include for example sequence alignment methods, for example, using a sequence alignment software or algorithm such as BLAST on a general purpose computer. A suitable unique nuclease target site can then be selected based on the results of the sequence comparison. In some embodiments, if the candidate nuclease target site differs from any other sequence within the genome by at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, or at least 10 nucleotides, the nuclease target site is selected as a unique site within the genome, whereas if the site does not fulfill this criteria, the site may be discarded. In some embodiments, once a site is selected based on the sequence comparison, as outlined above, a site-specific nuclease targeting the selected site is designed. For example, a zinc finger nuclease may be designed to target any selected nuclease target site by constructing a zinc finger array binding the target site, and conjugating the zinc finger array to a DNA cleavage domain. In embodiments where the DNA cleavage domain needs to dimerize in order to cleave DNA, to zinc finger arrays will be designed, each binding a half site of the nuclease target site, and each conjugated to a cleavage domain. In some embodiments, nuclease designing and/or generating is done by recombinant technology. Suitable recombinant technologies are well known to those of skill in the art, and the disclosure is not limited in this respect.

[0083] In some embodiments, a site-specific nuclease designed or generated according to aspects of this disclosure is isolated and/or purified. The methods and strategies for designing site-specific nucleases according to aspects of this disclosure can be applied to design or generate any site-specific nuclease, including, but not limited to Zinc Finger Nucleases, Transcription Activator-Like Effector Nucleases (TALENs), a homing endonuclease, an organic compound nuclease, or an enediyne antibiotic (e.g., dynemicin, neocarzinostatin, calicheamicin, esperamicin, bleomycin).

50 *Isolated Nucleases*

[0084] Some aspects of this disclosure provide isolated site-specific nucleases with enhanced specificity that are designed using the methods and strategies described herein. Some embodiments of this disclosure provide nucleic acids encoding such nucleases. Some embodiments of this disclosure provide expression constructs comprising such encoding nucleic acids. For example, in some embodiments an isolated nuclease is provided that has been engineered to cleave a desired target site within a genome, and has been evaluated according to a method provided herein to cut less than 1, less than 2, less than 3, less than 4, less than 5, less than 6, less than 7, less than 8, less than 9 or less than 10 off-target sites at a concentration effective for the nuclease to cut its intended target site. In some embodiments an isolated nuclease is provided that has been engineered to cleave a desired unique target site that has been selected

to differ from any other site within a genome by at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, or at least 10 nucleotide residues. In some embodiments, the isolated nuclease is an RNA-programmable nuclease, such as a Cas9 nuclease; a Zinc Finger Nuclease (ZFN); or a Transcription Activator-Like Effector Nuclease (TALEN), a homing endonuclease, an organic compound nuclease, or an enediyne antibiotic (e.g., dynemicin, neocarzinostatin, calicheamicin, esperamicin, bleomycin). In some embodiments, the isolated nuclease cleaves a target site within an allele that is associated with a disease or disorder. In some embodiments, the isolated nuclease cleaves a target site the cleavage of which results in treatment or prevention of a disease or disorder. In some embodiments, the disease is HIV/AIDS, or a proliferative disease. In some embodiments, the allele is a CCR5 (for treating HIV/AIDS) or a VEGFA allele (for treating a proliferative disease).

[0085] In some embodiments of the disclosure, the isolated nuclease is provided as part of a pharmaceutical composition. For example, some embodiments provide pharmaceutical compositions comprising a nuclease as provided herein, or a nucleic acid encoding such a nuclease, and a pharmaceutically acceptable excipient. Pharmaceutical compositions may optionally comprise one or more additional therapeutically active substances.

[0086] In some embodiments of the disclosure, compositions provided herein are administered to a subject, for example, to a human subject, in order to effect a targeted genomic modification within the subject. In some embodiments of the disclosure, cells are obtained from the subject and contacted with a nuclease or a nuclease-encoding nucleic acid ex vivo, and re-administered to the subject after the desired genomic modification has been effected or detected in the cells. Although the descriptions of pharmaceutical compositions provided herein are principally directed to pharmaceutical compositions which are suitable for administration to humans, it will be understood by the skilled artisan that such compositions are generally suitable for administration to animals of all sorts. Modification of pharmaceutical compositions suitable for administration to humans in order to render the compositions suitable for administration to various animals is well understood, and the ordinarily skilled veterinary pharmacologist can design and/or perform such modification with merely ordinary, if any, experimentation. Subjects to which administration of the pharmaceutical compositions is contemplated include, but are not limited to, humans and/or other primates; mammals, including commercially relevant mammals such as cattle, pigs, horses, sheep, cats, dogs, mice, and/or rats; and/or birds, including commercially relevant birds such as chickens, ducks, geese, and/or turkeys.

[0087] Formulations of the pharmaceutical compositions described herein may be prepared by any method known or hereafter developed in the art of pharmacology. In general, such preparatory methods include the step of bringing the active ingredient into association with an excipient and/or one or more other accessory ingredients, and then, if necessary and/or desirable, shaping and/or packaging the product into a desired single- or multi-dose unit.

[0088] Pharmaceutical formulations may additionally comprise a pharmaceutically acceptable excipient, which, as used herein, includes any and all solvents, dispersion media, diluents, or other liquid vehicles, dispersion or suspension aids, surface active agents, isotonic agents, thickening or emulsifying agents, preservatives, solid binders, lubricants and the like, as suited to the particular dosage form desired. Remington's The Science and Practice of Pharmacy, 21st Edition, A. R. Gennaro (Lippincott, Williams & Wilkins, Baltimore, MD, 2006) discloses various excipients used in formulating pharmaceutical compositions and known techniques for the preparation thereof. See also PCT application PCT/US2010/055131 for additional suitable methods, reagents, excipients and solvents for producing pharmaceutical compositions comprising a nuclease. Except insofar as any conventional excipient medium is incompatible with a substance or its derivatives, such as by producing any undesirable biological effect or otherwise interacting in a deleterious manner with any other component(s) of the pharmaceutical composition, its use is contemplated to be within the scope of this disclosure.

[0089] The function and advantage of these and other embodiments of the present invention will be more fully understood from the Examples below. The following Examples are intended to illustrate the benefits of the present invention and to describe particular embodiments, but are not intended to exemplify the full scope of the invention. Accordingly, it will be understood that the Examples are not meant to limit the scope of the invention.

EXAMPLES

Materials and Methods

[0090] *Oligonucleotides.* All oligonucleotides used in this study were purchased from Integrated DNA Technologies. Oligonucleotide sequences are listed in Table 9.

[0091] *Expression and Purification of S. pyogenes Cas9.* *E. coli* Rosetta (DE3) cells were transformed with plasmid pMJ806¹¹, encoding the *S. pyogenes cas9* gene fused to an N-terminal 6xHis-tag/maltose binding protein. The resulting expression strain was inoculated in Luria-Bertani (LB) broth containing 100 µg/mL of ampicillin and 30 µg/mL of chloramphenicol at 37 °C overnight. The cells were diluted 1:100 into the same growth medium and grown at 37 °C to OD₆₀₀ ~0.6. The culture was incubated at 18 °C for 30 min, and isopropyl β-D-1-thiogalactopyranoside (IPTG) was added at 0.2 mM to induce Cas9 expression. After ~17 h, the cells were collected by centrifugation at 8,000 g and resuspended

in lysis buffer (20 mM tris(hydroxymethyl)-aminomethane (Tris)-HCl, pH 8.0, 1 M KCl, 20 % glycerol, 1 mM tris (2-carboxyethyl)phosphine (TCEP)). The cells were lysed by sonication (10 sec pulse-on and 30 sec pulse-off for 10 min total at 6 W output) and the soluble lysate was obtained by centrifugation at 20,000 g for 30 min. The cell lysate was incubated with nickel-nitriloacetic acid (nickel-NTA) resin (Qiagen) at 4 °C for 20 min to capture His-tagged Cas9. The resin was transferred to a 20-mL column and washed with 20 column volumes of lysis buffer. Cas9 was eluted in 20 mM Tris-HCl (pH 8), 0.1 M KCl, 20 % glycerol, 1 mM TCEP, and 250 mM imidazole, and concentrated by Amicon ultra centrifugal filter (Millipore, 30-kDa molecular weight cut-off) to ~50 mg/mL. The 6xHis tag and maltose-binding protein were removed by TEV protease treatment at 4 °C for 20 h and captured by a second Ni-affinity purification step. The eluent, containing Cas9, was injected into a HiTrap SP FF column (GE Healthcare) in purification buffer containing 20 mM Tris-HCl (pH 8), 0.1 M KCl, 20 % glycerol, and 1 mM TCEP. Cas9 was eluted with purification buffer containing a linear KCl gradient from 0.1 M to 1 M over five column volumes. The eluted Cas9 was further purified by a HiLoad Superdex 200 column in purification buffer, snap-frozen in liquid nitrogen, and stored in aliquots at -80 °C.

[0092] *In Vitro RNA Transcription.* 100 pmol CLTA(# v2.1 fwd and v2.1 template rev were incubated at 95 °C and cooled at 0.1 °C/s to 37 °C in NEBuffer2 (50 mM sodium chloride, 10 mM Tris-HCl, 10 mM magnesium chloride, 1 mM dithiothreitol, pH 7.9) supplemented with 10 μM dNTP mix (Bio-Rad). 10 U of Klenow Fragment (3'→5' exo-) (NEB) were added to the reaction mixture and a double-stranded CLTA(# v2.1 template was obtained by overlap extension for 1 h at 37 °C. 200 nM CLTA(# v2.1 template alone or 100 nM CLTA(# template with 100 nM T7 promoter oligo was incubated overnight at 37 °C with 0.16 U/μL of T7 RNA Polymerase (NEB) in NEB RNAPol Buffer (40 mM Tris-HCl, pH 7.9, 6 mM magnesium chloride, 10 mM dithiothreitol, 2 mM spermidine) supplemented with 1 mM rNTP mix (1 mM rATP, 1 mM rCTP, 1 mM rGTP, 1 mM rUTP). *In vitro* transcribed RNA was precipitated with ethanol and purified by gel electrophoresis on a Criterion 10% polyacrylamide TBE-Urea gel (Bio-Rad). Gel-purified sgRNA was precipitated with ethanol and redissolved in water.

[0093] *In Vitro Library Construction.* 10 pmol of CLTA(# lib oligonucleotides were separately circularized by incubation with 100 units of CircLigase II ssDNA Ligase (Epicentre) in 1x CircLigase II Reaction Buffer (33 mM Tris-acetate, 66 mM potassium acetate, 0.5 mM dithiothreitol, pH 7.5) supplemented with 2.5 mM manganese chloride in a total reaction volume of 20 μL for 16 hours at 60 °C. The reaction mixture was incubated for 10 minutes at 85 °C to inactivate the enzyme. 5 μL (5 pmol) of the crude circular single-stranded DNA were converted into the concatemeric pre-selection libraries with the illustra TemplPhi Amplification Kit (GE Healthcare) according to the manufacturer's protocol. Concatemeric pre-selection libraries were quantified with the Quant-it PicoGreen dsDNA Assay Kit (Invitrogen).

[0094] *In Vitro Cleavage of On-Target and Off-Target Substrates.* Plasmid templates for PCR were constructed by ligation of annealed oligonucleotides CLTA(# site fwd/rev into *Hind*III/*Xba*I double-digested pUC19 (NEB). On-target substrate DNAs were generated by PCR with the plasmid templates and test fwd and test rev primers, then purified with the QIAquick PCR Purification Kit (Qiagen). Off-target substrate DNAs were generated by primer extension. 100 pmol off-target (#) fwd and off-target (#) rev primers were incubated at 95 °C and cooled at 0.1 °C/s to 37 °C in NEBuffer2 (50 mM sodium chloride, 10 mM Tris-HCl, 10 mM magnesium chloride, 1 mM dithiothreitol, pH 7.9) supplemented with 10 μM dNTP mix (Bio-Rad). 10 U of Klenow Fragment (3'→5' exo-) (NEB) were added to the reaction mixture and double-stranded off-target templates were obtained by overlap extension for 1 h at 37 °C followed by enzyme inactivation for 20 min at 75 °C, then purified with the QIAquick PCR Purification Kit (Qiagen). 200 nM substrate DNAs were incubated with 100 nM Cas9 and 100 nM (v1.0 or v2.1) sgRNA or 1000 nM Cas9 and 1000 nM (v1.0 or v2.1) sgRNA in Cas9 cleavage buffer (200 mM HEPES, pH 7.5, 1.5 M potassium chloride, 100 mM magnesium chloride, 1 mM EDTA, 5 mM dithiothreitol) for 10 min at 37 °C. On-target cleavage reactions were purified with the QIAquick PCR Purification Kit (Qiagen), and off-target cleavage reactions were purified with the QIAquick Nucleotide Removal Kit (Qiagen) before electrophoresis in a Criterion 5% polyacrylamide TBE gel (Bio-Rad).

[0095] *In Vitro Selection.* 200 nM concatemeric pre-selection libraries were incubated with 100 nM Cas9 and 100 nM sgRNA or 1000 nM Cas9 and 1000 nM sgRNA in Cas9 cleavage buffer (200 mM HEPES, pH 7.5, 1.5 M potassium chloride, 100 mM magnesium chloride, 1 mM EDTA, 5 mM dithiothreitol) for 10 min at 37 °C. Pre-selection libraries were also separately incubated with 2 U of BspMI restriction endonuclease (NEB) in NEBuffer 3 (100 mM NaCl, 50 mM Tris-HCl, 10 mM MgCl₂, 1 mM dithiothreitol, pH 7.9) for 1 h at 37 °C. Blunt-ended post-selection library members or sticky-ended pre-selection library members were purified with the QIAquick PCR Purification Kit (Qiagen) and ligated to 10 pmol adapter1/2(AACA) (Cas9:v2.1 sgRNA, 100 nM), adapter1/2(TTCA) (Cas9:v2.1 sgRNA, 1000 nM), adapter1/2 (Cas9:v2.1 sgRNA, 1000 nM), or lib adapter1/CLTA(# lib adapter 2 (pre-selection) with 1,000 U of T4 DNA Ligase (NEB) in NEB T4 DNA Ligase Reaction Buffer (50 mM Tris-HCl, pH 7.5, 10 mM magnesium chloride, 1 mM ATP, 10 mM dithiothreitol) overnight (> 10 h) at room temperature. Adapter-ligated DNA was purified with the QIAquick PCR Purification Kit and PCR-amplified for 10-13 cycles with Phusion Hot Start Flex DNA Polymerase (NEB) in Buffer HF (NEB) and primers CLTA(# sel PCR/PE2 short (post-selection) or CLTA(# lib seq PCR/lib fwd PCR (pre-selection). Amplified DNAs were gel purified, quantified with the KAPA Library Quantification Kit-Illumina (KAPA Biosystems), and subjected to single-read sequencing on an Illumina MiSeq or Rapid Run single-read sequencing on an Illumina HiSeq 2500 (Harvard University FAS Center for Systems Biology Core facility, Cambridge, MA).

[0096] *Selection Analysis.* Pre-selection and post-selection sequencing data were analyzed as previously described²¹, with modification (Algorithms) using scripts written in C++. Raw sequence data is not shown; see Table 2 for a curated summary. Specificity scores were calculated with the formulae: positive specificity score = (frequency of base pair at position[post-selection] - frequency of base pair at position[pre-selection]) / (1 - frequency of base pair at position[pre-selection]) and negative specificity score = (frequency of base pair at position[post-selection] - frequency of base pair at position[pre-selection]) / (frequency of base pair at position[pre-selection]). Normalization for sequence logos was performed as previously described²².

[0097] *Cellular Cleavage Assays.* HEK293T cells were split at a density of 0.8×10^5 per well (6-well plate) before transcription and maintained in Dulbecco's modified eagle medium (DMEM) supplemented with 10% fetal bovine serum (FBS) in a 37°C humidified incubator with 5% CO₂. After 1 day, cells were transiently transfected using Lipofectamine 2000 (Invitrogen) following the manufacturer's protocols. HEK293T cells were transfected at 70% confluence in each well of 6-well plate with 1.0 µg of the Cas9 expression plasmid (Cas9-HA-2xNLS-GFP-NLS) and 2.5 µg of the single-strand RNA expression plasmid pSilencer-CLTA (version 1.0 or 2.1). The transfection efficiencies were estimated to be ~70%, based on the fraction of GFP-positive cells observed by fluorescence microscopy. 48 h after transfection, cells were washed with phosphate buffered saline (PBS), pelleted and frozen at -80 °C. Genomic DNA was isolated from 200 µL cell lysate using the DNeasy Blood and Tissue Kit (Qiagen) according to the manufacturer's protocol.

[0098] *Off-Target Site Sequence Determination.* 100 ng genomic DNA isolated from cells treated with Cas9 expression plasmid and single-strand RNA expression plasmid (treated cells) or Cas9 expression plasmid alone (control cells) were amplified by PCR with 10 s 72°C extension for 35 cycles with primers CLTA(#)-(#)-(#) fwd and CLTA(#)-(#)-(#) rev and Phusion Hot Start Flex DNA Polymerase (NEB) in Buffer GC (NEB), supplemented with 3% DMSO. Relative amounts of crude PCR products were quantified by gel, and Cas9-treated (control) and Cas9:sgRNA-treated PCRs were separately pooled in equimolar concentrations before purification with the QIAquick PCR Purification Kit (Qiagen). Purified DNA was amplified by PCR with primers PE1-barcode# and PE2-barcode# for 7 cycles with Phusion Hot Start Flex DNA Polymerase (NEB) in Buffer HF (NEB). Amplified control and treated DNA pools were purified with the QIAquick PCR Purification Kit (Qiagen), followed by purification with Agencourt AMPure XP (Beckman Coulter). Purified control and treated DNAs were quantified with the KAPA Library Quantification Kit-Illumina (KAPA Biosystems), pooled in a 1:1 ratio, and subjected to paired-end sequencing on an Illumina MiSeq.

[0099] *Statistical Analysis.* Statistical analysis was performed as previously described²¹. P-values in Table 1 and Table 6 were calculated for a one-sided Fisher exact test.

Algorithms

[0100] All scripts were written in C++. Algorithms used in this study are as previous reported (reference) with modification.

[0101] *Sequence binning.* 1) designate sequence pairs starting with the barcode "AACAT" or "CCACAT" as post-selection library members. 2) for post-selection library members (with illustrated example): example read:

AACACATGGGTCGACACAAACACAACTCGGCAGGTACTTGAGATGTAGTCTTT
CCACATGGGTCGACACAAACACAACTCGGCAGGTATCTGTATGCC (SEQ ID NO:42)

i) search both paired reads for the positions, pos1 and pos2, of the constant sequence "**CTCGGCAGGT**" (SEQ ID NO:43). ii) keep only sequences that have identical sequences between the barcode and pos1 and preceding pos2. iii) keep the region between the two instances of the constant sequence (the region between the barcode and pos1 contains a cut half-site; the region that is between the two instances of the constant sequence contains a full site) example: ACTTGAGATGTAGTCTTCCACATGGGTCGACACAAACACAA (SEQ ID NO:44)

ii) search the sequence for a selection barcode ("TGTGTTGTGTT" (SEQ ID NO:45) for CLTA1, "AGAAGAAGAAGA" (SEQ ID NO:46) for CLTA2, "TTCTCTTCTCT" (SEQ ID NO:47) for CLTA3, "ACACAAACACAA" (SEQ ID NO:48) for CLTA4) example: ACTTGAGATGTAGTCTTCCACATGGGTCGACACAAACACAA (SEQ ID NO:49) - CLTA4

iii) the sequence before the barcode is the full post-selection library member (first four and last four nucleotides are fully randomized flanking sequence)
example: ACTT GCAGATGTAGTCTTCCACATGG GTCG (SEQ ID NO:50)

EP 3 486 318 B1

iv) parse the quality scores for the positions corresponding to the 23 nucleotide post-selection library member example read:

5 AACACATGGGTCGACACAAACACAACACTCGGCAGGTACTTCAGATGTAGTCTTC
CACATGGGTCGACACAAACACAACACTCGGCAGGTATCTGTATGCC (SEQ ID
NO:51)

10 CCCFFFFFHHHHHJJJJJJJJJJJJJJJJJJGIJJJJJJJHIIJJHHHGHAEF~~CD~~DDDDDDDD
DDDDDDDDDDDDDDDD?CDDEDD@DCCCD

15 v) keep sequences only if the corresponding quality score string (underlined) FASTQ quality characters for the sequence are '?' or higher in ASCII code (Phred quality score >= 30)

20 *NHEJ sequence calling*

[0102]

example read:

25 CAATCTCCCGCATGCGCTCAGTCCTCATCTCCCTCAAGCAGGCCGGTGCA
CTGAAGAGCCACCCTGTGAAACACTACATCTGCAATATCTTAATCCTACTCAGT
30 GAAGCTCTTCACAGTCATTGGATTAATTATGTTGAGTTCTTGGACCAAACC
(SEQ ID NO:52)

35 example quality scores:

CCBCCFFFCCCCGGGGGGGGGHHHHHHHHHHHHHHHHHHHHHGGGGGGGGGH
40 HHHHHHHHHHHHHHHHHHGHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
HHHHHFHHHG
FHHHHHF

45 1) identify the 20 base pairs flanking both sides of 20 base pair target site + three base pair PAM for each target site

example flanking sequences:

50 GCTGGTGCACTGAAGAGCCA (SEQ ID NO:53)
AATATCTTAATCCTACTCAG (SEQ ID NO:54)

55 2) search all sequence reads for the flanking sequences to identify the potential off-target site (the sequence between the flanking sequences)

example potential off-target site: **CCCTGTGAAACACTACATCTGC** (SEQ ID NO:55)

3) if the potential off-target site contains indels (length is less than 23), keep sequence as potential off-target site if all corresponding FASTQ quality characters for the sequence are '?' or higher in ASCII code (Phred quality score >= 30)

example potential off-target site length = 22 example corresponding FASTQ quality characters: HHGHHHHH-HHHHHHHHHHHHHHH

5 4) bin and manually inspect all sequences that pass steps 2 and 3 and keep sequences as potential modified sequences if they have at least one deletion involving position 16, 17, or 18 (of 20 counting from the non-PAM end) or if they have an insertion between position 17 and 18, consistent with the most frequent modifications observed for the intended target site (**Figure 3**)

example potential off-target site (reverse complement, with positions labeled) with reference sequence:

10 1111111112222

non-PAM end 12345678901234567890123 PAM end

GCAGATGTAGTGTTC-ACAGGG (SEQ ID NO:56)

GCAGATGTAGTGTCCACAGGG (SEQ ID NO:57)

15 4) repeat steps 1-3 for read2 and keep only if the sequence is the same

5) compare overall counts in Cas9+sgRNA treated sample to Cas9 alone sample to identify modified sites

Filter based on cleavage site (for post-selection sequences)

20 [0103]

- 25 1) tabulate the cleavage site locations across the recognition site by identifying the first position in the full sequenced recognition site (between the two constant sequences) that is identical to the first position in the sequencing read after the barcode (before the first constant sequence).
- 2) after tabulation, repeat step 1, keeping only sequences with cleavage site locations that are present in at least 5% of the sequencing reads.

30 *Results*

Broad off-target DNA cleavage profiling reveals RNA-programmed Cas9 nuclease specificity.

35 [0104] Sequence-specific endonucleases including zinc-finger nucleases (ZFNs) and transcription activator-like effector nucleases (TALENs) have become important tools to modify genes in induced pluripotent stem cells (iPSCs),¹⁻³ in multi-cellular organisms,⁴⁻⁸ and in *ex vivo* gene therapy clinical trials.^{9,10} Although ZFNs and TALENs have proved effective for such genetic manipulation, a new ZFN or TALEN protein must be generated for each DNA target site. In contrast, the RNA-guided Cas9 endonuclease uses RNA:DNA hybridization to determine target DNA cleavage sites, enabling a single monomeric protein to cleave, in principle, any sequence specified by the guide RNA.¹¹

40 [0105] Previous studies¹²⁻¹⁷ demonstrated that Cas9 mediates genome editing at sites complementary to a 20-nucleotide sequence in a bound guide RNA. In addition, target sites must include a protospacer adjacent motif (PAM) at the 3' end adjacent to the 20-nucleotide target site; for *Streptococcus pyogenes* Cas9, the PAM sequence is NGG. Cas9-mediated DNA cleavage specificity both *in vitro* and in cells has been inferred previously based on assays against small collections of potential single-mutation off-target sites. These studies suggested that perfect complementarity between guide RNA and target DNA is required in the 7-12 base pairs adjacent to the PAM end of the target site (3' end of the guide RNA) and mismatches are tolerated at the non-PAM end (5' end of the guide RNA).^{11, 12, 17-19}

45 [0106] Although such a limited number of nucleotides specifying Cas9:guide RNA target recognition would predict multiple sites of DNA cleavage in genomes of moderate to large size (> ~10⁷ bp), Cas9:guide RNA complexes have been successfully used to modify both cells^{12, 13, 15} and organisms.¹⁴ A study using Cas9:guide RNA complexes to modify zebrafish embryos observed toxicity at a rate similar to that of ZFNs and TALENs.¹⁴ A recent, broad study of the specificity of DNA binding (transcriptional repression) in *E. coli* of a catalytically inactive Cas9 mutant using high-throughput sequencing found no detectable off-target transcriptional repression in the relatively small *E. coli* transcriptome.²⁰ While these studies have substantially advanced our basic understanding of Cas9, a systematic and comprehensive profile of Cas9:guide RNA-mediated DNA cleavage specificity generated from measurements of Cas9 cleavage on a large number of related mutant target sites has not been described. Such a specificity profile is needed to understand and improve the potential of Cas9:guide RNA complexes as research tools and future therapeutic agents.

55 [0107] We modified our previously published *in vitro* selection,²¹ adapted to process the blunt-ended cleavage products produced by Cas9 compared to the overhang-containing products of ZFN cleavage, to determine the off-target DNA cleavage profiles of Cas9: single guide RNA (sgRNA)¹¹ complexes. Each selection experiment used DNA substrate

libraries containing $\sim 10^{12}$ sequences, a size sufficiently large to include ten-fold coverage of all sequences with eight or fewer mutations relative to each 22-base pair target sequence (including the two-base pair PAM) (FIG. 1). We used partially randomized nucleotide mixtures at all 22 target-site base pairs to create a binomially distributed library of mutant target sites with an expected mean of 4.62 mutations per target site. In addition, target site library members were flanked by four fully randomized base pairs on each side to test for specificity patterns beyond those imposed by the canonical 20-base pair target site and PAM

[0108] Pre-selection libraries of 10^{12} individual potential off-target sites were generated for each of four different target sequences in the human clathrin light chain A (*CLTA*) gene (FIG. 3). Synthetic 5'-phosphorylated 53-base oligonucleotides were self-ligated into circular single-stranded DNA *in vitro*, then converted into concatemeric 53-base pair repeats through rolling-circle amplification. The resulting pre-selection libraries were incubated with their corresponding Cas9:sgRNA complexes. Cleaved library members containing free 5' phosphates were separated from intact library members through the 5' phosphate-dependent ligation of non-phosphorylated double-stranded sequencing adapters. The ligation-tagged post-selection libraries were amplified by PCR. The PCR step generated a mixture of post-selection DNA fragments containing 0.5, 1.5, or 2.5, etc. repeats of library members cleaved by Cas9, resulting from amplification of an adapter-ligated cut half-site with or without one or more adjacent corresponding full sites (FIG. 1). Post-selection library members with 1.5 target-sequence repeats were isolated by gel purification and analyzed by high-throughput sequencing. In a final computational selection step to minimize the impact of errors during DNA amplification or sequencing, only sequences with two identical copies of the repeated cut half-site were analyzed.

[0109] Pre-selection libraries were incubated under enzyme-limiting conditions (200 nM target site library, 100 nM Cas9:sgRNA v2.1) or enzyme-saturating conditions (200 nM target site library, 1000 nM Cas9:sgRNA v2.1) for each of the four guide RNAs targets tested (*CLTA1*, *CLTA2*, *CLTA3*, and *CLTA4*) (FIG. 3C and 3D). A second guide RNA construct, sgRNA v1.0, which is less active than sgRNA v2.1, was assayed under enzyme-saturating conditions alone for each of the four guide RNA targets tested (200 nM target site library, 1000 nM Cas9:sgRNA v1.0). The two guide RNA constructs differ in their length (FIG. 3) and in their DNA cleavage activity level under the selection conditions, consistent with previous reports¹⁵ (FIG. 4). Both pre-selection and post-selection libraries were characterized by high-throughput DNA sequencing and computational analysis. As expected, library members with fewer mutations were significantly enriched in post-selection libraries relative to pre-selection libraries (FIG. 5).

[0110] *Pre- and Post-Selection Library Composition.* The pre-selection libraries for *CLTA1*, *CLTA2*, *CLTA3*, and *CLTA4* had observed mean mutation rates of 4.82 (n = 1,129,593), 5.06 (n = 847,618), 4.66 (n = 692,997), and 5.00 (n = 951,503) mutations per 22-base pair target site, including the two-base pair PAM, respectively. The post-selection libraries treated under enzyme-limiting conditions with Cas9 plus *CLTA1*, *CLTA2*, *CLTA3*, or *CLTA4* v.2.1 sgRNAs contained means of 1.14 (n = 1,206,268), 1.21 (n = 668,312), 0.91 (n = 1,138,568), and 1.82 (n = 560,758) mutations per 22-base pair target site. Under enzyme-excess conditions, the mean number of mutations among sequences surviving selection increased to 1.61 (n = 640,391), 1.86 (n = 399,560), 1.46 (n = 936,414), and 2.24 (n = 506,179) mutations per 22-base pair target site, respectively, for *CLTA1*, *CLTA2*, *CLTA3*, or *CLTA4* v2.1 sgRNAs. These results reveal that the selection significantly enriched library members with fewer mutations for all Cas9:sgRNA complexes tested, and that enzyme-excess conditions resulted in the putative cleavage of more highly mutated library members compared with enzyme-limiting conditions (FIG. 5).

[0111] We calculated specificity scores to quantify the enrichment level of each base pair at each position in the post-selection library relative to the pre-selection library, normalized to the maximum possible enrichment of that base pair. Positive specificity scores indicate base pairs that were enriched in the post-selection library and negative specificity scores indicate base pairs that were de-enriched in the post-selection library. For example, a score of +0.5 indicates that a base pair is enriched to 50% of the maximum enrichment value, while a score of -0.5 indicates that a base pair is de-enriched to 50% of the maximum deenrichment value.

[0112] In addition to the two base pairs specified by the PAM, all 20 base pairs targeted by the guide RNA were enriched in the sequences from the *CLTA1* and *CLTA2* selections (FIG. 2, FIG. 6 and 9, and Table 2). For the *CLTA3* and *CLTA4* selections (FIG. 7 and 8, and Table 2), guide RNA-specified base pairs were enriched at all positions except for the two most distal base pairs from the PAM (5' end of the guide RNA), respectively. At these non-specified positions farthest from the PAM, at least two of the three alternate base pairs were nearly as enriched as the specified base pair. Our finding that the entire 20 base-pair target site and two base pair PAM can contribute to Cas9:sgRNA DNA cleavage specificity contrasts with the results from previous single-substrate assays suggesting that only 7-12 base pairs and two base pair PAM are specified.^{11, 12, 15}

[0113] All single-mutant pre-selection (n \geq 14,569) and post-selection library members (n \geq 103,660) were computationally analyzed to provide a selection enrichment value for every possible single-mutant sequence. The results of this analysis (FIG. 2 and FIG. 6 and 8) show that when only single-mutant sequences are considered, the six to eight base pairs closest to the PAM are generally highly specified and the non-PAM end is poorly specified under enzyme-limiting conditions, consistent with previous findings.^{11, 12, 17-19} Under enzyme-saturating conditions, however, single mutations even in the six to eight base pairs most proximal to the PAM are tolerated, suggesting that the high specificity at the

PAM end of the DNA target site can be compromised when enzyme concentrations are high relative to substrate (FIG. 2). The observation of high specificity against single mutations close to the PAM only applies to sequences with a single mutation and the selection results do not support a model in which any combination of mutations is tolerated in the region of the target site farthest from the PAM (FIG. 10-15). Analyses of pre- and post-selection library composition are described elsewhere herein, position-dependent specificity patterns are illustrated in FIGs 18-20, PAM nucleotide specificity is illustrated in FIGs 21-24, and more detailed effects of Cas9:sgRNA concentration on specificity are described in FIG. 2G and FIG. 25).

[0114] *Specificity at the Non-PAM End of the Target Site.* To assess the ability of Cas9:v2.1 sgRNA under enzyme-excess conditions to tolerate multiple mutations distal to the PAM, we calculated maximum specificity scores at each position for sequences that contained mutations only in the region of one to 12 base pairs at the end of the target site most distal from the PAM (FIG. 10-17).

[0115] The results of this analysis show no selection (maximum specificity score ~ 0) against sequences with up to three mutations, depending on the target site, at the end of the molecule farthest from the PAM when the rest of the sequence contains no mutations. For example, when only the three base pairs farthest from the PAM are allowed to vary (indicated by dark bars in FIG. 11C) in the CLTA2 target site, the maximum specificity scores at each of the three variable positions are close to zero, indicating that there was no selection for any of the four possible base pairs at each of the three variable positions. However, when the eight base pairs farthest from the PAM are allowed to vary (FIG. 11H), the maximum specificity scores at positions 4-8 are all greater than +0.4, indicating that the Cas9:sgRNA has a sequence preference at these positions even when the rest of the substrate contains preferred, on-target base pairs.

[0116] We also calculated the distribution of mutations (FIG. 15-17), in both pre-selection and v2.1 sgRNA-treated post-selection libraries under enzyme-excess conditions, when only the first 1-12 base pairs of the target site are allowed to vary. There is significant overlap between the pre-selection and post-selection libraries for only a subset of the data (FIG. 15-17, a-c), demonstrating minimal to no selection in the post-selection library for sequences with mutations only in the first three base pairs of the target site. These results collectively show that Cas9:sgRNA can tolerate a small number of mutations (~one to three) at the end of the sequence farthest from the PAM when provided with maximal sgRNA:DNA interactions in the rest of the target site.

[0117] *Specificity at the PAM End of the Target Site.* We plotted positional specificity as the sum of the magnitudes of the specificity scores for all four base pairs at each position of each target site, normalized to the same sum for the most highly specified position (FIG. 18-20). Under both enzyme-limiting and enzyme-excess conditions, the PAM end of the target site is highly specified. Under enzyme-limiting conditions, the PAM end of the molecule is almost absolutely specified (specificity score \geq +0.9 for guide RNA-specified base pairs) by CLTA1, CLTA2, and CLTA3 guide RNAs (FIG. 2 and FIG. 6-9), and highly specified by CLTA4 guide RNA (specificity score of +0.7 to +0.9). Within this region of high specificity, specific single mutations, consistent with wobble pairing between the guide RNA and target DNA, that are tolerated. For example, under enzyme-limiting conditions for single-mutant sequences, a dA:dT off-target base pair and a guide RNA-specified dG:dC base pair are equally tolerated at position 17 out of 20 (relative to the non-PAM end of the target site) of the CLTA3 target site. At this position, an rG:dT wobble RNA:DNA base pair may be formed, with minimal apparent loss of cleavage activity.

[0118] Importantly, the selection results also reveal that the choice of guide RNA hairpin affects specificity. The shorter, less-active sgRNA v1.0 constructs are more specific than the longer, more-active sgRNA v2.1 constructs when assayed under identical, enzyme-saturating conditions that reflect an excess of enzyme relative to substrate in a cellular context (FIG. 2 and FIGs. 5-8). The higher specificity of sgRNA v1.0 over sgRNA v2.1 is greater for CLTA1 and CLTA2 (~40-90% difference) than for CLTA3 and CLTA4 (< 40% difference). Interestingly, this specificity difference is localized to different regions of the target site for each target sequence (FIGs. 2H and 26). Collectively, these results indicate that different guide RNA architectures result in different DNA cleavage specificities, and that guide RNA-dependent changes in specificity do not affect all positions in the target site equally. Given the inverse relationship between Cas9:sgRNA concentration and specificity described above, we speculate that the differences in specificity between guide RNA architectures arises from differences in their overall level of DNA-cleavage activities.

[0119] *Effects of Cas9:sgRNA Concentration on DNA Cleavage Specificity.* To assess the effect of enzyme concentration on patterns of specificity for the four target sites tested, we calculated the concentration-dependent difference in positional specificity and compared it to the maximal possible change in positional specificity (FIG. 25). In general, specificity was higher under enzyme-limiting conditions than enzyme-excess conditions. A change from enzyme-excess to enzyme-limiting conditions generally increased the specificity at the PAM end of the target by \geq 80% of the maximum possible change in specificity. Although a decrease in enzyme concentration generally induces small (~30%) increases in specificity at the end of the target sites farthest from the PAM, concentration decreases induce much larger increases in specificity at the end of the target site nearest the PAM. For CLTA4, a decrease in enzyme concentration is accompanied by a small (~30%) decrease in specificity at some base pairs near the end of the target site farthest from the PAM

[0120] *Specificity of PAM Nucleotides.* To assess the contribution of the PAM to specificity, we calculated the abundance of all 16 possible PAM dinucleotides in the pre-selection and post-selection libraries, considering all observed post-

selection target site sequences (FIG. 21) or considering only post-selection target site sequences that contained no mutations in the 20 base pairs specified by the guide RNA (FIG. 22). Considering all observed post-selection target site sequences, under enzyme-limiting conditions, GG dinucleotides represented 99.8%, 99.9%, 99.8%, and 98.5% of the post-selection PAM dinucleotides for selections with CLTA1, CLTA2, CLTA3, and CLTA4 v2.1 sgRNAs, respectively.

In contrast, under enzyme-excess conditions, GG dinucleotides represented 97.7%, 98.3%, 95.7%, and 87.0% of the post-selection PAM dinucleotides for selections with CLTA1, CLTA2, CLTA3, and CLTA4 v2.1 sgRNAs, respectively. These data demonstrate that an increase in enzyme concentration leads to increased cleavage of substrates containing non-canonical PAM dinucleotides.

[0121] To account for the pre-selection library distribution of PAM dinucleotides, we calculated specificity scores for the PAM dinucleotides (FIG. 23). When only on-target post-selection sequences are considered under enzyme-excess conditions (FIG. 24), non-canonical PAM dinucleotides with a single G rather than two Gs are relatively tolerated. Under enzyme-excess conditions, Cas9:CLTA4 sgRNA 2.1 exhibited the highest tolerance of non-canonical PAM dinucleotides of all the Cas9:sgRNA combinations tested. AG and GA dinucleotides were the most tolerated, followed by GT, TG, and CG PAM dinucleotides. In selections with Cas9:CLTA1, 2, or 3 sgRNA 2.1 under enzyme-excess conditions, AG was the predominate non-canonical PAM (FIG. 23 and 24). Our results are consistent with another recent study of PAM specificity, which shows that Cas9:sgRNA can recognize AG PAM dinucleotides²³. In addition, our results show that under enzyme-limiting conditions, GG PAM dinucleotides are highly specified, and under enzyme-excess conditions, non-canonical PAM dinucleotides containing a single G can be tolerated, depending on the guide RNA context.

[0122] To confirm that the *in vitro* selection results accurately reflect the cleavage behavior of Cas9 *in vitro*, we performed discrete cleavage assays of six CLTA4 off-target substrates containing one to three mutations in the target site. We calculated enrichment values for all sequences in the post-selection libraries for the Cas9:CLTA4 v2.1 sgRNA under enzyme-saturating conditions by dividing the abundance of each sequence in the post-selection library by the calculated abundance in the pre-selection library. Under enzyme-saturating conditions, the single one, two, and three mutation sequences with the highest enrichment values (27.5, 43.9, and 95.9) were cleaved to $\geq 71\%$ completion (FIG. 27). A two-mutation sequence with an enrichment value of 1.0 was cleaved to 35%, and a two-mutation sequence with an enrichment value near zero (0.064) was not cleaved. The three-mutation sequence, which was cleaved to 77% by CLTA4 v2.1 sgRNA, was cleaved to a lower efficiency of 53% by CLTA4 v1.0 sgRNA (FIG. 28). These results indicate that the selection enrichment values of individual sequences are predictive of *in vitro* cleavage efficiencies.

[0123] To determine if results of the *in vitro* selection and *in vitro* cleavage assays pertain to Cas9:guide RNA activity in human cells, we identified 51 off-target sites (19 for CLTA1 and 32 for CLTA4) containing up to eight mutations that were both enriched in the *in vitro* selection and present in the human genome (Tables 3-5). We expressed Cas9:CLTA1 sgRNA v1.0, Cas9:CLTA1 sgRNA v2.1, Cas9:CLTA4 sgRNA v1.0, Cas9:CLTA4 sgRNA v2.1, or Cas9 without sgRNA in HEK293T cells by transient transfection and used genomic PCR and high-throughput DNA sequencing to look for evidence of Cas9:sgRNA modification at 46 of the 51 off-target sites as well as at the on-target loci; no specific amplified DNA was obtained for five of the 51 predicted off-target sites (three for CLTA1 and two for CLTA4).

[0124] Deep sequencing of genomic DNA isolated from HEK293T cells treated with Cas9:CLTA1 sgRNA or Cas9:CLTA4 sgRNA identified sequences evident of non-homologous end-joining (NHEJ) at the on-target sites and at five of the 49 tested off-target sites (CLTA1-1-1, CLTA1-2-2, CLTA4-3-1, CLTA4-3-3, and CLTA4-4-8) (Tables 1 and 6-8). The CLTA4 target site was modified by Cas9:CLTA4 v2.1 sgRNA at a frequency of 76%, while off-target sites, CLTA4-3-1 CLTA4-3-3, and CLTA4-4-8, were modified at frequencies of 24%, 0.47% and 0.73%, respectively. The CLTA1 target site was modified by Cas9:CLTA1 v2.1 sgRNA at a frequency of 0.34%, while off-target sites, CLTA1-1-1 and CLTA1-2-2, were modified at frequencies of 0.09% and 0.16%, respectively.

[0125] Under enzyme-saturating conditions with the v2.1 sgRNA, the two verified CLTA1 off-target sites, CLTA1-1-1 and CLTA1-2-2, were two of the three most highly enriched sequences identified in the *in vitro* selection. CLTA4-3-1 and CLTA4-3-3 were the highest and third-highest enriched sequences of the seven CLTA4 three-mutation sequences enriched in the *in vitro* selection that are also present in the genome. The *in vitro* selection enrichment values of the four-mutation sequences were not calculated, since 12 out of the 14 CLTA4 sequences in the genome containing four mutations, including CLTA4-4-8, were observed at a level of only one sequence count in the post-selection library. Taken together, these results confirm that several of the off-target substrates identified in the *in vitro* selection that are present in the human genome are indeed cleaved by Cas9:sgRNA complexes in human cells, and also suggest that the most highly enriched genomic off-target sequences in the selection are modified in cells to the greatest extent.

[0126] The off-target sites we identified in cells were among the most-highly enriched in our *in vitro* selection and contain up to four mutations relative to the intended target sites. While it is possible that heterochromatin or covalent DNA modifications could diminish the ability of a Cas9:guide RNA complex to access genomic off-target sites in cells, the identification of five out of 49 tested cellular off-target sites in this study, rather than zero or many, strongly suggests that Cas9-mediated DNA cleavage is not limited to specific targeting of only a 7-12-base pair target sequence, as suggested in recent studies.^{11, 12, 19}

[0127] The cellular genome modification data are also consistent with the increase in specificity of sgRNA v1.0 com-

pared to sgRNA v2.1 sgRNAs observed in the *in vitro* selection data and discrete assays. Although the CLTA1-2-2, CLTA 4-3-3, and CLTA 4-4-8 sites were modified by the Cas9:sgRNA v2.1 complexes, no evidence of modification at any of these three sites was detected in Cas9:sgRNA v1.0-treated cells. The CLTA4-3-1 site, which was modified at 32% of the frequency of on-target CLTA4 site modification in Cas9:v2.1 sgRNA-treated cells, was modified at only 0.5% of the on-target modification frequency in v1.0 sgRNA-treated cells, representing a 62-fold change in selectivity. Taken together, these results demonstrate that guide RNA architecture can have a significant influence on Cas9 specificity in cells. Our specificity profiling findings present an important caveat to recent and ongoing efforts to improve the overall DNA modification activity of Cas9:guide RNA complexes through guide RNA engineering.^{11, 15}

[0128] Overall, the off-target DNA cleavage profiling of Cas9 and subsequent analyses show that (i) Cas9:guide RNA recognition extends to 18-20 specified target site base pairs and a two-base pair PAM for the four target sites tested; (ii) increasing Cas9:guide RNA concentrations can decrease DNA-cleaving specificity *in vitro*; (iii) using more active sgRNA architectures can increase DNA-cleavage specificity both *in vitro* and in cells but impair DNA-cleavage specificity both *in vitro* and in cells; and (iv) as predicted by our *in vitro* results, Cas9:guide RNA can modify off-target sites in cells with up to four mutations relative to the on-target site. Our findings provide key insights to our understanding of RNA-programmed Cas9 specificity, and reveal a previously unknown role for sgRNA architecture in DNA-cleavage specificity. The principles revealed in this study may also apply to Cas9-based effectors engineered to mediate functions beyond DNA cleavage.

EQUIVALENTS AND SCOPE

[0129] Those skilled in the art will recognize, or be able to ascertain using no more than routine experimentation, many equivalents to the specific embodiments of the invention described herein. The scope of the present invention is not intended to be limited to the above description, but rather is as set forth in the appended claims.

[0130] In the claims articles such as "a," "an," and "the" may mean one or more than one unless indicated to the contrary or otherwise evident from the context. Claims or descriptions that include "or" between one or more members of a group are considered satisfied if one, more than one, or all of the group members are present in, employed in, or otherwise relevant to a given product or process unless indicated to the contrary or otherwise evident from the context. The invention includes embodiments in which exactly one member of the group is present in, employed in, or otherwise relevant to a given product or process. The invention also includes embodiments in which more than one, or all of the group members are present in, employed in, or otherwise relevant to a given product or process.

[0131] Furthermore, it is to be understood that the disclosure encompasses all variations, combinations, and permutations in which one or more limitations, elements, clauses, descriptive terms, etc., from one or more of the claims or from relevant portions of the description is introduced into another claim. For example, any claim that is dependent on another claim can be modified to include one or more limitations found in any other claim that is dependent on the same base claim.

[0132] Where elements are presented as lists, e.g., in Markush group format, it is to be understood that each subgroup of the elements is also disclosed, and any element(s) can be removed from the group. It is also noted that the term "comprising" is intended to be open and permits the inclusion of additional elements or steps. It should be understood that, in general, where the invention, or aspects of the invention, is/are referred to as comprising particular elements, features, steps, etc., certain embodiments of the invention or aspects of the invention consist such elements, features, steps, etc. For purposes of simplicity those embodiments have not been specifically set forth *in haec verba* herein. Thus for each embodiment of the invention that comprises one or more elements, features, steps, etc., the invention also provides embodiments that consist of those elements, features, steps, etc.

[0133] Where ranges are given, endpoints are included. Furthermore, it is to be understood that unless otherwise indicated or otherwise evident from the context and/or the understanding of one of ordinary skill in the art, values that are expressed as ranges can assume any specific value within the stated ranges in different embodiments of the invention, to the tenth of the unit of the lower limit of the range, unless the context clearly dictates otherwise. It is also to be understood that unless otherwise indicated or otherwise evident from the context and/or the understanding of one of ordinary skill in the art, values expressed as ranges can assume any subrange within the given range, wherein the endpoints of the subrange are expressed to the same degree of accuracy as the tenth of the unit of the lower limit of the range.

[0134] In addition, it is to be understood that any particular embodiment of the present disclosure may be explicitly excluded from any one or more of the claims. Where ranges are given, any value within the range may explicitly be excluded from any one or more of the claims. Any embodiment, element, feature, application, or aspect of the compositions and/or methods of the disclosure can be excluded from any one or more claims. For purposes of brevity, all of the embodiments in which one or more elements, features, purposes, or aspects is excluded are not set forth explicitly herein.

TABLES

[0135]

Table 1: Cellular modification induced by Cas9:CLTA4 sgRNA. 33 human genomic DNA sequences were identified that were enriched in the Cas9:CLTA4 v2.1 sgRNA *in vitro* selections under enzyme-limiting or enzyme-saturating conditions. Sites shown with underline contain insertions or deletions (indels) that are consistent with significant Cas9:sgRNA-mediated modification in HEK293T cells. *In vitro* enrichment values for selections with Cas9:CLTA4 v1.0 sgRNA or Cas9:CLTA4 v2.1 sgRNA are shown for sequences with three or fewer mutations. Enrichment values were not calculated for sequences with four or more mutations due to low numbers of *in vitro* selection sequence counts. Modification frequencies (number of sequences with indels divided by total number of sequences) in HEK293T cells treated with Cas9 without sgRNA ("no sgRNA"), Cas9 with CLTA4 v1.0 sgRNA, or Cas9 with CLTA4 v2.1 sgRNA. P-values are listed for those sites that show significant modification in v1.0 sgRNA- or v2.1 sgRNA-treated cells compared to cells treated with Cas9 without sgRNA. "Not tested (n.t.)" indicates that PCR of the genomic sequence failed to provide specific amplification products.

Table 2: Raw selection sequence counts. Positions -4 to -1 are the four nucleotides preceding the 20-base pair target site. PAM1, PAM2, and PAM3 are the PAM positions immediately following the target site. Positions +4 to +7 are the four nucleotides immediately following the PAM

Table 3: CLTA1 genomic off-target sequences. 20 human genomic DNA sequences were identified that were enriched in the Cas9:CLTA1 v2.1 sgRNA *in vitro* selections under enzyme-limiting or enzyme-excess conditions. "m" refers to number of mutations from on-target sequence with mutations shown in lower case. Sites shown with underline contain insertions or deletions (indels) that are consistent with significant Cas9:sgRNA-mediated modification in HEK293T cells. Human genome coordinates are shown for each site (assembly GRCh37). CLTA1-0-1 is present at two loci, and sequence counts were pooled from both loci. Sequence counts are shown for amplified and sequenced DNA for each site from HEK293T cells treated with Cas9 without sgRNA ("no sgRNA"), Cas9 with CLTA1 v1.0 sgRNA, or Cas9 with CLTA1 v2.1 sgRNA.

Table 4: CLTA4 genomic off-target sequences. 33 human genomic DNA sequences were identified that were enriched in the Cas9:CLTA4 v2.1 sgRNA *in vitro* selections under enzyme-limiting or enzyme-excess conditions. "m" refers to number of mutations from on-target sequence with mutations shown in lower case. Sites shown with underline contain insertions or deletions (indels) that are consistent with significant Cas9:sgRNA-mediated modification in HEK293T cells. Human genome coordinates are shown for each site (assembly GRCh37). Sequence counts are shown for amplified and sequenced DNA for each site from HEK293T cells treated with Cas9 without sgRNA ("no sgRNA"), Cas9 with CLTA4 v1.0 sgRNA, or Cas9 with CLTA4 v2.1 sgRNA.

Table 5: Genomic coordinates of CLTA1 and CLTA4 off-target sites. 54 human genomic DNA sequences were identified that were enriched in the Cas9:CLTA1 v2.1 sgRNA and Cas9:CLTA4 v2.1 sgRNA *in vitro* selections under enzyme-limiting or enzyme-excess conditions. Human genome coordinates are shown for each site (assembly GRCh37).

Table 6: Cellular modification induced by Cas9:CLTA1 sgRNA. 20 human genomic DNA sequences were identified that were enriched in the Cas9:CLTA1 v2.1 sgRNA *in vitro* selections under enzyme-limiting or enzyme-excess conditions. Sites shown with underline contain insertions or deletions (indels) that are consistent with significant Cas9:sgRNA-mediated modification in HEK293T cells. *In vitro* enrichment values for selections with Cas9:CLTA1 v1.0 sgRNA or Cas9:CLTA1 v2.1 sgRNA are shown for sequences with three or fewer mutations. Enrichment values were not calculated for sequences with four or more mutations due to low numbers of *in vitro* selection sequence counts. Modification frequencies (number of sequences with indels divided by total number of sequences) in HEK293T cells treated with Cas9 without sgRNA ("no sgRNA"), Cas9 with CLTA1 v1.0 sgRNA, or Cas9 with CLTA1 v2.1 sgRNA. P-values of sites that show significant modification in v1.0 sgRNA- or v2.1 sgRNA-treated cells compared to cells treated with Cas9 without sgRNA were 1.1E-05 (v1.0) and 6.9E-55 (v2.1) for CLTA1-0-1, 2.6E-03 (v1.0) and 2.0E-10 (v2.1) for CLTA1-1-1, and 4.6E-08 (v2.1) for CLTA1-2-2. P-values were calculated using a one-sided Fisher exact test. "Not tested (n.t.)" indicates that the site was not tested or PCR of the genomic sequence failed to provide specific amplification products.

Table 7: CLTA1 genomic off-target indel sequences. Insertion and deletion-containing sequences from cells treated with amplified and sequenced DNA for the on-target genomic sequence (CLTA1-0-1) and each modified off-target site from HEK293T cells treated with Cas9 without sgRNA ("no sgRNA"), Cas9 with CLTA1 v1.0 sgRNA, or Cas9 with CLTA1 v2.1 sgRNA. "ref" refers to the human genome reference sequence for each site, and the modified sites are listed below. Mutations relative to the on-target genomic sequence are shown in lowercase letters. Insertions and deletions are shown in underlined bold letters or dashes, respectively. Modification percentages are shown for those conditions (v1.0 sgRNA or v2.1 sgRNA) that show statistically significant enrichment of modified sequences compared to the control (no sgRNA).

Table 8: CLTA4 genomic off-target indel sequences. Insertion and deletion-containing sequences from cells treated with amplified and sequenced DNA for the on-target genomic sequence (CLTA4-0-1) and each modified off-target site from HEK293T cells treated with Cas9 without sgRNA ("no sgRNA"), Cas9 with CLTA4 v1.0 sgRNA, or Cas9 with CLTA4 v2.1 sgRNA. "ref" refers to the human genome reference sequence for each site, and the modified sites are listed below.

EP 3 486 318 B1

Mutations relative to the on-target genomic sequence are shown in lowercase letters. Insertions and deletions are shown in underlined bold letters or dashes, respectively. Modification percentages are shown for those conditions (v1.0 sgRNA or v2.1 sgRNA) that show statistically significant enrichment of modified sequences compared to the control (no sgRNA).

Table 9: Oligonucleotides used in this study. All oligonucleotides were purchased from Integrated DNA Technologies.

5 An asterisk (*) indicates that the preceding nucleotide was incorporated as a hand mix of phosphoramidites consisting of 79 mol% of the phosphoramidite corresponding to the preceding nucleotide and 4 mol% of each of the other three canonical phosphoramidites. "/5Phos/" denotes a 5' phosphate group installed during synthesis.

10

15

20

25

30

35

40

45

50

55

Table 1

#of Mutations	sequence	SEQ ID NO.	gene	in vitro enrichment		modification frequency in HEK293T cells		P-value
				v1.0	v2.1	no sgRNA	v1.0	
CLTA4-0-1	<u>GCAGATGTAGTGTTC</u> <u>ACAGGG</u>	58	CLTA	20	7.95	0.021%	11%	<1E-55
CLTA4-3-1	<u>aCAATGTAGTatTTCCA</u> <u>CAGGG</u>	59	SEQ ID NO: 59	16.5	12.5	0.006%	0.055%	24%
CLTA4-3-2	GCAATATGTAGTGTTC aATGt	60	SEQ ID NO: 60	2.99	6.97	0.017%	0%	0.014%
CLTA4-3-3	<u>cCAGATGTAGTatTTCCA</u> <u>CAGGG</u>	61	CELF1	1.00	4.95	0%	0%	0.469%
CLTA4-3-4	<u>GCAGttTAGTGTTC</u> AGGG	62	BC073807	0.79	3.12	0%	0%	0%
CLTA4-3-5	GCAGAGtTAGTGTTC CACaG	63	MPPED2	0	1.22	0.005%	0.015%	0.018%
CLTA4-3-6	GCAGATGgAGgGTtCA CAGGG	64	DCHS2	1.57	1.17	0.015%	0.023%	0.021 %

(continued)

	# of Mutations	sequence	SEQ ID NO.	gene	v1.0	v2.1	in vitro enrichment			modification frequency in HEK293T cells		P-value
							no sgRNA	v1.0	v2.1	0.012%	0.005%	
CLTA4-3-7	3	GgAaATTAGTGTTCACAGGG	65		0.43	0.42	0.005%	0.012%	0.010%	0.003%	0.003%	
CLTA4-4-1	4	aaAGAAgTAGTaTTTCCA CATGG	66			n.t.	n.t.	n.t.	n.t.	n.t.	n.t.	
CLTA4-4-2	4	aaAGATGTAgtCTTTCCACAAAGG	67				0.004%	0%	0%	0.005%	0.005%	
CLTA4-4-3	4	aaAtATGTAgtCTTTCCACAGGG	68				0.004%	0.009%	0.009%	0%	0%	
CLTA4-4-4	4	atAGATGTAgtCTTTCCAAAGGa	69	NR1H4			0.032%	0.006%	0.006%	0.052%	0.052%	
CLTA4-4-5	4	ccAGAgGTAGTGctTcccACAGGG	70				0.005%	0.006%	0.006%	0.007%	0.007%	
CLTA4-4-6	4	ccAGATGTagGTTCACAAAGG	71	XKR6			0.018%	0%	0%	0.007%	0.007%	

(continued)

	# of Mutations	sequence	SEQ ID NO.	gene	v1.0	v2.1	in vitro enrichment			modification frequency in HEK293T cells	P-value
							no sgRNA	v1.0	v2.1		
CLTA4-4-7	4	ctAcATGTTAGTGTTCAt ATGG	SEQ ID NO: 72	HKR1			0.006%	0%	0.008%		
CLTA4-4-8	4	ctAGATGaaAGTGtcTTCCA CATGG	SEQ ID NO: 73	CDK8			0.009%	0.013%	0.730%		9.70E-21
CLTA4-4-9	4	GAAaATGgAGTGTtAca CATGG	SEQ ID NO: 74				0%	0%	0.004%		
CLTA4-4-10	4	GCAaATGTAatTaTTTCCA CAAGG	SEQ ID NO: 75				0.004%	0%	0%		
CLTA4-4-11	4	GCAaATGTAAGctTTTgAC CtAGG	SEQ ID NO: 76	NOV			0%	0.00%	0%		
CLTA4-4-12	4	GCAGATGTAGGctTTTgAC ATGG	SEQ ID NO: 77				0%	0.00%	0%		
CLTA4-4-13	4	GCAGCTtaAGTGTTTtCAC ATGG	SEQ ID NO: 78	GRHL2			0.020%	0.02%	0.030%		

(continued)

# of Mutations	sequence	SEQ ID NO.	gene	in vitro enrichment		modification frequency in HEK293T cells		P-value
				v1.0	v2.1	no sgRNA	v1.0	
CLTA4-4-14	ttAcTGTAGTGTTCaCAC ACGG	LINC00535				n.t.	n.t.	
CLTA4-4-1	GaAGAgGaaAGTGTgCc CAGGG	SEQ ID NO: 80	RNHI			0.004%	0.01%	0.006%
CLTA4-5-1	GaAGATGTgGaGTgCaA CATGG	SEQ ID NO: 81	FZD3			0.004%	0.00%	0%
CLTA4-5-2	GCAGAAaGTActGTggtAC AAGG	SEQ ID NO: 82				0.002%	0.00%	0.003%
CLTA4-5-3	GCAGATGTgGaaTTaCaA CAGGG	SEQ ID NO: 83	SLC9A2			0%	0.00%	0%
CLTA4-5-4	GCAAGtaTAGTGtaTaCaA CATGG	SEQ ID NO: 84				0.004%	0.00%	0.005%
CLTA4-5-5	taAGATGTAGTaTTCCA aAAAt	SEQ ID NO: 85				0.007%	0.01%	0%

(continued)

5
10
15
20
25
30
35
40
45
50
55

#of Mutations	sequence	SEQ ID NO.	gene	in vitro enrichment			modification frequency in HEK293T cells		P-value
				v1.0	v2.1	no sgRNA	v1.0	v2.1	
CLTA4-6-1	6	GCAGctTGgcaTttcTCCACACGG					n.t.	n.t.	
CLTA4-6-2	6	GgAGATcTgaTGgTTTCAAGG	SEQ ID NO: 87				0.007%	0.00%	0.009%
CLTA4-6-3	6	taAaATTGcAGTGTatTCCAtATGG	SEQ ID NO: 88	SMA4			0.015%	0.00%	0%
CLTA4-7-1	7	GCcagaaTAGTTTCCaACAAAGG	SEQ ID NO: 89	SEPHS2			0%	0.00%	0.007%
CLTA4-7-2	8	tgtATTAGaGaTTtgCACAAGG	SEQ ID NO: 90	RORB			0%	0.00%	0%

EP 3 486 318 B1

Table 2

100 nM Cas9:CLTA V2.1 sRNA			100 nM Cas9:CLTA V2.1 gRNA			100 nM Cas9:CLTA V2.1 gRNA		
Position	-4	-1	-4	-1	-1	-4	-1	-4
A	202151	245368	9138+04	2	3	3	1	2
B	2105120	241259	274819	737834	13618	13830	5616+03	107832
C	182376	241259	18774	189614	186532	9388+05	28572	107420
D	121701	18774	187534	186532	186533	130495	132543	293767
E	460761	461193	475464	913834	913834	13275	13275	329375
F	942707	372834	387466	1108406	112024	348565	137704	108370
G	1000 nM Cas9:CLTA V1.0 sRNA			1000 nM Cas9:CLTA V1.0 gRNA			1000 nM Cas9:CLTA V2.1 gRNA	
Position	-4	-3	-2	-1	1	2	3	4
A	219833	263464	207913	97886	20663	39724	1151200	42986
B	226763	20815	311444	353414	38986	11093	115475	118540
C	300611	467791	470094	576062	575771	112930	12594	112350
D	1000 nM Cas9:CLTA V2.1 sRNA			1000 nM Cas9:CLTA V2.1 gRNA			1000 nM Cas9:CLTA V1.0 gRNA	
Position	-4	-3	-2	-1	1	2	3	4
A	168775	205343	165197	207798	75243	67180	24410	36549
B	197800	285445	243886	241777	80375	80375	80366	80366
C	197806	285323	189323	581721	80368	80368	80367	80367
D	164077	361452	361452	361452	322595	322595	322595	322595
E	770275	770275	302144	770275	322595	322595	322595	322595
F	357359	357359	357359	357359	357359	357359	357359	357359
G	357359	357359	357359	357359	357359	357359	357359	357359
H	CLTA pre-selection library			CLTA pre-selection library			CLTA pre-selection library	
Position	-4	-3	-2	-1	1	2	3	4
A	1422	1322	1222	1122	1013	9103	8103	7103
B	142705	132554	132549	132544	132535	132535	132535	132535
C	137801	132829	132829	132829	132830	132830	132830	132830
D	284866	261316	261315	261315	244854	244854	244854	244854
E	244854	244854	244854	244854	586511	586511	586511	586511
F	573870	488914	488914	488914	573870	488914	488914	573870
G	573870	488914	488914	488914	573870	488914	488914	573870
H	100 nM Cas9:CLTA V2.1 gRNA			100 nM Cas9:CLTA V2.1 gRNA			100 nM Cas9:CLTA V2.1 gRNA	
Position	-4	-3	-2	-1	1	2	3	4
A	59030	76101	78987	81833	81833	81833	81833	81833
B	168401	182082	134980	134985	22088	428623	44498	1216+04
C	85032	75735	829459	133275	416562	470714	470714	470714
D	240525	244794	259452	251515	42090	26966	26966	26966
E	26966	26966	26966	26966	54511	54511	54511	54511
F	54511	54511	54511	54511	54511	54511	54511	54511
G	54511	54511	54511	54511	54511	54511	54511	54511
H	100 nM Cas9:CLTA V1.0 sRNA			100 nM Cas9:CLTA V1.0 gRNA			100 nM Cas9:CLTA V2.1 gRNA	
Position	-4	-3	-2	-1	1	2	3	4
A	98188	141261	142569	141850	151224	116745	528773	50295
B	130605	247215	260075	243819	34420	56151	527725	526660
C	300524	153473	153473	153473	742232	85777	86777	86777
D	1438973	1438973	421524	439767	388061	711197	51206	31918
E	532570	532570	532570	532570	532570	532570	532570	532570
F	532570	532570	532570	532570	532570	532570	532570	532570
G	532570	532570	532570	532570	532570	532570	532570	532570
H	532570	532570	532570	532570	532570	532570	532570	532570
I	CLTA pre-selection library			CLTA pre-selection library			CLTA pre-selection library	
Position	-4	-3	-2	-1	1	2	3	4
A	47374	70467	71355	72686	12180	27654	45457	17740
B	130798	13157	13157	13157	12180	32159	25159	21818
C	222051	214104	224296	224296	159320	41376	39159	36119
D	147974	147974	147974	147974	370924	42360	45387	40917
E	45387	45387	45387	45387	45387	45387	45387	45387
F	45387	45387	45387	45387	45387	45387	45387	45387
G	45387	45387	45387	45387	45387	45387	45387	45387
H	45387	45387	45387	45387	45387	45387	45387	45387

Table 2 (continued)

			no sgRNA	v1.0 sgRNA	v2.1 sgRNA
	m	sequence	modified sequences total sequences	modified sequences total sequences	modified sequences total sequences
CLTA1-0-1	0	<u>AGTCCTCATCTCCCTCAAGCAGG</u> (SEQ ID NO:91)	2 58889	18 42683	178 52845
CLTA1-1-1	1	<u>AGTCCTCAaCTCCCTCAAGCAGG</u> (SEQ ID NO:92)	1 39804	9 29000	37 40588
CLTA1-2-1	2	AGcCCTCATTCCCTCAAGCAGG (SEQ ID NO:93)	0 16276	0 15032	0 18277
CLTA1-2-2	2	<u>AcTCCTCATCcCCCTCAAGCCGG</u> (SEQ ID NO:94)	3 21267	1 20042	33 22579
CLTA1-2-3	2	AGTCatCATCTCCCTCAAGCAGa (SEQ ID NO:95)	0 0	0 0	0 0
CLTA1-3-1	3	cGTCCCTCcTCTCCCCaAGCAGG (SEQ ID NO:96)	2 53901	0 42194	0 52205
CLTA1-3-2	3	tGTCCCTCtTCTCCCTCAAGCAGa (SEQ ID NO:97)	0 14890	0 14231	0 15937
CLTA1-4-1	4	AagCtTCATCTCTCAAGCTGG (SEQ ID NO:98)	0 49579	2 31413	0 41234
CLTA1-4-2	4	AGTaCTCttTCCCTCAgGCTGG (SEQ ID NO:99)	2 30013	1 23470	4 26999
CLTA1-4-3	4	AGTctTaAatTCCCTCAAGCAGG (SEQ ID NO:100)	2 63792	0 52321	1 73007
CLTA1-4-4	4	AGTgCTCATCTaCCagAAGCTGG (SEQ ID NO:101)	1 12585	0 11339	0 12066
CLTA1-4-5	4	ccTCCTCATCTCCCTgcAGCAGG (SEQ ID NO:102)	4 30568	1 23810	0 27870
CLTA1-4-6	4	ctaCaTCATCTCCCTCAAGCTGG (SEQ ID NO:103)	0 13200	1 12886	2 12843
CLTA1-4-7	4	gGTCCTCATCTCCCTaAAacAGa (SEQ ID NO:104)	1 8697	3 8188	0 8783
CLTA1-4-8	4	tGTCCCTCATCggCCTCAgGCAGG (SEQ ID NO:105)	0 13169	0 8805	2 12830
CLTA1-5-1	5	AGaCacCATCTCCCTtgAGCTGG (SEQ ID NO:106)	0 46109	1 32515	2 35567
CLTA1-5-2	5	AGgCaTCATCTaCaTCAAGtTGG (SEQ ID NO:107)	0 41280	0 28896	0 35152
CLTA1-5-3	5	AGTaaTCActTCCaTCAAGCCGG (SEQ ID NO:108)	0 0	0 0	0 0
CLTA1-5-4	5	tccCCTCAcCTCCCTaAAGCAGG (SEQ ID NO:109)	2 24169	5 17512	1 23483
CLTA1-5-5	5	tGTctTtATTCCCTCtAGCTGG (SEQ ID NO:110)	0 11527	0 10481	1 11027
CLTA1-6-1	6	AGTCCTCATCTCCCTCAAGCAGG (SEQ ID NO:111)	0 6537	0 5654	0 6741

Table 3

			no sgRNA	v1.0 sgRNA	v2.1 sgRNA	
	m	sequence	modified sequences total sequences	modified sequences total sequences	modified sequences total sequences	
5						
10						
CLTA4-0-1	0	<u>GCAGATGTAGTGT</u> TTCCACAGGG (SEQ ID NO:112)	6 <u>29191</u>	2005 <u>18640</u>	14970 <u>19661</u>	
CLTA4-3-1	3	<u>aC</u> atATGTAGT <u>a</u> TTTCCACAGGG (SEQ ID NO:113)	2 <u>34165</u>	11 <u>20018</u>	3874 <u>16082</u>	
CLTA4-3-2	3	GCAtATGTAGTGT <u>TTCCA</u> ATGt (SEQ ID NO:114)	3 17923	0 11688	2 13880	
CLTA4-3-3	3	<u>c</u> CAGATGTAGT <u>a</u> TT <u>c</u> CCACAGGG (SEQ ID NO:115)	0 <u>16559</u>	0 <u>12007</u>	52 <u>11082</u>	
CLTA4-3-4	3	GCAGT <u>T</u> TAGTGT <u>TTt</u> CACAGGG (SEQ ID NO:116)	0 21722	0 12831	0 15726	
CLTA4-3-5	3	GCAGA <u>g</u> tTAGTGT <u>TTCC</u> ACACaG (SEQ ID NO:117)	1 21222	2 13555	3 16425	
CLTA4-3-6	3	GCAGATG <u>g</u> AGgGT <u>TTt</u> CACAGGG (SEQ ID NO:118)	3 20342	3 12804	3 14068	
CLTA4-3-7	3	GgAaAT <u>t</u> TAGTGT <u>TTCC</u> CACAGGG (SEQ ID NO:119)	2 38894	3 24017	1 29347	
CLTA4-4-1	4	aaAGAaGTAGT <u>a</u> TTTCCACATGG (SEQ ID NO:120)	0 0	0 0	0 0	
CLTA4-4-2	4	aaAGATGTAGT <u>c</u> TTCCACAAGG (SEQ ID NO:121)	1 27326	0 17365	1 18941	
CLTA4-4-3	4	aaAtATGTAGT <u>c</u> TTTCCACAGGG (SEQ ID NO:122)	2 46232	3 32264	0 32638	
CLTA4-4-4	4	atAGATGTAGTGT <u>TTCCA</u> aAGGa (SEQ ID NO:123)	9 27821	1 16223	8 15388	
CLTA4-4-5	4	<u>c</u> CAGA <u>g</u> GTAGT <u>G</u> c <u>T</u> CCACAGGG (SEQ ID NO:124)	1 20979	1 15674	1 15086	
CLTA4-4-6	4	<u>c</u> CAGATGT <u>g</u> agGT <u>TTCC</u> CACAA <u>GG</u> (SEQ ID NO:125)	4 22021	0 15691	1 14253	
CLTA4-4-7	4	ctAcATGTAGTGT <u>TTCC</u> AtATGG (SEQ ID NO:126)	2 35942	0 23076	1 11867	
CLTA4-4-8	4	<u>ct</u> AGAT <u>G</u> aAGT <u>G</u> c <u>TTCC</u> CACATGG (SEQ ID NO:127)	1 10692	1 7609	59 8077	
CLTA4-4-9	4	GaAaATG <u>g</u> AGTGT <u>TTTa</u> CACATGG (SEQ ID NO:128)	0 34616	0 22302	1 24671	
CLTA4-4-10	4	GCAaAT <u>G</u> aAGTGT <u>ca</u> CCACA <u>AGG</u> (SEQ ID NO:129)	1 25210	0 16187	0 16974	
CLTA4-4-11	4	GCAaATGT <u>A</u> t <u>T</u> TTCC <u>ACT</u> AGG (SEQ ID NO:130)	0 34144	1 24770	0 22547	
CLTA4-4-12	4	GCAGATGTAG <u>c</u> t <u>TTgt</u> ACATGG (SEQ ID NO:131)	0 14254	0 9616	0 9994	
CLTA4-4-13	4	GCAG <u>c</u> T <u>ta</u> AGTGT <u>TTt</u> CACATGG (SEQ ID NO:132)	8 39466	1 7609	5 16525	
CLTA4-4-14	4	ttAcATGTAGTGT <u>TTTa</u> CACACGG (SEQ ID NO:133)	0 0	0 22302	0 0	
CLTA4-5-1	5	GaAGA <u>g</u> GaAGTGT <u>TTg</u> C <u>c</u> CAGGG (SEQ ID NO:134)	1 27616	1 16319	1 16140	
CLTA4-5-2	5	GaAGATGT <u>g</u> GaGTT <u>ga</u> CACATGG (SEQ ID NO:135)	1 22533	0 14292	0 15013	

5	CLTA4-5-3	5	GCAGAaGTAcTGTGttACAAGG (SEQ ID NO:136)	1	44243	1	29391	1	29734
10	CLTA4-5-4	5	GCAGATGTgGaaTTaCaACAGGG (SEQ ID NO:137)	0	27321	0	13640	0	14680
15	CLTA4-5-5	5	GCAGtcaTAGTGtaTaCACATGG (SEQ ID NO:138)	1	26538	0	18449	1	20559
20	CLTA4-5-6	5	taAGATGTAGTaTTTCCAaAAGt (SEQ ID NO:139)	1	15145	1	8905	0	7911
25	CLTA4-6-1	6	GCAGcTGgcaTtTcTCCACACGG (SEQ ID NO:140)	0	2	0	0	0	0
30	CLTA4-6-2	6	GgAGATcTgaTGTTCTACAAGG (SEQ ID NO:141)	2	27797	0	19450	2	21709
35	CLTA4-6-3	6	taAaATGcAGTGtaTCCAtATGG (SEQ ID NO:142)	4	27551	0	18424	0	18783
40	CLTA4-7-1	7	GCcagaaTAGTtTTCaACAAGG (SEQ ID NO:143)	0	20942	0	13137	1	13792
45	CLTA4-7-2	8	ttgtATTAGaGaTTgCACAAGG (SEQ ID NO:144)	0	28470	0	18104	0	20416

Table 4**Table 5**

Off-target site	Human genome coordinates
CLTA1-0-1	9(+): 36,211,732-36,211,754 12(+): 7,759,893-7,759,915
CLTA1-1-1	8(-): 15,546,437-15,546,459
CLTA1-2-1	3(-): 54,223,111-54,223,133
CLTA1-2-2	15(+): 89,388,670-89,388,692
CLTA1-2-3	5(+): 88716920-88,716,942
CLTA1-3-1	21(+): 27,972,462-27,972,484
CLTA1-3-2	4(-): 17,179,924-17,179,946
CLTA1-4-1	1(+): 147,288,742-147,288,764
CLTA1-4-2	10(+): 97,544,444-97,544,466
CLTA1-4-3	2(-): 161,873,870-161,873,892
CLTA1-4-4	1(+): 196,172,702-196,172,724
CLTA1-4-5	13(+): 56,574,636-56,574,658
CLTA1-4-6	2(+): 241,357,827-241,357,849
CLTA1-4-7	3(+): 121,248,627-121,248,649
CLTA1-4-8	12(+): 132,937,319-132,937,341
CLTA1-5-1	9(-): 80,930,919 -80,930,941
CLTA1-5-2	2(+): 140,901,875-14,0901,897
CLTA1-5-3	3(+): 45,016,841-45,016,863
CLTA1-5-4	X(+): 40,775,684-40,775,706
CLTA1-5-5	2(-): 185,151,622-185,151,644
CLTA1-6-1	X(+): 150,655,097-150,655,119
CLTA4-0-1	9(-): 36,211,779-36,211,801

(continued)

	Off-target site	Human genome coordinates
5	CLTA4-3-1	12(-): 50,679,419-50,679,441
	CLTA4-3-2	X(-): 143,939,483-143,939,505
	CLTA4-3-3	11(-): 47,492,611-47,492,633
10	CLTA4-3-4	3(-): 162,523,715-162,523,737
	CLTA4-3-5	11(+): 30,592,975-30,592,997
	CLTA4-3-6	4(-): 155,252,699-155,252,721
	CLTA4-3-7	18(+): 39,209,441-39,209,463
15	CLTA4-4-1	17(-): 36,785,650-36,785,672
	CLTA4-4-2	1(-): 241,537,119-241,537,141
	CLTA4-4-3	8(-): 120,432,103-120,432,125
	CLTA4-4-4	6(-): 106,204,600-106,204,622
20	CLTA4-4-5	8(+): 102,527,804-102,527,826
	CLTA4-4-6	8(-): 94,685,538-94,685,560
	CLTA4-4-7	2(+): 35,820,054-35,820,076
25	CLTA4-4-8	3(-): 36,590,352-36,590,374
	CLTA4-4-9	12(+): 100,915,498-100,915,520
	CLTA4-4-10	21(+): 33,557,705-33,557,727
	CLTA4-4-11	8(+): 10,764,183-10,764,205
30	CLTA4-4-12	19(+): 37,811,645-37,811,667
	CLTA4-4-13	13(-): 26,832,673-26,832,695
	CLTA4-4-14	6(+): 19,349,572-19,349,594
35	CLTA4-5-1	11 (-): 502,300-502,322
	CLTA4-5-2	8(-): 28,389,683-28,389,705
	CLTA4-5-3	2(-): 118,557,405-118,557,427
40	CLTA4-5-4	2(-): 103,248,360-103,248,382
	CLTA4-5-5	21(-): 42,929,085-42,929,107
	CLTA4-5-6	13(-): 83,097,278-83,097,300
	CLTA4-6-1	2(+): 43,078,423-43,078,445
45	CLTA4-6-2	7(-): 11,909,384-11,909,406
	CLTA4-6-3	5(-): 69,775,482-69,775,504
	CLTA4-7-1	16(+): 30,454,945-30,454,967
50	CLTA4-7-2	9(-): 77,211,328-77,211,350

Table 6

				<i>in vitro</i> enrichment	modification frequency in HEK293T cells
	number of mutations	sequence	gene	v1.0	v2.1
CLTA1-0-1	0	AGTCCTCATCTCCCTCAAGCAGG (SEQ ID NO:145)	<u>CLTA</u>	<u>41.4</u>	<u>2.3.3</u>
CLTA1-1-1	1	AGTCCTCAaCTCCCTCAAGCAGG (SEQ ID NO:146)	<u>TUSC3</u>	<u>25.9</u>	<u>14</u>
CLTA1-2-1	2	AGCCCTCATTTCCCTCAAGCAGG (SEQ ID NO:147)	CACNA2D3	15.4	0%
CLTA1-2-2	2	AcTCCCTCATCCCTCAAGCAGG (SEQ ID NO:148)	<u>ACAN</u>	<u>29.2</u>	<u>18.8</u>
CLTA1-2-3	2	AGTCATCATCTCCCTCAAGCAGA (SEQ ID NO:149)		0.06	1.27
CLTA1-3-1	3	cGTCCTCCCTCCCAAGCAGG (SEQ ID NO:150)		0	2.07
CLTA1-3-2	3	tGTCCCTtTCTCCCTCAAGCAGA (SEQ ID NO:151)	BC029598	0	1.47
CLTA1-4-1	4	AagCtTCATCTCTCAAGCTGG (SEQ ID NO:152)			0%
CLTA1-4-2	4	AGTaCTCTTCCCTCAGGCTGG (SEQ ID NO:153)	ENTPD1		0.007%
CLTA1-4-3	4	AGTCttaAatTCCTCAAGCAGG (SEQ ID NO:154)			0.003%
CLTA1-4-4	4	AGTgCTCATCTaCCagAAGCTGG (SEQ ID NO:155)			0.008%
CLTA1-4-5	4	ccTCCCTCATCTCCCTtgAGCAGG (SEQ ID NO:156)			0.013%
CLTA1-4-6	4	ctacAtCATCTCCCTCAAGCTGG (SEQ ID NO:157)			0%
CLTA1-4-7	4	gGTCCTCATCTCCCTtaAAaCAGA (SEQ ID NO:158)	POLQ (coding)		0.011%
CLTA1-4-8	4	tGTCCTCATCtgGCCCTAGGCTGG (SEQ ID NO:159)			0%
CLTA1-5-1	5	AGaCacCATCTCCCTtgAGCTGG (SEQ ID NO:160)	PSAT1		0%
CLTA1-5-2	5	AGgCaTCATCTaCaTCAAAGtTGG (SEQ ID NO:161)			0%
CLTA1-5-3	5	AGTaTCACtCCatTCAAAGCCGG (SEQ ID NO:162)	ZDHHC3, EXOSC7	n.t.	n.t.
CLTA1-5-4	5	tccCCTCACCCCTaAAGCAGG (SEQ ID NO:163)		0.008%	0.029%
CLTA1-5-5	5	tGTCTttaATTCCCTCTAGCTGG (SEQ ID NO:164)		0%	0.004%
CLTA1-6-1	6	AGTCCTCATCTCCCTCAAGCAGG (SEQ ID NO:165)		0%	0.009%

Table 7

CLTA1-0-1		# of sequences		
	sequence	no sgRNA	v1.0 sgRNA	v2.1 sgRNA
5	ref AGTCCTCATCTCCCTCAAGCAGG (SEQ ID NO:166)	58,887	42,665	52,667
	AGTCCTCATCTCCCTCA <u>A</u> AGCAGG (SEQ ID NO:167)	0	0	66
10	AGTCCTCATCTCCCTC-AGCAGG (SEQ ID NO:168)	0	2	28
	AGTCCTCAT-----	0	0	13
15	AGTCCTCATCTCCCT <u>C</u> AGCAGG (SEQ ID NO:169)	0	0	11
	AGTCCTCAT-----AGCAGG (SEQ ID NO:170)	0	0	9
20	AGTCCTCATCT-----AGCAGG (SEQ ID NO:171)	0	0	8
	AGTCCTCA-----AGCAGG (SEQ ID NO:172)	0	0	6
25	AGTCCTCATCTCCCTCA <u>AAGGCAGTGTGTTGTT</u> <u>ACTTGAGTTGTC</u> AGCAGG (SEQ ID NO:173)	0	0	4
	AGTCCTCATCTCCCTC <u>T</u> ATTAGCAGG (SEQ ID NO:174)	0	0	4
30	AGTCCTCATCTCCCTCA <u>GGGCTTGTGTTACAGC</u> <u>TCACCTTGAATTG</u> CACAAGCGTGCA <u>AGCAG</u> G (SEQ ID NO:175)	0	0	3
	AGTCCTCATCTCCCT-AGCAGG (SEQ ID NO:176)	0	11	0
35	AGTCCTCATCCCTC-AAGCAGG (SEQ ID NO:177)	0	3	0
	AGTCCTCATCTCCCT-AAGCAGG (SEQ ID NO:178)	1	2	0
40	other	1	0	26
	modified total	2	18 (0.042%)	178 (0.34%)
CLTA1-1-1		# of sequences		
	sequence	no sgRNA	v1.0 sgRNA	v2.1 sgRNA
45	ref AGTCCTCAaCTCCCTCAAGCAGG (SEQ ID NO:179)	39,803	28,991	40,551
	AGTCCTCAaCTCCCTCA <u>A</u> AGCAGG (SEQ ID NO:180)	0	4	13
50	AGTCCTCAaCTCCCTCA----- (SEQ ID NO:181)	0	0	12
	AGTCCTCAaCTCCCTC-AGCAGG (SEQ ID NO:182)	0	2	4
55	AGTCCTCAaCTCCCTCA <u>AGAAAGGTGTTGAAAAA</u> <u>TCAGAAAGAGAGAAACA</u> AGCAGG (SEQ ID NO:183)	0	0	3
	AGTCCTCAaCTCCCTCA <u>ATCTACGGTCCATTCC</u> <u>CGTTTCCACTCACCTTGCGCCGC</u> AGCAGG (SEQ ID NO:184)	0	0	2
	AGTCCTCAaCTCCCT-AAGCAGG (SEQ ID NO:185)	0	3	1
	AGTCCTCAaCTCCCTCA <u>ACCAACTTTAACATCC</u> <u>TGCTGGTTCTGTCAATTAAAGTTGAA</u> AGCAGG (SEQ ID NO:186)	0	0	1
	AGTCCTCAaCTCCCTCA <u>CAGCAAATAAAAAAGT</u> <u>TGTTTATGCATATTCAAGATAAGCAA</u> AGCAGG (SEQ ID NO:187)	0	0	1

(continued)

CLTA1-1-1		# of sequences		
	sequence	no sgRNA	v1.0 sgRNA	v2.1 sgRNA
5	AGTCCTCAaCTCCC-AAGCAGG (SEQ ID NO:188)	1	0	0
	modified total	1	9 (0.031 %)	37 (0.091 %)
CLTA1-2-2		# of sequences		
10	sequence	no sgRNA	v1.0 sgRNA	v2.1 sgRNA
ref	AcTCCTCATCcCCCTCAAGCCGG (SEQ ID NO:189)	21,264	20,041	22,546
	AcTCCTCATCcCCCT <u>A</u> AGCCGG (SEQ ID NO:190)	0	0	8
15	AcTCCTCATCcCCCT <u>C</u> <u>A</u> GAGCCGG (SEQ ID NO:191)	0	0	7
	AcTCCTC-----AGCCGG (SEQ ID NO:192)	0	0	5
	AcTCCTCATCcCCCT <u>A</u> <u>A</u> AGCCGG (SEQ ID NO:193)	0	0	2
20	AcTCCTCATCcCCCT <u>C</u> <u>A</u> GAGCCGG (SEQ ID NO:194)	0	0	2
	AcTCCTCATCcCCCT <u>C</u> <u>T</u> AGCCGG (SEQ ID NO:195)	0	0	2
	AcTCCTCATCcCCCT <u>C</u> <u>T</u> <u>C</u> AGCCGG (SEQ ID NO:196)	0	0	2
25	AcTCCTCATCcC-----AGCCGG (SEQ ID NO:197)	0	0	2
	AcTCCTCATCcCCCTA-AGCCGG (SEQ ID NO:198)	3	1	1
	AcTCCTCATCcCCCT <u>A</u> <u>T</u> AGCCGG (SEQ ID NO:199)	0	0	1
30	AcTCCTCACCC <u>C</u> <u>A</u> GAGCCGG (SEQ ID NO:200)	0	0	1
	modified total	3	1	33 (0.15%)

Table 8

CLTA4-0-1		# of sequences		
	sequence	control	v1.0 sqRNA	v2.1 sqRNA
35	GCAGATGTAGTGTTCACAGGG (SEQ ID NO:201)	29,185	16,635	17,555
	GCAGATGTAGTGTTC-ACAGGG (SEQ ID NO:202)	1	891	5,937
40	GCAGATGTAGTGT <u>T</u> <u>C</u> CACAGGG (SEQ ID NO:203)	0	809	5,044
	GCAGATGTAGTG----CACAGGG (SEQ ID NO:204)	0	14	400
	GCAGATGTAGTGTTC-CAGGG (SEQ ID NO:205)	0	19	269
45	GCAGATGTAC-----ACAGGG (SEQ ID NO:206)	0	17	262
	GCAGATGTAGTGTCA---CAGGG (SEQ ID NO:207)	2	6	254
	GCAGATGTAGTGTCA-CAGGG (SEQ ID NO:208)	0	21	229
50	GCAGATGTAGTGTTC-CAGGG (SEQ ID NO:209)	1	14	188
	GCAGATGTAGT----CACAGGG (SEQ ID NO:210)	0	0	152
	GCAGATGT-----AGGG (SEQ ID NO:211)	0	6	129
55	other	2	208	2,106
	modified total	6	2,005 (11 %)	14,970 (76%)

(continued)

CLTA4-3-1		# of sequences		
	sequence	control	v1.0 sqRNA	v2.1 sqRNA
5	ref aCAtATGTAGTaTTTCCACAGGG (SEQ ID NO:212)	34,163	20,007	12,208
	a CAtATGTAGTaTTT CC CACAGGG (SEQ ID NO:213)	0	8	1779
10	aCAtATGTAGTaTTTCA-CAGGG (SEQ ID NO:214)	1	0	293
	aCAtATGTAGTaTTTC-CAGGG (SEQ ID NO:215)	1	0	227
15	aCAtAT-----CACAGGG (SEQ ID NO:216)	0	0	117
	a-----CAGGG	0	0	96
20	aCAt-----CACAGGG (SEQ ID NO:217)	0	0	78
	aCAtATGTAGT----CACAGGG (SEQ ID NO:218)	0	0	77
25	aCAtATGTAGTaTTTCC-(SEQ ID NO:219)	0	0	76
	aCAtATGT-----AGGG (SEQ ID NO:220)	0	0	68
30	aCAtATGTAG-----CACAGGG (SEQ ID NO:221)	0	0	64
	other	0	3	999
35	modified total	2	11 (0.055%)	3874 (24%)
CLTA4-3-3		# of sequences		
	sequence	control	v1.0 sqRNA	v2.1 sqRNA
40	ref ccAGATGTAGTaTTcccACAGGG (SEQ ID NO:222)	16,559	12,007	11,030
	ccAGATGTAGTaTT cc ACAGGG (SEQ ID NO:223)	0	0	35
45	ccAGATGTAGTaT---ACAGGG (SEQ ID NO:224)	0	0	5
	ccAGATGTAGTaT--CACAGGG (SEQ ID NO:225)	0	0	3
50	ccAGATGTAGTaTT cc AACACAGGG (SEQ ID NO:226)	0	0	2
	ccAGATGTAGTaTT-CACAGGG (SEQ ID NO:227)	0	0	2
55	ccAGATGTAGTaTTcC-CAGGG (SEQ ID NO:228)	0	0	2
	ccAGATGTA-----	0	0	2
	ccAGATGTAGTaTTcC-ACAGGG (SEQ ID NO:229)	0	0	1
	modified total	0	0	52 (0.47%)
CLTA4-4-8		# of sequences		
	sequence	control	v1.0 sqRNA	v2.1 sqRNA
55	ref ctAGATGaAGTGcTTCCACATGG (SEQ ID NO:230)	10,691	7,608	8,018
	ctAGATGaAGTGcTT CC CACATGG (SEQ ID NO:231)	0	0	49
	ctAGATGaAGTGcTTC-ACATGG (SEQ ID NO:232)	0	0	6
	ctAGATGaAGTG-(SEQ ID NO:233)	0	0	2
	ctAGATGaAGTGcTT CC AACACATGG (SEQ ID NO:234)	0	0	1
	ctAGATGaAGTGcTTC-CATGG (SEQ ID NO:235)	1	0	0
	ctAGATGaAGTGcTTCC-CATGG (SEQ ID NO:236)	0	1	0
	modified total	1	1	59 (0.73%)

Table 9

oligonucleotide name	oligonucleotide sequence (5'→3')
CLTA1 v2.1 template fwd	TAA TAC GAC TCA CTA TAG GAG TCC TCA TCT CCC TCA AGC GTT TTA GAG CTA TGC TG (SEQ ID NO:237)
CLTA2 v2.1 template fwd	TAA TAC GAC TCA CTA TAG GCT CCC TCA AGC AGG CCC CGC GTT TTA GAG CTA TGC TG (SEQ ID NO:238)
CLTA3 v2.1 template fwd	TAA TAC GAC TCA CTA TAG GTG TGA AGA GCT TCA CTG AGT GTT TTA GAG CTA TGC TG (SEQ ID NO:239)
CLTA4 v2.1 template fwd	TAA TAC GAC TCA CTA TAG GGC AGA TGT AGT GTT TCC ACA GTT TTA GAG CTA TGC TG (SEQ ID NO:240)
v2.1 template rev	GAT AAC GGA CTA GCC TTA TTT TAA CTT GCT ATG CTT TTC AGC ATA GCT CTA AAA C (SEQ ID NO:241)
CLTA1 v1.0 template	CGG ACT AGC CTT ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC GCT TGA GGG AGA TGA GGA CTC CTA TAG TGA GTC GTA TTA (SEQ ID NO:242)
CLTA2 v1.0 template	CGG ACT AGC CTT ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC GCG GGG CCT GCT TGA GGG AGC CTA TAG TGA GTC GTA TTA (SEQ ID NO:243)
CLTA3 v1.0 template	CGG ACT AGC CTT ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC ACT CAG TGA AGC TCT TCA CAC CTA TAG TGA GTC GTA TTA (SEQ ID NO:244)
CLTA4 v1.0 template	CGG ACT AGC CTT ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC TGT GGA AAC ACT ACA TCT GCC CTA TAG TGA GTC GTA TTA (SEQ ID NO:245)
T7 promoter oligo	TAA TAC GAC TCA CTA TAG G (SEQ ID NO:246)
CLTA1 lib	/5Phos/AAC ACA NNN NC*C* NG*C* T*T*G* A*G*G* G*A*G* A*T*G* A*G*G* A*C*T* NNN NAC CTG CCG AGA ACA CA (SEQ ID NO:247)
CLTA2 lib	/5Phos/TCT TCT NNN NC*C* NG*C* G*G*G* G*C*C* T*G*C* T*T*G* A*G*G* G*A*G* NNN NAC CTG CCG AGT CTT CT (SEQ ID NO:248)
CLTA3 lib	/5Phos/AGA GAA NNN NC*C* NA*C* T*C*A* G*T*G* A*A*G* C*T*C* T*T*C* A*C*A* NNN NAC CTG CCG AGA GAG AA (SEQ ID NO:249)
CLTA4 lib	/5Phos/TTG TGT NNN NC*C* NT*G* T*G*G* A*A*A* C*A*C* T*A*C* A*T*C* T*G*C* NNN NAC CTG CCG AGT TGT GT (SEQ ID NO:250)
CLTA1 site fwd	CTA GCA GTC CTC ATC TCC CTC AAG CAG GC (SEQ ID NO:251)
CLTA1 site rev	AGC TGC CTG CTT GAG GGA GAT GAG GAC TG (SEQ ID NO:252)
CLTA2 site fwd	CTA GTC TCC CTC AAG CAG GCC CCG CTG GT (SEQ ID NO:253)
CLTA2 site rev	AGC TAC CAG CGG GGC CTG CTT GAG GGA GA (SEQ ID NO:254)
CLTA3 site fwd	CTA GCT GTG AAG AGC TTC ACT GAG TAG GA (SEQ ID NO:255)
CLTA3 site rev	AGC TTC CTA CTC AGT GAA GCT CTT CAC AG (SEQ ID NO:256)
CLTA4 site fwd	CTA GTG CAG ATG TAG TGT TTC CAC AGG GT (SEQ ID NO:257)
CLTA4 site rev	AGC TAC CCT GTG GAA ACA CTA CAT CTG CA (SEQ ID NO:258)
test fwd	GCG ACA CGG AAA TGT TGA ATA CTC AT (SEQ ID NO:259)
test rev	GGA GTC AGG CAA CTA TGG ATG AAC G (SEQ ID NO:260)
off-target CLTA4-0 fwd	ACT GTG AAG AGC TTC ACT GAG TAG GAT TAA GAT ATT GCA GAT GTA GTG TTT CCA CAG GGT (SEQ ID NO:261)
off-target CLTA4-1 fwd	ACT GTG AAG AGC TTC ACT GAG TAG GAT TAA GAT ATT GAA GAT GTA GTG TTT CCA CAG GGT (SEQ ID NO:262)

(continued)

	oligonucleotide name	oligonucleotide sequence (5'→3')
5	off-target CLTA4-2a fwd	ACT GTG AAG AGC TTC ACT GAG TAG GAT TAA GAT ATT GAA GAT GTA GTG TTT CCA CTG GGT (SEQ ID NO:263)
10	off-target CLTA4-2b fwd	ACT GTG AAG AGC TTC ACT GAG TAG GAT TAA GAT ATT GCA GAT GGA GGG TTT CCA CAG GGT (SEQ ID NO:264)
15	off-target CLTA4-2c fwd	ACT GTG AAG AGC TTC ACT GAG TAG GAT TAA GAT ATT GCA GAT GTA GTG TTA CCA GAG GGT (SEQ ID NO:265)
20	off-target CLTA4-3 fwd	ACT GTG AAG AGC TTC ACT GAG TAG GAT TAA GAT ATT GGG GAT GTA GTG TTT CCA CTG GGT (SEQ ID NO:266)
25	off-target CLTA4-0 rev	TCC CTC AAG CAG GCC CCG CTG GTG CAC TGA AGA GCC ACC CTG TGG AAA CAC TAC ATC TGC (SEQ ID NO:267)
30	off-target CLTA4-1 rev	TCC CTC AAG CAG GCC CCG CTG GTG CAC TGA AGA GCC ACC CTG TGG AAA CAC TAC ATC TTC (SEQ ID NO:268)
35	off-target CLTA4-2a rev	TCC CTC AAG CAG GCC CCG CTG GTG CAC TGA AGA GCC ACC CAG TGG AAA CAC TAC ATC TTC (SEQ ID NO:269)
40	off-target CLTA4-2b rev	TCC CTC AAG CAG GCC CCG CTG GTG CAC TGA AGA GCC ACC CTG TGG AAA CCC TCC ATC TGC (SEQ ID NO:270)
45	off-target CLTA4-2c rev	TCC CTC AAG CAG GCC CCG CTG GTG CAC TGA AGA GCC ACC CTC TGG TAA CAC TAC ATC TGC (SEQ ID NO:271)
50	off-target CLTA4-3 rev	TCC CTC AAG CAG GCC CCG CTG GTG CAC TGA AGA GCC ACC CAG TGG AAA CAC TAC ATC CCC (SEQ ID NO:272)
55	adapter1 (AACAA)	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC TAA CA (SEQ ID NO:273)
	adapter2(AACAA)	TGT TAG ATC GGA AGA GCG TCG TGT AGG GAA AGA GTG TAG ATC TCG GTG G (SEQ ID NO:274)
	adapter1 (TTCA)	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC TTT CA (SEQ ID NO:275)
	adapter2(TTCA)	TGA AAG ATC GGA AGA GCG TCG TGT AGG GAA AGA GTG TAG ATC TCG GTG G (SEQ ID NO:276)
	adapter1	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC T (SEQ ID NO:277)
	adapter2	AGA TCG GAA GAG CGT CGT GTA GGG AAA GAG TGT AGA TCT CGG TGG (SEQ ID NO:278)
	lib adapter1	GAC GGC ATA CGA GAT (SEQ ID NO:279)
	CLTA1 lib adapter2	AAC AAT CTC GTA TGC CGT CTT CTG CTT G (SEQ ID NO:280)
	CLTA2 lib adapter2	TCT TAT CTC GTA TGC CGT CTT CTG CTT G (SEQ ID NO:281)
	CLTA3 lib adapter2	AGA GAT CTC GTA TGC CGT CTT CTG CTT G (SEQ ID NO:282)
	CLTA4 lib adapter2	TTG TAT CTC GTA TGC CGT CTT CTG CTT G (SEQ ID NO:283)
	CLTA1 sel PCR	CAA GCA GAA GAC GGC ATA CGA GAT TGT GTT CTC GGC AGG T (SEQ ID NO:284)
	CLTA2 sel PCR	CAA GCA GAA GAC GGC ATA CGA GAT AGA AGA CTC GGC AGG T (SEQ ID NO:285)
	CLTA3 sel PCR	CAA GCA GAA GAC GGC ATA CGA GAT TTC TCT CTC GGC AGG T (SEQ ID NO:286)
	CLTA4 sel PCR	CAA GCA GAA GAC GGC ATA CGA GAT ACA CAA CTC GGC AGG T (SEQ ID NO:287)
	PE2 short	AAT GAT ACG GCG ACC ACC GA (SEQ ID NO:288)

(continued)

	oligonucleotide name	oligonucleotide sequence (5'→3')
5	CLTA1 lib seq PCR	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC TNN NNA CCT ACC TGC CGA GAA CAC A (SEQ ID NO:289)
10	CLTA2 lib seq PCR	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC TNN NNA CCT ACC TGC CGA GTC TTC T (SEQ ID NO:290)
15	CLTA3 lib seq PCR	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC TNN NNA CCT ACC TGC CGA GAG AGA A (SEQ ID NO:291)
20	CLTA4 lib seq PCR	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC TNN NNA CCT ACC TGC CGA GTT GTG T (SEQ ID NO:292)
25	lib fwd PCR	CAA GCA GAA GAC GGC ATA CGA GAT (SEQ ID NO:293)
30	CLTA1-0-1 (Chr. 9) fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT CAA GTC TAG CAA GCA GGC CA (SEQ ID NO:294)
35	CLTA1-0-1 (Chr. 12) fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT CAG GCA CTG AGT GGG AAA GT (SEQ ID NO:295)
40	CLTA1-1-1 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT TAA CCC CAA GTC AGC AAG CA (SEQ ID NO:296)
45	CLTA1-2-1 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT TTG CTG GTC AAT ACC CTG GC (SEQ ID NO:297)
50	CLTA1-2-2 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT TGA GTA CCC CTG AAA TGG GC (SEQ ID NO:298)
55	CLTA1-3-1 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT TCG CTA CCA ATC AGG GCT TT (SEQ ID NO:299)
60	CLTA1-3-2 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT CCA TTG CCA CTT GTT TGC AT (SEQ ID NO:300)
65	CLTA1-4-1 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT CCT ACC CCC ACA ACT TTG CT (SEQ ID NO:301)
70	CLTA1-4-2 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT GTG TAC ATC CAG TGC ACC CA (SEQ ID NO:302)
75	CLTA1-4-3 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT TCG GAA AGG ACT TTG AAT ACT TGT (SEQ ID NO:303)
80	CLTA1-4-4 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT CGG CCC AAG ACC TCA TTC AC (SEQ ID NO:304)
85	CLTA1-4-5 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT GTC CTC TCT GGG GCA GAA GT (SEQ ID NO:305)
90	CLTA1-4-6 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT AGC TGA GTC ATG AGT TGT CTC C (SEQ ID NO:306)
95	CLTA1-4-7 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT CTG CCA GCT TCT CAC ACC AT (SEQ ID NO:307)
100	CLTA1-4-8 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT CTG AAG GAC AAA GGC GGG AA (SEQ ID NO:308)
105	CLTA1-5-1 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT AAG GTG CTA AAG GCT CCA CG (SEQ ID NO:309)
110	CLTA1-5-2 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT GAC CAT TGG TGA GCC CAG AG (SEQ ID NO:310)

(continued)

	oligonucleotide name	oligonucleotide sequence (5'->3')
5	CLTA1-5-3 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT TTT TTC GGG CAA CTG CTC AC (SEQ ID NO:311)
10	CLTA1-5-4 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT GCA AGC CTT CTC TCC TCA GA (SEQ ID NO:312)
15	CLTA1-5-5 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT ACA CAA ACT TCC CTG AGA CCC (SEQ ID NO:313)
20	CLTA1-6-1 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT TGA GTT AGC CCT GCT GTT CA (SEQ ID NO:314)
25	CLTA4-0-1 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT TGA AGA GCT TCA CTG AGT AGG A (SEQ ID NO:315)
30	CLTA4-3-1 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT TCC CCT TAC AGC CAA TTT CGT (SEQ ID NO:316)
35	CLTA4-3-2 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT TGC TGA TGA AAT GCA ATT AAG AGG T (SEQ ID NO:317)
40	CLTA4-3-3 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT GGT CCC TGC AAG CCA GTA TG (SEQ ID NO:318)
45	CLTA4-3-4 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT ATC AAA GCC TTG TAT CAC AGT T (SEQ ID NO:319)
50	CLTA4-3-5 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT CCC AAA TAA TGC AGG AGC CAA (SEQ ID NO:320)
55	CLTA4-3-6 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT CTG CCT TTA GTG GGA CAG ACT T (SEQ ID NO:321)
	CLTA4-3-7 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT AGT AAC CCT AGT AGC CCT CCA (SEQ ID NO:322)
	CLTA4-4-1 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT CAT TGC AGT GAG CCG AGA TTG (SEQ ID NO:323)
	CLTA4-4-2 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT TGG CAA AGT TCA CTT CCA TGT (SEQ ID NO:324)
	CLTA4-4-3 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT TGC TCT GTG ATG TCT GCC AC (SEQ ID NO:325)
	CLTA4-4-4 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT TGT GTA GGA TTG TGA ACC AGC A (SEQ ID NO:326)
	CLTA4-4-5 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT TCC CAG CCC AGC ATT TTT CT (SEQ ID NO:327)
	CLTA4-4-6 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT AGG TTG CTT TGT GCA CAG TC (SEQ ID NO:328)
	CLTA4-4-7 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT CCT GGC TTG GGA TGT TGG AA (SEQ ID NO:329)
	CLTA4-4-8 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT TTG CCC AAG GTC ATA CTG CT (SEQ ID NO:330)
	CLTA4-4-9 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT ACC CAC TAG GTA GCC ATA ATC CA (SEQ ID NO:331)
	CLTA4-4-10 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT CGG TCA TGT CGC TTG GAA GA (SEQ ID NO:332)

(continued)

	oligonucleotide name	oligonucleotide sequence (5'->3')
5	CLTA4-4-11 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT TTG GCC CAT ATT GCT TTA TGC TG (SEQ ID NO:333)
10	CLTA4-4-12 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT ATT AGG GGT TGG CTG CAT GA (SEQ ID NO:334)
15	CLTA4-4-13 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT CCA AGA CGT GTT GCA TGC TG (SEQ ID NO:335)
20	CLTA4-4-14 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT TGG GAG GTG ATA AAT TCC CTA AAT (SEQ ID NO:336)
25	CLTA4-5-1 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT CCA GAG ACA AAG GTG GGG AG (SEQ ID NO:337)
30	CLTA4-5-2 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT TCA TAC AGA AGA GCA AAG TAC CA (SEQ ID NO:338)
35	CLTA4-5-3 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT CAA AGA GGG GTA TCG GGA GC (SEQ ID NO:339)
40	CLTA4-5-4 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT AAA TGG AAG AAC CAA GTA GAT GAA (SEQ ID NO:340)
45	CLTA4-5-5 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT TTT TGG TTG ACA GAT GGC CAC A (SEQ ID NO:341)
50	CLTA4-5-6 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT TCT TAC TTG TGT GAT TTT AGA ACA A (SEQ ID NO:342)
55	CLTA4-6-1 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT GAT GGT TCA TGC AGA GGG CT (SEQ ID NO:343)
	CLTA4-6-2 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT GCT GGT CTT TCC TGA GCT GT (SEQ ID NO:344)
	CLTA4-6-3 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT CTC CAT CAG ATA CCT GTA CCC A (SEQ ID NO:345)
	CLTA4-7-1 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT GGG AAA ACA CTC TCT CTC TGC T (SEQ ID NO:346)
	CLTA4-7-2 fwd	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT GGA GGC CAC GAC ACA CAA TA (SEQ ID NO:347)
	CLTA1-0-1 (Chr. 9) rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT CAC AGG GTG GCT CTT CAG TG (SEQ ID NO:348)
	CLTA1-0-1 (Chr. 12) rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT TGC ACA TGT TTC CAC AGG GT (SEQ ID NO:349)
	CLTA1-1-1 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT AGT GTT TCC AGG AGC GGT TT (SEQ ID NO:350)
	CLTA1-2-1 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT AAG CCT CAG GCA CAA CTC TG (SEQ ID NO:351)
	CLTA1-2-2 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT TAG GGG AGG GGC AAA GAC A (SEQ ID NO:352)
	CLTA1-3-1 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT GGG AAC AGT GGT ATG CTG GT (SEQ ID NO:353)
	CLTA1-3-2 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT AGT GTG GAC ACT GAC AAG GAA (SEQ ID NO:354)

(continued)

	oligonucleotide name	oligonucleotide sequence (5'->3')
5	CLTA1-4-1 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT TCA CTG CCT GGG TGC TTT AG (SEQ ID NO:355)
10	CLTA1-4-2 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT TAC CCC AGC CTC CAG CTT TA (SEQ ID NO:356)
15	CLTA1-4-3 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT TGA CTA CTG GGG AGC GAT GA (SEQ ID NO:357)
20	CLTA1-4-4 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT AGG CTG TTA TGC AGG AAA GGA A (SEQ ID NO:358)
25	CLTA1-4-5 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT GCG GTT GAG GTG GAT GGA AG (SEQ ID NO:359)
30	CLTA1-4-6 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT GGC AGC ATC CCT TAC ATC CT (SEQ ID NO:360)
35	CLTA1-4-7 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT AGA AAA AGC TTC CCC AGA AAG GA (SEQ ID NO:361)
40	CLTA1-4-8 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT CTG CAC CAA CCT CTA CGT CC (SEQ ID NO:362)
45	CLTA1-5-1 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT CTG GAG AGG GCA TAG TTG GC (SEQ ID NO:363)
50	CLTA1-5-2 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT TGG AAG GCT CTT TGT GGG TT (SEQ ID NO:364)
55	CLTA1-5-3 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT TTC CTA GCG GGA ACT GGA AA (SEQ ID NO:365)
	CLTA1-5-4 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT AGG CTA ATG GGG TAG GGG AT (SEQ ID NO:366)
	CLTA1-5-5 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT TGT CCA TGT TGG CTG AGG TG (SEQ ID NO:367)
	CLTA1-6-1 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT CAG GCC AAC CTT GAC AAC TT (SEQ ID NO:368)
	CLTA4-0-1 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT AGC AGG CCA AAG ATG TCT CC (SEQ ID NO:369)
	CLTA4-3-1 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT TCT GCT CTT GAG GTT ATT TGT CC (SEQ ID NO:370)
	CLTA4-3-2 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT GGG ACC AAT TTG CTA CTC ATG G (SEQ ID NO:371)
	CLTA4-3-3 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT TGG AGG CTG TAA ACG TCC TG (SEQ ID NO:372)
	CLTA4-3-4 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT TGC TAT GAT TTG CTG AAT TAC TCC T (SEQ ID NO:373)
	CLTA4-3-5 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT GCA ATT TTG CAG ACC ACC ATC (SEQ ID NO:374)
	CLTA4-3-6 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT GGC AGC TTG CAA CCT TCT TG (SEQ ID NO:375)
	CLTA4-3-7 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT TCA TGA GAG TTT CCC CAA CA (SEQ ID NO:376)

(continued)

	oligonucleotide name	oligonucleotide sequence (5'->3')
5	CLTA4-4-1 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT ACT TGA GGG GGA AAA AGT TTC TTA (SEQ ID NO:377)
10	CLTA4-4-2 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT TGG TCC CTG TCT GTC ATT GG (SEQ ID NO:378)
15	CLTA4-4-3 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT AAG CGA GTG ACT GTC TGG GA (SEQ ID NO:379)
20	CLTA4-4-4 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT CAT GGG TGG GAC ACG TAG TT (SEQ ID NO:380)
25	CLTA4-4-5 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT GGC TTT CCT GGA CAC CCT ATC (SEQ ID NO:381)
30	CLTA4-4-6 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT AGA GCG AGG GAG CGA TGT A (SEQ ID NO:382)
35	CLTA4-4-7 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT TTG TGG ACC ACT GCT TAG TGC (SEQ ID NO:383)
40	CLTA4-4-8 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT CAA CTA CCC TGA GGC CAC C (SEQ ID NO:384)
45	CLTA4-4-9 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT GGT CAG CAC TCC TCA GCT TT (SEQ ID NO:385)
50	CLTA4-4-10 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT TGG AGG ATG CAT GCC ACA TT (SEQ ID NO:386)
55	CLTA4-4-11 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT CCC AGC CTC TTT GAC CCT TC (SEQ ID NO:387)
	CLTA4-4-12 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT CCC ACA CCA GGC TGT AAG G (SEQ ID NO:388)
	CLTA4-4-13 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT TAG ATA TAT GGG TGT GTC TGT ACG (SEQ ID NO:389)
	CLTA4-4-14 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT TTC CAA AGT GGC TGA ACC AT (SEQ ID NO:390)
	CLTA4-5-1 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT CCC ACA GGG CTG ATG TTT CA (SEQ ID NO:391)
	CLTA4-5-2 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT TTG TAA TGC AAC CTC TGT CAT GC (SEQ ID NO:392)
	CLTA4-5-3 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT CCA GCT CCA GCA ATC CAT GA (SEQ ID NO:393)
	CLTA4-5-4 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT TTT GGG AAA GAT AGC CCT GGA (SEQ ID NO:394)
	CLTA4-5-5 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT CAA TGA AAC AGC GGG GAG GT (SEQ ID NO:395)
	CLTA4-5-6 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT ACA ATC ACG TGT CCT TCA CT (SEQ ID NO:396)
	CLTA4-6-1 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT CAG ATC CCT CCT GGG CAA TG (SEQ ID NO:397)
	CLTA4-6-2 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT GTC AGG AGG CAA GGA GGA AC (SEQ ID NO:398)

(continued)

	oligonucleotide name	oligonucleotide sequence (5'→3')
5	CLTA4-6-3 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT ACT TCC TTC CTT TTG AGA CCA AGT (SEQ ID NO:399)
10	CLTA4-7-1 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT GCG GCA GAT TCC TGG TGA TT (SEQ ID NO:400)
15	CLTA4-7-2 rev	GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCT GGT CAC CAT CAG CAC AGT CA (SEQ ID NO:401)
20	PE1-barcode1	CAA GCA GAA GAC GGC ATA CGA GAT ATA TCA GTG TGA CTG GAG TTC AGA CGT GTG CT (SEQ ID NO:402)
25	PE 1-barcode2	CAA GCA GAA GAC GGC ATA CGA GAT TTT CAC CGG TGA CTG GAG TTC AGA CGT GTG CT (SEQ ID NO:403)
30	PE1-barcode3	CAA GCA GAA GAC GGC ATA CGA GAT CCA CTC ATG TGA CTG GAG TTC AGA CGT GTG CT (SEQ ID NO:404)
35	PE1-barcode4	CAA GCA GAA GAC GGC ATA CGA GAT TAC GTA CGG TGA CTG GAG TTC AGA CGT GTG CT (SEQ ID NO:405)
40	PE1-barcode5	CAA GCA GAA GAC GGC ATA CGA GAT CGA AAC TCG TGA CTG GAG TTC AGA CGT GTG CT (SEQ ID NO:406)
45	PE 1-barcode6	CAA GCA GAA GAC GGC ATA CGA GAT ATC AGT ATG TGA CTG GAG TTC AGA CGT GTG CT (SEQ ID NO:407)
50	PE2-barcode1	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACA TTA CTC GAC ACT CTT TCC CTA CAC GAC (SEQ ID NO:408)
55	PE2-barcode2	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CCG GAG AAC ACT CTT TCC CTA CAC GAC (SEQ ID NO:409)
	PE2-barcode3	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACC GCT CAT TAC ACT CTT TCC CTA CAC GAC (SEQ ID NO:410)

REFERENCES**[0136]**

1. Hockemeyer, D. et al. Genetic engineering of human pluripotent cells using TALE nucleases. *Nature biotechnology* 29, 731-734 (2011).
2. Zou, J. et al. Gene targeting of a disease-related gene in human induced pluripotent stem and embryonic stem cells. *Cell stem cell* 5, 97-110 (2009).
3. Hockemeyer, D. et al. Efficient targeting of expressed and silent genes in human ESCs and iPSCs using zinc-finger nucleases. *Nature biotechnology* 27, 851-857 (2009).
4. Doyon, Y. et al. Heritable targeted gene disruption in zebrafish using designed zinc-finger nucleases. *Nature biotechnology* 26, 702-708 (2008).
5. Meng, X., Noyes, M.B., Zhu, L.J., Lawson, N.D. & Wolfe, S.A. Targeted gene inactivation in zebrafish using engineered zinc-finger nucleases. *Nature biotechnology* 26, 695-701 (2008).
6. Sander, J.D. et al. Targeted gene disruption in somatic zebrafish cells using engineered TALENs. *Nature biotechnology* 29, 697-698 (2011).
7. Tesson, L. et al. Knockout rats generated by embryo microinjection of TALENs. *Nature biotechnology* 29, 695-696 (2011).
8. Cui, X. et al. Targeted integration in rat and mouse embryos with zinc-finger nucleases. *Nature biotechnology* 29, 64-67 (2011).
9. Perez, E.E. et al. Establishment of HIV-1 resistance in CD4+ T cells by genome editing using zinc-finger nucleases. *Nature biotechnology* 26, 808-816 (2008).
10. NCT00842634, NCT01044654, NCT01252641, NCT01082926.
11. Jinek, M. et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*

- 337, 816-821 (2012).
 12. Cong, L. et al. Multiplex genome engineering using CRISPR/Cas systems. *Science* 339, 819-823 (2013).
 13. Mali, P. et al. RNA-guided human genome engineering via Cas9. *Science* 339, 823-826 (2013).
 14. Hwang, W.Y. et al. Efficient genome editing in zebrafish using a CRISPR-Cas system. *Nature biotechnology* 31, 227-229 (2013).
 5
 15. Jinek, M. et al. RNA-programmed genome editing in human cells. *eLife* 2, e00471 (2013).
 16. Dicarlo, J.E. et al. Genome engineering in *Saccharomyces cerevisiae* using CRISPR-Cas systems. *Nucleic acids research* (2013).
 10
 17. Jiang, W., Bikard, D., Cox, D., Zhang, F. & Marraffini, L.A. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nature biotechnology* 31, 233-239 (2013).
 18. Saprauskas, R. et al. The *Streptococcus thermophilus* CRISPR/Cas system provides immunity in *Escherichia coli*. *Nucleic acids research* 39, 9275-9282 (2011).
 15
 19. Semenova, E. et al. Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. *Proceedings of the National Academy of Sciences of the United States of America* 108, 10098-10103 (2011).
 20
 20. Qi, L.S. et al. Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression. *Cell* 152, 1173-1183 (2013).
 21. Pattanayak, V., Ramirez, C.L., Joung, J.K. & Liu, D.R. Revealing off-target cleavage specificities of zinc-finger nucleases by in vitro selection. *Nature methods* 8, 765-770 (2011).
 22. Doyon, J.B., Pattanayak, V., Meyer, C.B. & Liu, D.R. Directed evolution and substrate specificity profile of homing endonuclease I-SceI. *Journal of the American Chemical Society* 128, 2477-2484 (2006).
 23. Jiang, W., Bikard, D., Cox, D., Zhang, F. & Marraffini, L.A. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nature biotechnology* 31, 233-239 (2013).
 24. Pattanayak, V., Ramirez, C.L., Joung, J.K. & Liu, D.R. Revealing off-target cleavage specificities of zinc-finger nucleases by in vitro selection. *Nature methods* 8, 765-770 (2011).
 25
 25. Schneider, T.D. & Stephens, R.M. Sequence logos: a new way to display consensus sequences. *Nucleic acids research* 18, 6097-6100 (1990).

SEQUENCE LISTING

[0137]

<110> President and Fellows of Harvard College

35 <120> NUCLEASE PROFILING SYSTEM

<130> P103019EP1

<140>

40 <141> 2014-08-08

<150> US 61/864,289

<151> 2013-08-09

45 <150> US 14/320,370

<151> 2014-06-30

<150> US 14/320,413

50 <151> 2013-06-30

<160> 410

<170> PatentIn version 3.5

55 <210> 1

<211> 23

<212> DNA

<213> Artificial Sequence

5 <220>
 <223> Synthetic Polynucleotide
 <220>
 10 5 <221> misc_feature
 <222> (21)..(21)
 <223> n is a, c, g, or t
 <400> 1
 agtcctatc tccctcaagc ngg 23
 <210> 2
 <211> 48
 <212> DNA
 15 15 <213> Artificial Sequence
 <220>
 <223> Synthetic Oligonucleotide
 20 20 <400> 2
 atgtctcccg catgcgctca gtcctcatct ccctcaagca ggccccgc 48
 <210> 3
 <211> 48
 25 25 <212> DNA
 <213> Artificial Sequence
 <220>
 <223> Synthetic Oligonucleotide
 30 30 <400> 3
 gcggggcctg cttagggag atgaggactg agcgcatgcg ggagacat 48
 <210> 4
 <211> 48
 35 35 <212> DNA
 <213> Artificial Sequence
 <220>
 40 40 <223> Synthetic Oligonucleotide
 <400> 4
 tggtgactg aagagccacc ctgtggaaac actacatctg caatatct 48
 <210> 5
 <211> 48
 45 45 <212> DNA
 <213> Artificial Sequence
 <220>
 50 50 <223> Synthetic Oligonucleotide
 <400> 5
 agatattgca gatgttgttccacaggg tggctttca gtgcacca 48
 55 55 <210> 6
 <211> 48
 <212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

5

<400> 6

taatcctact cagtgaagtc cttcacagtc attggattaa ttatgttg 48

10 <210> 7

<211> 48

<212> DNA

<213> Artificial Sequence

15 <220>

<223> Synthetic Oligonucleotide

<400> 7

caacataatt aatccaatga ctgtgaagag cttcactgag taggatta 48

20 <210> 8

<211> 23

<212> DNA

<213> Artificial Sequence

25 <220>

<223> Synthetic Oligonucleotide

<400> 8

agtccctatc tccctcaagc agg 23

30

<210> 9

<211> 23

<212> DNA

<213> Artificial Sequence

35 <220>

<223> Synthetic Oligonucleotide

<400> 9

cctgcttgag ggagatgagg act 23

40 <210> 10

<211> 60

<212> RNA

45 <213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

50 <400> 10

gaguccucau cucccucaag cguuuuagag cuagaaaauag cuaaaaauaa ggcuaguccg 60

<210> 11

<211> 23

55 <212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

<400> 11
ctccctcaag caggccccgc tgg 23

5
<210> 12
<211> 23
<212> DNA
<213> Artificial Sequence

10
<220>
<223> Synthetic Oligonucleotide

15 <400> 12
ccagcggggc ctgcttgagg gag 23

<210> 13
<211> 60
<212> RNA
20 <213> Artificial Sequence

<220>
<223> Synthetic Oligonucleotide

25 <400> 13
gcucccucaa gcaggcccg cguuuuagag cuagaaauag cuaaaaauaa ggcuaguccg 60

<210> 14
<211> 23
30 <212> DNA
<213> Artificial Sequence

<220>
<223> Synthetic Oligonucleotide

35 <400> 14
tgtgaagagc ttcaactgagt agg 23

<210> 15
<211> 23
<212> DNA
40 <213> Artificial Sequence

<220>
<223> Synthetic Oligonucleotide

45 <400> 15
cctactcagt gaagctttc aca 23

<210> 16
<211> 60
<212> RNA
50 <213> Artificial Sequence

55 <220>
<223> Synthetic Oligonucleotide

<400> 16

gugugaagag cuucacugag uguuuuagag cuagaaauag cuaaaaauaa ggcuaguccg 60

5 <210> 17
 <211> 23
 <212> DNA
 <213> Artificial Sequence

10 <220>
 <223> Synthetic Oligonucleotide

15 <400> 17
 gcagatgtag tgttccaca ggg 23

20 <210> 18
 <211> 23
 <212> DNA
 <213> Artificial Sequence

25 <220>
 <223> Synthetic Oligonucleotide

30 <400> 18
 ccctgtggaa acactacatc tgc 23

35 <210> 19
 <211> 60
 <212> RNA
 <213> Artificial Sequence

40 <220>
 <223> Synthetic Oligonucleotide

45 <400> 19
 ggcagaugua guguuuccac aguuuuagag cuagaaauag cuaaaaauaa ggcuaguccg 60

50 <210> 20
 <211> 23
 <212> DNA
 <213> Artificial Sequence

55 <220>
 <223> Synthetic Oligonucleotide

60 <400> 21
 cctgcttgag ggagatgagg act 23

65 <210> 22

5 <211> 73
 <212> RNA
 <213> Artificial Sequence

 10 <220>
 <223> Synthetic Oligonucleotide

 15 <400> 22

 20 10 gaguccucau cucccuacaag cgaaaaaagg cuaugcugaa aagcauagcu aaaaauuaagg 60
 cuaguccguu auc

 25 15 <210> 23
 <211> 23
 <212> DNA
 <213> Artificial Sequence

 30 20 <220>
 <223> Synthetic Oligonucleotide

 35 25 <400> 23
 ctccctcaag caggccccgc tgg 23

 40 30 <210> 24
 <211> 23
 <212> DNA
 <213> Artificial Sequence

 45 35 <220>
 <223> Synthetic Oligonucleotide

 50 40 <400> 24
 ccagcggggc ctgtttgagg gag 23

 55 45 <210> 25
 <211> 73
 <212> RNA
 <213> Artificial Sequence

 <220>
 <223> Synthetic Oligonucleotide

 50 50 <400> 25

 55 55 10 gcucccucaa gcaggccccg cgaaaaaagg cuaugcugaa aagcauagcu aaaaauuaagg 60
 cuaguccguu auc 73

 60 60 <210> 26
 <211> 23
 <212> DNA
 <213> Artificial Sequence

 65 65 <220>
 <223> Synthetic Oligonucleotide

5 <400> 26
 tgtaagagc ttcaactgagt agg 23
 <210> 27
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 10 <220>
 <223> Synthetic Oligonucleotide
 <400> 27
 cctactcagt gaagctcttc aca 23
 15 <210> 28
 <211> 73
 <212> RNA
 <213> Artificial Sequence
 20 <220>
 <223> Synthetic Oligonucleotide
 <400> 28
 25 gugugaagag cuucacugag uguuuuagag cuaugcugaa aagcauagcu uaaaaauaagg 60
 cuaguccguu auc 73
 30 <210> 29
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 35 <220>
 <223> Synthetic Oligonucleotide
 <400> 29
 gcagatgtag tggccaca ggg 23
 40 <210> 30
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 45 <220>
 <223> Synthetic Oligonucleotide
 <400> 30
 ccctgtggaa acactacatc tgc 23
 <210> 31
 <211> 73
 <212> RNA
 <213> Artificial Sequence
 55 <220>
 <223> Synthetic Oligonucleotide

<400> 31

5	ggcagaugua guguuuccac aguuuuagag cuaugcugaa aagcauagcu uaaaaauaagg	60
	cuaguccguu auc	73

<210> 32

<211> 23

10 <212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

15 <220>

<221> misc_feature

<222> (21)..(21)

<223> n is a, c, g, or t

20

<400> 32

ctccctcaag caggccccgc ngg 23

<210> 33

25 <211> 23

<212> DNA

<213> Artificial Sequence

<220>

30 <223> Synthetic Oligonucleotide

<220>

<221> misc_feature

<222> (21)..(21)

35 <223> n is a, c, g, or t

<400> 33

gcagatgtag tgttccaca ngg 23

40 <210> 34

<211> 23

<212> DNA

<213> Artificial Sequence

45 <220>

<223> Synthetic Oligonucleotide

<400> 34

50 gcagatgtag tgttccaca ggg 23

55 <210> 35

<211> 23

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

1 <400> 35
 2 gaagatgtag tgttccaca ggg 23
 3
 4 <210> 36
 5 <211> 23
 6 <212> DNA
 7 <213> Artificial Sequence
 8
 9 <220>
 10 <223> Synthetic Oligonucleotide
 11
 12 <400> 36
 13 gaagatgtag tgttccact ggg 23
 14
 15 <210> 37
 16 <211> 23
 17 <212> DNA
 18 <213> Artificial Sequence
 19
 20 <220>
 21 <223> Synthetic Oligonucleotide
 22
 23 <400> 37
 24 gcagatggag ggttccaca ggg 23
 25
 26 <210> 38
 27 <211> 23
 28 <212> DNA
 29 <213> Artificial Sequence
 30
 31 <220>
 32 <223> Synthetic Oligonucleotide
 33
 34 <400> 38
 35 gcagatgtag tggaccaga ggg 23
 36
 37 <210> 39
 38 <211> 23
 39 <212> DNA
 40 <213> Artificial Sequence
 41
 42 <220>
 43 <223> Synthetic Oligonucleotide
 44
 45 <400> 39
 46 gggatgtag tgttccact ggg 23
 47
 48 <210> 40
 49 <211> 4104
 50 <212> DNA
 51 <213> Artificial Sequence
 52
 53 <220>
 54 <223> Synthetic Oligonucleotide
 55
 56 <400> 40

EP 3 486 318 B1

	atggataaga aataactcaat aggcttagat atcggcacaa atagcgtcgg atggcggtg	60
5	atcactgatg attataagg tccgtctaaa aagttcaagg ttctggaaa tacagaccgc	120
	cacagtatca aaaaaaatct tatagggct ctttatttgc gcagtggaga gacagcgaa	180
	gcgactcgtc tcaaacggac agctcgtaga aggtatacac gtcggaagaa tcgtatttgc	240
10	tatctacagg agatTTTC aaatgagatg gcgaaagtag atgatagttt ctttcatcga	300
	cttgaagagt ctttttggt ggaagaagac aagaagcatg aacgtcatcc tattttgga	360
	aatatagtag atgaagttgc ttatcatgag aaatatccaa ctatctatca tctgcgaaaa	420
15	aaattggcag attctactga taaagcggat ttgcgtttaa tctatttgc cttagcgcatt	480
	atgattaagt ttctgttca tttttgatt gagggagatt taaatcctga taatagtgt	540
	gtggacaaac tatttatcca gttggtacaa atctacaatc aattatttga agaaaaccct	600
20	attaacgcaa gtagagtaga tgctaaagcg attcttctg cacgatttag taaatcaaga	660
	cgattagaaa atctcattgc tcagctcccc ggtgagaaga gaaatggctt gtttggaaat	720
25	ctcattgtt tgcattggg attgaccctt aattttaaat caaattttga tttggcagaa	780
	gatgctaaat tacagcttc aaaagatact tacgtatgtt atttagataa tttatttgcg	840
	caaattggag atcaatatgc tgatttgtt ttggcagcta agaatttatac agatgctatt	900
30	ttactttcag atatcctaag agtaaatagt gaaataacta aggctcccattt atcagctca	960
	atgattaagc gctacgatga acatcatcaa gacttgactc tttaaaagc ttttagttcga	1020
	caacaacttc cagaaaagta taaagaaatc tttttgatc aatcaaaaaa cgatatgc	1080
35	ggttatatttgc atgggggagc tagccaagaa gaattttata aatttatcaa accaatttt	1140
	gaaaaaatgg atggtaactga ggaattatttgc gtgaaactaa atcgtgaaga tttgctgcgc	1200

40

45

50

55

EP 3 486 318 B1

	aagcaacgga ccttgacaa cggtcttatt cccatcaa ttcacttggg tgagctgcatt	1260
5	gctatttga gaagacaaga agactttat ccattttaa aagacaatcg tgagaagatt	1320
	gaaaaaatct tgactttcg aattccttat tatgttggc cattggcgcg tggcaatagt	1380
	cgtttgcatt ggatgactcg gaagtctgaa gaaacaatta cccatggaa ttttgaagaa	1440
10	gttgcata aaggtgcttc agctaatca tttattgaac gcatgacaaa ctgtataaa	1500
	aatcttccaa atgaaaaagt actaccaaaa catagttgc tttatgagta ttttacggtt	1560
	tataacgaat tgacaaaggt caaatatgtt actgaggaa tgcgaaaacc agcatttctt	1620
15	tcaggtgaac agaagaaagc cattgttgc ttactcttca aaacaaatcg aaaagtaacc	1680
	gttaagcaat taaaagaaga ttatttcaaa aaaatagaat gtttgcata tggtgaaatt	1740
	tcaggagttg aagatagatt taatgcttca ttaggcgcct accatgattt gctaaaaatt	1800
20	attaaagata aagattttt ggataatgaa gaaaatgaag atatcttgc ggatattgtt	1860
	ttaacattga ccttatttga agataggggg atgattgagg aaagacttaa aacatatgct	1920
	caccttttgc atgataaggt gatgaaacag cttaaacgtc gccgttatac tggggggaa	1980
25	cgttgttgc gaaaattgtat taatggattt agggataagc aatctggcaa aacaatatttt	2040
	gatttttgc aatcagatgg ttttgcattt cgcatttttgc tgcagctgtat ccatgtatgt	2100
	agtttgcacat taaaagaaga tattcaaaaa gcacaggtgt ctggacaagg ccatagtttt	2160
30	catgaacaga ttgctaactt agctggcagt cctgcttattt aaaaaggtat ttacagact	2220
	gtaaaaatttgc ttgatgaact ggtcaagta atggggcata agccagaaaa tatcgatttt	2280
	gaaatggcac gtgaaaatca gacaactcaa aaggcccaga aaaattcgcg agagcgtatg	2340
35	aaacgaatcg aagaaggat caaagaatta ggaagtcaga ttcttaaaga gcatcctgtt	2400
	gaaaatactc aattgcacaa tgaaaagctc tatctctattt atctacaaaa tggaaagagac	2460
	atgtatgtgg accaagaattt agatattttt cgtttaaatgtt attatgtatgtt cgatcacatt	2520
40	gttccacaaa gtttcattttt agacgattca atagacaata aggtactaac gcgttctgtat	2580
	aaaaatcgatg gtaaatcgga taacgttcca agtgaagaag tagtcaaaaa gatgaaaaac	2640
	tattggagac aacttctaaa cgccaaatgtt atcactcaac gtaagttgc taatttacgc	2700
45	aaagctgaac gtggagggtt gagtgaactt gataaagctg gttttatcaa acgccaatttgc	2760
	gttggaaactc gccaaatcac taagcatgtt gcacaaattt tggatagtcg catgaataact	2820
	aaatacgatg aaaatgataa acttatttgcg gaggttaaag tgattacattt aaaaatctaaa	2880
50	ttatgttctgc acttccgaaa agatttccaa ttctataaag tacgtgagat taacaattttt	2940
	catcatgccc atgatgcgtt tctaaatgcc gtcgttggaa ctgcatttgc taaatgtat	3000
55	ccaaaaacttgc aatcgaggatt tgtctatgtt gattataaag tttatgtatgt tcgtaaaatgc	3060
	attgcttaatgtt ctgagcaaga aataggcaaa gcaaccgcacaa aatatttctt ttactcttaat	3120

	atcatgaact tcttcaaaac agaaattaca cttgcaaatg gagagattcg caaacgcct	3180
	ctaatcgaaa ctaatgggaa aactggagaa attgtctggg ataaaggcg agatttgcc	3240
5	acagtgcgca aagtattgtc catgccccaa gtcaatattt tcaagaaaac agaagtacag	3300
	acaggcggat tctccaagga gtcaattttt ccaaaaagaa attcggacaa gcttattgct	3360
10	cgtaaaaaaag actgggatcc aaaaaaatat ggtggtttg atagtccac ggttagcttat	3420
	tcagtcctag tggttgctaa ggtggaaaaa gggaaatcga agaagttaaa atccgttaaa	3480
	gagttactag ggatcacaat tatggaaaga agttcctttg aaaaaaatcc gattgactt	3540
15	ttagaagcta aaggatataa ggaagttaaa aaagacttaa tcattaaact acctaaatat	3600
	agtcttttg agttagaaaaa cggtcgtaaa cggatgctgg ctatgtccgg agaattacaa	3660
	aaaggaaatg agctggctct gccaaagcaaa tatgtgaatt ttttatattt agctagtcatt	3720
20	tatgaaaagt tgaagggtag tccagaagat aacgaacaaa aacaattgtt tgtggagcag	3780
	cataaggcatt atttagatga gattattgag caaatcagtg aattttctaa gcgtgttatt	3840
25	ttagcagatg ccaattttaga taaagttctt agtgcataata acaaacatag agacaaacca	3900
	atacgtgaac aagcagaaaaa tattattcat ttattnacgt tgacgaatct tggagctccc	3960
	gctgctttta aatattttga tacaacaatt gatcgtaaac gatatacgtc tacaaaagaa	4020
30	gttttagatg ccactcttccatcaatcc atcaactggtc tttatgaaac acgcattgtat	4080
	ttgagtcagc taggaggtga ctga	4104

<210> 41

<211> 1367

<212> PRT

<213> Artificial Sequence

<220>

<223> Synthetic Polypeptide

<400> 41

45

50

55

EP 3 486 318 B1

Met Asp Lys Lys Tyr Ser Ile Gly Leu Asp Ile Gly Thr Asn Ser Val
1 5 10 15

5 Gly Trp Ala Val Ile Thr Asp Asp Tyr Lys Val Pro Ser Lys Lys Phe
20 25 30

10 Lys Val Leu Gly Asn Thr Asp Arg His Ser Ile Lys Lys Asn Leu Ile
35 40 45

15 Gly Ala Leu Leu Phe Gly Ser Gly Glu Thr Ala Glu Ala Thr Arg Leu
50 55 60

Lys Arg Thr Ala Arg Arg Tyr Thr Arg Arg Lys Asn Arg Ile Cys

20

25

30

35

40

45

50

55

EP 3 486 318 B1

	65	70	75	80
5	Tyr Leu Gln Glu Ile Phe Ser Asn Glu Met Ala Lys Val Asp Asp Ser 85		90	95
10	Phe Phe His Arg Leu Glu Glu Ser Phe Leu Val Glu Glu Asp Lys Lys 100	105		110
15	His Glu Arg His Pro Ile Phe Gly Asn Ile Val Asp Glu Val Ala Tyr 115	120	125	
20	His Glu Lys Tyr Pro Thr Ile Tyr His Leu Arg Lys Lys Leu Ala Asp 130	135	140	
25	Ser Thr Asp Lys Ala Asp Leu Arg Leu Ile Tyr Leu Ala Leu Ala His 145	150	155	160
30	Met Ile Lys Phe Arg Gly His Phe Leu Ile Glu Gly Asp Leu Asn Pro 165	170		175
35	Asp Asn Ser Asp Val Asp Lys Leu Phe Ile Gln Leu Val Gln Ile Tyr 180	185	190	
40	Asn Gln Leu Phe Glu Glu Asn Pro Ile Asn Ala Ser Arg Val Asp Ala 195	200	205	
45	Lys Ala Ile Leu Ser Ala Arg Leu Ser Lys Ser Arg Arg Leu Glu Asn 210	215	220	
50	Leu Ile Ala Gln Leu Pro Gly Glu Lys Arg Asn Gly Leu Phe Gly Asn 225	230	235	240
55	Leu Ile Ala Leu Ser Leu Gly Leu Thr Pro Asn Phe Lys Ser Asn Phe 245	250	255	
	Asp Leu Ala Glu Asp Ala Lys Leu Gln Leu Ser Lys Asp Thr Tyr Asp 260	265	270	
	Asp Asp Leu Asp Asn Leu Leu Ala Gln Ile Gly Asp Gln Tyr Ala Asp 275	280	285	
	Leu Phe Leu Ala Ala Lys Asn Leu Ser Asp Ala Ile Leu Leu Ser Asp 290	295	300	
	Ile Leu Arg Val Asn Ser Glu Ile Thr Lys Ala Pro Leu Ser Ala Ser 305	310	315	320

EP 3 486 318 B1

Met Ile Lys Arg Tyr Asp Glu His His Gln Asp Leu Thr Leu Leu Lys
325 330 335

5 Ala Leu Val Arg Gln Gln Leu Pro Glu Lys Tyr Lys Glu Ile Phe Phe
340 345 350

Asp Gln Ser Lys Asn Gly Tyr Ala Gly Tyr Ile Asp Gly Gly Ala Ser
355 360 365

10 Gln Glu Glu Phe Tyr Lys Phe Ile Lys Pro Ile Leu Glu Lys Met Asp
370 375 380

15 Gly Thr Glu Glu Leu Leu Val Lys Leu Asn Arg Glu Asp Leu Leu Arg
385 390 395 400

20 Lys Gln Arg Thr Phe Asp Asn Gly Ser Ile Pro His Gln Ile His Leu
405 410 415

Gly Glu Leu His Ala Ile Leu Arg Arg Gln Glu Asp Phe Tyr Pro Phe
420 425 430

25 Leu Lys Asp Asn Arg Glu Lys Ile Glu Lys Ile Leu Thr Phe Arg Ile
435 440 445

30 Pro Tyr Tyr Val Gly Pro Leu Ala Arg Gly Asn Ser Arg Phe Ala Trp
450 455 460

35 Met Thr Arg Lys Ser Glu Glu Thr Ile Thr Pro Trp Asn Phe Glu Glu
465 470 475 480

Val Val Asp Lys Gly Ala Ser Ala Gln Ser Phe Ile Glu Arg Met Thr
485 490 495

40 Asn Phe Asp Lys Asn Leu Pro Asn Glu Lys Val Leu Pro Lys His Ser
500 505 510

45 Leu Leu Tyr Glu Tyr Phe Thr Val Tyr Asn Glu Leu Thr Lys Val Lys
515 520 525

Tyr Val Thr Glu Gly Met Arg Lys Pro Ala Phe Leu Ser Gly Glu Gln
530 535 540

50 Lys Lys Ala Ile Val Asp Leu Leu Phe Lys Thr Asn Arg Lys Val Thr
545 550 555 560

55 Val Lys Gln Leu Lys Glu Asp Tyr Phe Lys Lys Ile Glu Cys Phe Asp
565 570 575

EP 3 486 318 B1

Ser Val Glu Ile Ser Gly Val Glu Asp Arg Phe Asn Ala Ser Leu Gly
580 585 590

5 Ala Tyr His Asp Leu Leu Lys Ile Ile Lys Asp Lys Asp Phe Leu Asp
595 600 605

10 Asn Glu Glu Asn Glu Asp Ile Leu Glu Asp Ile Val Leu Thr Leu Thr
610 615 620

15 Leu Phe Glu Asp Arg Gly Met Ile Glu Glu Arg Leu Lys Thr Tyr Ala
625 630 635 640

20 His Leu Phe Asp Asp Lys Val Met Lys Gln Leu Lys Arg Arg Arg Tyr
645 650 655

25 Thr Gly Trp Gly Arg Leu Ser Arg Lys Leu Ile Asn Gly Ile Arg Asp
660 665 670

30 Lys Gln Ser Gly Lys Thr Ile Leu Asp Phe Leu Lys Ser Asp Gly Phe
675 680 685

35 Ala Asn Arg Asn Phe Met Gln Leu Ile His Asp Asp Ser Leu Thr Phe
690 695 700

40 Lys Glu Asp Ile Gln Lys Ala Gln Val Ser Gly Gln Gly His Ser Leu
705 710 715 720

45 His Glu Gln Ile Ala Asn Leu Ala Gly Ser Pro Ala Ile Lys Lys Gly
725 730 735

Ile Leu Gln Thr Val Lys Ile Val Asp Glu Leu Val Lys Val Met Gly
740 745 750

50 His Lys Pro Glu Asn Ile Val Ile Glu Met Ala Arg Glu Asn Gln Thr
755 760 765

55 Thr Gln Lys Gly Gln Lys Asn Ser Arg Glu Arg Met Lys Arg Ile Glu
770 775 780

Glu Gly Ile Lys Glu Leu Gly Ser Gln Ile Leu Lys Glu His Pro Val
785 790 795 800

Glu Asn Thr Gln Leu Gln Asn Glu Lys Leu Tyr Leu Tyr Tyr Leu Gln
805 810 815

Asn Gly Arg Asp Met Tyr Val Asp Gln Glu Leu Asp Ile Asn Arg Leu
820 825 830

EP 3 486 318 B1

Ser Asp Tyr Asp Val Asp His Ile Val Pro Gln Ser Phe Ile Lys Asp
835 840 845

5 Asp Ser Ile Asp Asn Lys Val Leu Thr Arg Ser Asp Lys Asn Arg Gly
850 855 860

10 Lys Ser Asp Asn Val Pro Ser Glu Glu Val Val Lys Lys Met Lys Asn
865 870 875 880

15 Tyr Trp Arg Gln Leu Leu Asn Ala Lys Leu Ile Thr Gln Arg Lys Phe
885 890 895

Asp Asn Leu Thr Lys Ala Glu Arg Gly Gly Leu Ser Glu Leu Asp Lys
900 905 910

20 Ala Gly Phe Ile Lys Arg Gln Leu Val Glu Thr Arg Gln Ile Thr Lys
915 920 925

25 His Val Ala Gln Ile Leu Asp Ser Arg Met Asn Thr Lys Tyr Asp Glu
930 935 940

Asn Asp Lys Leu Ile Arg Glu Val Lys Val Ile Thr Leu Lys Ser Lys
945 950 955 960

30 Leu Val Ser Asp Phe Arg Lys Asp Phe Gln Phe Tyr Lys Val Arg Glu
965 970 975

35 Ile Asn Asn Tyr His His Ala His Asp Ala Tyr Leu Asn Ala Val Val
980 985 990

40 Gly Thr Ala Leu Ile Lys Lys Tyr Pro Lys Leu Glu Ser Glu Phe Val
995 1000 1005

Tyr Gly Asp Tyr Lys Val Tyr Asp Val Arg Lys Met Ile Ala Lys
1010 1015 1020

45 Ser Glu Gln Glu Ile Gly Lys Ala Thr Ala Lys Tyr Phe Phe Tyr
1025 1030 1035

50 Ser Asn Ile Met Asn Phe Phe Lys Thr Glu Ile Thr Leu Ala Asn
1040 1045 1050

Gly Glu Ile Arg Lys Arg Pro Leu Ile Glu Thr Asn Gly Glu Thr
1055 1060 1065

55 Gly Glu Ile Val Trp Asp Lys Gly Arg Asp Phe Ala Thr Val Arg

EP 3 486 318 B1

1070

1075

1080

5 Lys Val Leu Ser Met Pro Gln Val Asn Ile Val Lys Lys Thr Glu
 1085 1090 1095

Val Gln Thr Gly Gly Phe Ser Lys Glu Ser Ile Leu Pro Lys Arg
 1100 1105 1110

10

Asn Ser Asp Lys Leu Ile Ala Arg Lys Lys Asp Trp Asp Pro Lys
 1115 1120 1125

15

Lys Tyr Gly Gly Phe Asp Ser Pro Thr Val Ala Tyr Ser Val Leu
 1130 1135 1140

20

Val Val Ala Lys Val Glu Lys Gly Lys Ser Lys Lys Leu Lys Ser
 1145 1150 1155

Val Lys Glu Leu Leu Gly Ile Thr Ile Met Glu Arg Ser Ser Phe
 1160 1165 1170

25

Glu Lys Asn Pro Ile Asp Phe Leu Glu Ala Lys Gly Tyr Lys Glu
 1175 1180 1185

30

Val Lys Lys Asp Leu Ile Ile Lys Leu Pro Lys Tyr Ser Leu Phe
 1190 1195 1200

35

Glu Leu Glu Asn Gly Arg Lys Arg Met Leu Ala Ser Ala Gly Glu
 1205 1210 1215

40

Leu Gln Lys Gly Asn Glu Leu Ala Leu Pro Ser Lys Tyr Val Asn
 1220 1225 1230

Phe Leu Tyr Leu Ala Ser His Tyr Glu Lys Leu Lys Gly Ser Pro
 1235 1240 1245

50

Glu Asp Asn Glu Gln Lys Gln Leu Phe Val Glu Gln His Lys His
 1250 1255 1260

Tyr Leu Asp Glu Ile Ile Glu Gln Ile Ser Glu Phe Ser Lys Arg
 1265 1270 1275

Val Ile Leu Ala Asp Ala Asn Leu Asp Lys Val Leu Ser Ala Tyr
 1280 1285 1290

55

Asn Lys His Arg Asp Lys Pro Ile Arg Glu Gln Ala Glu Asn Ile
 1295 1300 1305

EP 3 486 318 B1

Ile His Leu Phe Thr Leu Thr Asn Leu Gly Ala Pro Ala Ala Phe
1310 1315 1320

5 Lys Tyr Phe Asp Thr Thr Ile Asp Arg Lys Arg Tyr Thr Ser Thr
1325 1330 1335

10 Lys Glu Val Leu Asp Ala Thr Leu Ile His Gln Ser Ile Thr Gly
1340 1345 1350

15 Leu Tyr Glu Thr Arg Ile Asp Leu Ser Gln Leu Gly Gly Asp
1355 1360 1365

<210> 42

<211> 100

<212> DNA

<213> Artificial Sequence

20 <220>

<223> Synthetic Oligonucleotide

25 <400> 42

aacacatggg tcgacacaaa cacaactcggt caggtacttg cagatgttagt ctttccacat 60

gggtcgacac aaacacaact cggcaggat ctcgtatgcc 100

30 <210> 43

<211> 10

<212> DNA

<213> Artificial Sequence

35 <220>

<223> Synthetic Oligonucleotide

40 <400> 43

ctcggcaggat 10

<210> 44

<211> 43

<212> DNA

<213> Artificial Sequence

45 <220>

<223> Synthetic Oligonucleotide

50 <400> 44

acttgcagat gtagtctttc cacatgggtc gacacaaaca caa 43

<210> 45

<211> 12

<212> DNA

<213> Artificial Sequence

55 <220>

<223> Synthetic Oligonucleotide

<400> 45

tgtgttgtg tt 12

5

<210> 46

<211> 12

<212> DNA

<213> Artificial Sequence

10

<220>

<223> Synthetic Oligonucleotide

<400> 46

15

agaagaagaa ga 12

<210> 47

<211> 12

<212> DNA

20

<213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

25

<400> 47

ttctcttct ct 12

<210> 48

<211> 12

30

<212> DNA

<213> Artificial Sequence

35

<220>

<223> Synthetic Oligonucleotide

<400> 48

acacaaacac aa

12

40

<210> 49

<211> 43

<212> DNA

<213> Artificial Sequence

45

<220>

<223> Synthetic Oligonucleotide

<400> 49

acttcagat gtagtcattc cacatgggtc gacacaaaca caa

43

50

<210> 50

<211> 31

<212> DNA

<213> Artificial Sequence

55

<220>

<223> Synthetic Oligonucleotide

<400> 50

acttgcagat gtagtcgttcc acatgggtc g 31
 <210> 51
 <211> 100
 5 <212> DNA
 <213> Artificial Sequence

 <220>
 <223> Synthetic Oligonucleotide
 10 <400> 51

 aacacatggg tcgacacaaa cacaactcg caggtacttg cagatgtagt cttccacat 60
 15 gggtcgacac aaacacaact cgccaggtat ctcttatgcc 100

 <210> 52
 <211> 162
 20 <212> DNA
 <213> Artificial Sequence

 <220>
 <223> Synthetic Oligonucleotide
 25 <400> 52

 caatctccccg catgcgctca gtcctcatct ccctcaagca ggcccccgtg gtgcactgaa 60
 30 gagccaccct gtgaaacact acatctgaa tatcttaatc ctactcgtg aagctttca 120
 cagtcattgg attaattatg ttgagttctt ttggaccaaaa cc 162

 <210> 53
 35 <211> 20
 <212> DNA
 <213> Artificial Sequence

 <220>
 40 <223> Synthetic Oligonucleotide

 <400> 53
 gctggtgac tgaagagcca 20

 45 <210> 54
 <211> 20
 <212> DNA
 <213> Artificial Sequence

 50 <220>
 <223> Synthetic Oligonucleotide

 <400> 54
 aatatctaa tcctactcag 20
 55 <210> 55
 <211> 22
 <212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

5

<400> 55

ccctgtaaaa cactacatct gc 22

10 <210> 56

<211> 22

<212> DNA

<213> Artificial Sequence

15 <220>

<223> Synthetic Oligonucleotide

<400> 56

gcagatgtag tgttcacag gg 22

20 <210> 57

<211> 23

<212> DNA

<213> Artificial Sequence

25 <220>

<223> Synthetic Oligonucleotide

<400> 57

gcagatgtag tgttccaca ggg 23

30

<210> 58

<211> 23

<212> DNA

<213> Artificial Sequence

35 <220>

<223> Synthetic Oligonucleotide

<400> 58

40 gcagatgtag tgttccaca ggg 23

<210> 59

<211> 23

<212> DNA

45 <213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

50 <400> 59

acatatgtag tatttccaca ggg 23

<210> 60

<211> 23

55 <212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

5 <400> 60
gcatatgtag tggccaaa tgt 23

<210> 61
<211> 23
<212> DNA
<213> Artificial Sequence

10 <220>
<223> Synthetic Oligonucleotide

15 <400> 61
ccagatgtag tattccaca ggg 23

<210> 62
<211> 23
<212> DNA
20 <213> Artificial Sequence

<220>
<223> Synthetic Oligonucleotide

25 <400> 62
gcagtttag tgtttcaca ggg 23

<210> 63
<211> 23
30 <212> DNA
<213> Artificial Sequence

<220>
<223> Synthetic Oligonucleotide

35 <400> 63
gcagatgtag tgttccaca cag 23

<210> 64
<211> 23
<212> DNA
40 <213> Artificial Sequence

<220>
<223> Synthetic Oligonucleotide

45 <400> 64
gcagatggag ggtttcaca ggg 23

<210> 65
<211> 23
<212> DNA
50 <213> Artificial Sequence

55 <220>
<223> Synthetic Oligonucleotide

<400> 65

ggaaattag tgttccaca ggg 23
 <210> 66
 <211> 23
 5 <212> DNA
 <213> Artificial Sequence

 <220>
 <223> Synthetic Oligonucleotide
 10 <400> 66
 aaagaatgtatccacata tgg 23

 <210> 67
 <211> 23
 15 <212> DNA
 <213> Artificial Sequence

 <220>
 20 <223> Synthetic Oligonucleotide

 <400> 67
 aaagatgtatccacata agg 23

 25 <210> 68
 <211> 23
 <212> DNA
 <213> Artificial Sequence

 30 <220>
 <223> Synthetic Oligonucleotide

 <400> 68
 aaatatgtatccacata ggg 23
 35 <210> 69
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 40 <220>
 <223> Synthetic Oligonucleotide

 <400> 69
 45 atagatgtatccacaaa gga 23

 <210> 70
 <211> 23
 <212> DNA
 50 <213> Artificial Sequence

 <220>
 <223> Synthetic Oligonucleotide

 <400> 70
 55 ccagaggtat tgctccacata ggg 23

 <210> 71

<211> 23
<212> DNA
<213> Artificial Sequence

5 <220>
 <223> Synthetic Oligonucleotide

 <400> 71
ccagatgtga ggttccaca agg 23

10 <210> 72
<211> 23
<212> DNA
<213> Artificial Sequence

15 <220>
 <223> Synthetic Oligonucleotide

 <400> 72
ctacatgtag tgttccata tgg 23

20 <210> 73
<211> 22
<212> DNA
25 <213> Artificial Sequence

 <220>
 <223> Synthetic Oligonucleotide

30 <400> 73
tagatgaagt gcttccacat gg 22

 <210> 74
<211> 23

35 <212> DNA
<213> Artificial Sequence

 <220>
 <223> Synthetic Oligonucleotide

40 <400> 74
gaaaatggag tgttcacaca tgg 23

 <210> 75
<211> 23

45 <212> DNA
<213> Artificial Sequence

 <220>
 <223> Synthetic Oligonucleotide

50 <400> 75
gcaaatgaag tgtcaccaca agg 23

 <210> 76
<211> 23

55 <212> DNA
<213> Artificial Sequence

<220>
<223> Synthetic Oligonucleotide
5 <400> 76
 gcaaatgtat tatttccact agg 23

 <210> 77
 <211> 23
 <212> DNA
10 <213> Artificial Sequence

 <220>
 <223> Synthetic Oligonucleotide
15 <400> 77
 gcagatgtag ctttgtaca tgg 23

 <210> 78
 <211> 23
20 <212> DNA
 <213> Artificial Sequence

 <220>
 <223> Synthetic Oligonucleotide
25 <400> 78
 gcagcttaag tgtttcaca tgg 23

 <210> 79
 <211> 23
 <212> DNA
30 <213> Artificial Sequence

 <220>
 <223> Synthetic Oligonucleotide
35 <400> 79
 ttacatgtag tgtttacaca cg 23

 <210> 80
 <211> 23
 <212> DNA
40 <213> Artificial Sequence

45 <220>
 <223> Synthetic Oligonucleotide

 <400> 80
 gaagaggaag tgttgcccc ggg 23
50 <210> 81
 <211> 23
 <212> DNA
 <213> Artificial Sequence

55 <220>
 <223> Synthetic Oligonucleotide

1 <400> 81
 gaagatgtgg agttgacaca tgg 23
 5 <210> 82
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 10 <220>
 <223> Synthetic Oligonucleotide
 <400> 82
 gcagaaggac tggtttaca agg 23
 15 <210> 83
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 20 <220>
 <223> Synthetic Oligonucleotide
 <400> 83
 gcagatgtgg aattacaaca ggg 23
 25 <210> 84
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 30 <220>
 <223> Synthetic Oligonucleotide
 <400> 84
 35 gcagtcatag tgtatacaca tgg 23
 <210> 85
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 40 <220>
 <223> Synthetic Oligonucleotide
 <400> 85
 taagatgttag tatttccaaa agt 23
 45 <210> 86
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 50 <220>
 <223> Synthetic Oligonucleotide
 <400> 86
 55 gcagctggca ttctccaca cg 23

5 <210> 87
 <211> 23
 <212> DNA
 <213> Artificial Sequence

10 <220>
 <223> Synthetic Oligonucleotide

 <400> 87
ggagatctga tggttctaca agg 23

15 <210> 88
 <211> 23
 <212> DNA
 <213> Artificial Sequence

 <220>
 <223> Synthetic Oligonucleotide

20 <400> 88
taaaatgcag tgtatccata tgg 23

 <210> 89
 <211> 23
25 <212> DNA
 <213> Artificial Sequence

 <220>
 <223> Synthetic Oligonucleotide

30 <400> 89
gccagaatag ttttcaaca agg 23

 <210> 90
 <211> 23
35 <212> DNA
 <213> Artificial Sequence

 <220>
 <223> Synthetic Oligonucleotide

40 <400> 90
ttgtattnag agattgcaca agg 23

 <210> 91
 <211> 23
45 <212> DNA
 <213> Artificial Sequence

 <220>
 <223> Synthetic Oligonucleotide

 <400> 91
agtccatc tccctcaagg agg 23
55 <210> 92
 <211> 23
 <212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

5

<400> 92

agtccctcaac tccctcaagc agg 23

10 <210> 93

<211> 23

<212> DNA

<213> Artificial Sequence

15 <220>

<223> Synthetic Oligonucleotide

<400> 93

agccctcatt tccctcaagc agg 23

20 <210> 94

<211> 23

<212> DNA

<213> Artificial Sequence

25 <220>

<223> Synthetic Oligonucleotide

<400> 94

actcctcatc cccctcaagc cg 23

30

<210> 95

<211> 23

<212> DNA

<213> Artificial Sequence

35 <220>

<223> Synthetic Oligonucleotide

<400> 95

agtcatcatc tccctcaagc aga 23

40 <210> 96

<211> 23

<212> DNA

45 <213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

50 <400> 96

cgtcctcctc tcccccaagc agg 23

<210> 97

<211> 23

55 <212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

<400> 97

tgtcccttc tccctcaagc aga 23

5

<210> 98

<211> 23

<212> DNA

<213> Artificial Sequence

10

<220>

<223> Synthetic Oligonucleotide

<400> 98

15

aagtttcatc tctctcaaggc tgg 23

<210> 99

<211> 23

<212> DNA

20

<213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

25

<400> 99

agtactcttt tccctcaggc tgg 23

<210> 100

<211> 23

30

<212> DNA

<213> Artificial Sequence

35

<220>

<223> Synthetic Oligonucleotide

<400> 100

40

agtcttaaat tccctcaaggc agg 23

<210> 101

<211> 23

<212> DNA

<213> Artificial Sequence

45

<220>

<223> Synthetic Oligonucleotide

<400> 101

agtgctcatc taccagaaggc tgg 23

50

<210> 102

<211> 23

<212> DNA

<213> Artificial Sequence

55

<220>

<223> Synthetic Oligonucleotide

<400> 102

cctcctcatc tccctgcagc agg 23
 <210> 103
 <211> 23
 5 <212> DNA
 <213> Artificial Sequence
 <220>
 <223> Synthetic Oligonucleotide
 10 <400> 103
 ctacatcatc tccctaaggc tgg 23
 <210> 104
 15 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220>
 20 <223> Synthetic Oligonucleotide
 <400> 104
 ggtcctcatc tccctaaaac aga 23
 25 <210> 105
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 30 <220>
 <223> Synthetic Oligonucleotide
 <400> 105
 tgtcctcatc ggcctcaggc agg 23
 35 <210> 106
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 40 <220>
 <223> Synthetic Oligonucleotide
 <400> 106
 45 agacaccatc tccctgagc tgg 23
 <210> 107
 <211> 23
 <212> DNA
 50 <213> Artificial Sequence
 <220>
 <223> Synthetic Oligonucleotide
 <400> 107
 55 aggcatcatc tacatcaagt tgg 23
 <210> 108

<211> 23
<212> DNA
<213> Artificial Sequence

5 <220>
 <223> Synthetic Oligonucleotide

 <400> 108
 agtaatcaact tccatcaagc cgg 23

10 <210> 109
 <211> 23
 <212> DNA
 <213> Artificial Sequence

15 <220>
 <223> Synthetic Oligonucleotide

 <400> 109
 tcccctcacc tccctaaagc agg 23

20 <210> 110
 <211> 23
 <212> DNA
 <213> Artificial Sequence

25 <220>
 <223> Synthetic Oligonucleotide

 <400> 110
 tgtcttatt tccctctagc tgg 23

30 <210> 111
 <211> 23
 <212> DNA
 <213> Artificial Sequence

35 <220>
 <223> Synthetic Oligonucleotide

 <400> 111
 agtcctcatc tccctcaagc agg 23

40 <210> 112
 <211> 23
 <212> DNA
 <213> Artificial Sequence

45 <220>
 <223> Synthetic Oligonucleotide

 <400> 112
 gcagatgtag tgttccaca ggg 23

50 <210> 113
 <211> 23
 <212> DNA
 <213> Artificial Sequence

<220>
<223> Synthetic Oligonucleotide
5 <400> 113
 acatatgtat tattccaca ggg 23

<210> 114
<211> 23
<212> DNA
10 <213> Artificial Sequence

<220>
<223> Synthetic Oligonucleotide
15 <400> 114
 gcatatgtat tgttccaaa tgt 23

<210> 115
<211> 23
20 <212> DNA
 <213> Artificial Sequence

<220>
<223> Synthetic Oligonucleotide
25 <400> 115
 ccagatgtat tattccaca ggg 23

<210> 116
30 <211> 23
 <212> DNA
 <213> Artificial Sequence

<220>
35 <223> Synthetic Oligonucleotide

<400> 116
 gcagtttag tgtttcaca ggg 23

40 <210> 117
 <211> 23
 <212> DNA
 <213> Artificial Sequence

45 <220>
 <223> Synthetic Oligonucleotide

<400> 117
50 gcagagttag tgttccaca cag 23

<210> 118
<211> 23
<212> DNA
55 <213> Artificial Sequence

<220>
55 <223> Synthetic Oligonucleotide

1 <400> 118
 2 gcagatggag ggtttcacca ggg 23
 3
 4 <210> 119
 5 <211> 23
 6 <212> DNA
 7 <213> Artificial Sequence
 8
 9 <220>
 10 <223> Synthetic Oligonucleotide
 11
 12 <400> 119
 13 ggaaattag tgttccacca ggg 23
 14
 15 <210> 120
 16 <211> 23
 17 <212> DNA
 18 <213> Artificial Sequence
 19
 20 <220>
 21 <223> Synthetic Oligonucleotide
 22
 23 <400> 120
 24 aaagaatgtat tattccacca tgg 23
 25
 26 <210> 121
 27 <211> 23
 28 <212> DNA
 29 <213> Artificial Sequence
 30
 31 <220>
 32 <223> Synthetic Oligonucleotide
 33
 34 <400> 121
 35 aaagatgtat tcattccacca agg 23
 36
 37 <210> 122
 38 <211> 23
 39 <212> DNA
 40 <213> Artificial Sequence
 41
 42 <220>
 43 <223> Synthetic Oligonucleotide
 44
 45 <400> 122
 46 aaatatgtat tctttccacca ggg 23
 47
 48 <210> 123
 49 <211> 23
 50 <212> DNA
 51 <213> Artificial Sequence
 52
 53 <220>
 54 <223> Synthetic Oligonucleotide
 55
 56 <400> 123
 57 atagatgtat tgtttccaaa gga 23

5 <210> 124
 <211> 23
 <212> DNA
 <213> Artificial Sequence

10 <220>
 <223> Synthetic Oligonucleotide

 <400> 124
 ccagaggtag tgctccaca ggg 23

15 <210> 125
 <211> 23
 <212> DNA
 <213> Artificial Sequence

 <220>
 <223> Synthetic Oligonucleotide

20 <400> 125
 ccagatgtga ggttccaca agg 23

 <210> 126
 <211> 23
 <212> DNA
 <213> Artificial Sequence

 <220>
 <223> Synthetic Oligonucleotide

25 <400> 126
 ctacatgtag tgttccata tgg 23

 <210> 127
 <211> 23
 <212> DNA
 <213> Artificial Sequence

 <220>
 <223> Synthetic Oligonucleotide

30 <400> 126
 ctacatgtag tgttccata tgg 23

 <210> 127
 <211> 23
 <212> DNA
 <213> Artificial Sequence

 <220>
 <223> Synthetic Oligonucleotide

35 <400> 127
 cttagatgaag tgcttccaca tgg 23

 <210> 128
 <211> 23
 <212> DNA
 <213> Artificial Sequence

 <220>
 <223> Synthetic Oligonucleotide

40 <400> 127
 cttagatgaag tgcttccaca tgg 23

 <210> 128
 <211> 23
 <212> DNA
 <213> Artificial Sequence

 <220>
 <223> Synthetic Oligonucleotide

45 <400> 128
 gaaaatggag tgtttacaca tgg 23

 <210> 129
 <211> 23
 <212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

5

<400> 129

gcaaatgaag tgtcaccaca agg 23

10 <210> 130

<211> 23

<212> DNA

<213> Artificial Sequence

15 <220>

<223> Synthetic Oligonucleotide

<400> 130

gcaaatgtat tatttccact agg 23

20 <210> 131

<211> 23

<212> DNA

<213> Artificial Sequence

25 <220>

<223> Synthetic Oligonucleotide

<400> 131

gcagatgtag ctttgtaca tgg 23

30

<210> 132

<211> 23

<212> DNA

<213> Artificial Sequence

35 <220>

<223> Synthetic Oligonucleotide

<400> 132

gcagcttaag tgtttcaca tgg 23

40 <210> 133

<211> 23

<212> DNA

45 <213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

50 <400> 133

ttacatgtag tgttcacaca cg 23

<210> 134

<211> 23

55 <212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

<400> 134

gaagaggaag tggccca ggg 23

5

<210> 135

<211> 23

<212> DNA

<213> Artificial Sequence

10

<220>

<223> Synthetic Oligonucleotide

<400> 135

15

gaagatgtgg agttgacaca tgg 23

<210> 136

<211> 23

<212> DNA

20

<213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

25

<400> 136

gcagaagtac tggtttaca agg 23

<210> 137

<211> 23

30

<212> DNA

<213> Artificial Sequence

35

<220>

<223> Synthetic Oligonucleotide

40

<400> 137

gcagatgtgg aattacaaca ggg 23

<210> 138

<211> 23

<212> DNA

<213> Artificial Sequence

45

<220>

<223> Synthetic Oligonucleotide

<400> 138

gcagtcatag tgtatacaca tgg 23

50

<210> 139

<211> 23

<212> DNA

<213> Artificial Sequence

55

<220>

<223> Synthetic Oligonucleotide

<400> 139

taagatgtag tatttccaaa agt 23
 <210> 140
 <211> 23
 5 <212> DNA
 <213> Artificial Sequence
 <220>
 <223> Synthetic Oligonucleotide
 10 <400> 140
 gcagctggca ttctccaca cg 23
 <210> 141
 15 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220>
 20 <223> Synthetic Oligonucleotide
 <400> 141
 ggagatctga tggttctaca agg 23
 25 <210> 142
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 30 <220>
 <223> Synthetic Oligonucleotide
 <400> 142
 taaaatgcag tgttatccata tgg 23
 35 <210> 143
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 40 <220>
 <223> Synthetic Oligonucleotide
 <400> 143
 45 gccagaatag ttttcaaca agg 23
 <210> 144
 <211> 23
 <212> DNA
 50 <213> Artificial Sequence
 <220>
 <223> Synthetic Oligonucleotide
 <400> 144
 ttgtatTTAG agattgcaca agg 23
 55 <210> 145

<211> 23
 <212> DNA
 <213> Artificial Sequence
 5 <220>
 <223> Synthetic Oligonucleotide

 <400> 145
 agtcctcatc tccctcaagg agg 23
 10 <210> 146
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 15 <220>
 <223> Synthetic Oligonucleotide

 <400> 146
 20 agtcctcaac tccctcaagg agg 23

 <210> 147
 <211> 23
 <212> DNA
 25 <213> Artificial Sequence

 <220>
 <223> Synthetic Oligonucleotide

 30 <400> 147
 agccctcatt tccctcaagg agg 23

 <210> 148
 <211> 23
 35 <212> DNA
 <213> Artificial Sequence

 <220>
 <223> Synthetic Oligonucleotide
 40 <400> 148
 actcctcatc cccctcaagg cgg 23

 <210> 149
 <211> 23
 <212> DNA
 45 <213> Artificial Sequence

 <220>
 <223> Synthetic Oligonucleotide

 50 <400> 149
 agtcatcatc tccctcaagg aga 23

 <210> 150
 <211> 23
 <212> DNA
 55 <213> Artificial Sequence

<220>
<223> Synthetic Oligonucleotide
5 <400> 150
 cgtcctcc tc ccccaagg agg 23

 <210> 151
 <211> 23
 <212> DNA
10 <213> Artificial Sequence

 <220>
 <223> Synthetic Oligonucleotide

15 <400> 151
 tgtccttc tccctcaagg aga 23

 <210> 152
 <211> 23
20 <212> DNA
 <213> Artificial Sequence

 <220>
 <223> Synthetic Oligonucleotide
25 <400> 152
 aagcttcatc tctctcaagg tgg 23

 <210> 153
 <211> 23
30 <212> DNA
 <213> Artificial Sequence

 <220>
35 <223> Synthetic Oligonucleotide

 <400> 153
 agtactctt tccctcaggc tgg 23

40 <210> 154
 <211> 23
 <212> DNA
 <213> Artificial Sequence

45 <220>
 <223> Synthetic Oligonucleotide

 <400> 154
50 agtctaaat tccctcaagg agg 23

 <210> 155
 <211> 23
 <212> DNA
 <213> Artificial Sequence

55 <220>
 <223> Synthetic Oligonucleotide

1 <400> 155
 agtgctcatc taccagaagc tgg 23
 5 <210> 156
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 10 <220>
 <223> Synthetic Oligonucleotide
 <400> 156
 cctcctcatc tccctgcagc agg 23
 15 <210> 157
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 20 <220>
 <223> Synthetic Oligonucleotide
 <400> 157
 ctacatcatc tccctaagg tgg 23
 25 <210> 158
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 30 <220>
 <223> Synthetic Oligonucleotide
 <400> 158
 35 ggtcctcatc tccctaaaac aga 23
 <210> 159
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 40 <220>
 <223> Synthetic Oligonucleotide
 <400> 159
 tgccctcatc ggcctcaggc agg 23
 45 <210> 160
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 50 <220>
 <223> Synthetic Oligonucleotide
 <400> 160
 agacaccatc tcccttgagc tgg 23
 55

<210> 161
<211> 23
<212> DNA
<213> Artificial Sequence
5
<220>
<223> Synthetic Oligonucleotide

<400> 161
10 aggcatcatc tacatcaagt tgg 23

<210> 162
<211> 23
<212> DNA
15 <213> Artificial Sequence

<220>
<223> Synthetic Oligonucleotide

<400> 162
20 agtaatcact tccatcaagc cg 23

<210> 163
<211> 23
25 <212> DNA
<213> Artificial Sequence

<220>
<223> Synthetic Oligonucleotide
30
<400> 163
tccccctacc tccctaaagc agg 23

<210> 164
35 <211> 23
<212> DNA
<213> Artificial Sequence

<220>
40 <223> Synthetic Oligonucleotide

<400> 164
tgttttatt tccctcttagc tgg 23

<210> 165
45 <211> 23
<212> DNA
<213> Artificial Sequence

<220>
50 <223> Synthetic Oligonucleotide

<400> 165
agtccatc tccctcaagc agg 23
55
<210> 166
<211> 23
<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

5

<400> 166

agtccatc tccctcaagg agg 23

10 <210> 167

<211> 24

<212> DNA

<213> Artificial Sequence

15 <220>

<223> Synthetic Oligonucleotide

<400> 167

agtccatc tccctcaaag cagg 24

20 <210> 168

<211> 22

<212> DNA

<213> Artificial Sequence

25 <220>

<223> Synthetic Oligonucleotide

<400> 168

agtccatc tccctcagca gg 22

30

<210> 169

<211> 24

<212> DNA

<213> Artificial Sequence

35 <220>

<223> Synthetic Oligonucleotide

<400> 169

40 agtccatc tccctcatag cagg 24

<210> 170

<211> 15

<212> DNA

45 <213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

50 <400> 170

agtccatcata gcagg 15

<210> 171

<211> 17

55 <212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

<400> 171

agtccatc tagcagg 17

5

<210> 172

<211> 14

<212> DNA

<213> Artificial Sequence

10

<220>

<223> Synthetic Oligonucleotide

<400> 172

15

agtccatcaag cagg 14

<210> 173

<211> 51

<212> DNA

20

<213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

25

<400> 173

agtccatc tccctcaaag gcagtgttg ttacttgagt ttgtcagcag g 51

<210> 174

<211> 25

30

<212> DNA

<213> Artificial Sequence

35

<220>

<223> Synthetic Oligonucleotide

40

<400> 174

agtccatc tccctatttgcagg 25

<210> 175

<211> 65

<212> DNA

<213> Artificial Sequence

45

<220>

<223> Synthetic Oligonucleotide

<400> 175

50

agtccatc tccctcagggttttaca gtcacccttt gaatttgcac aagcgtgcaa 60

60

gcagg

65

55

<210> 176

<211> 21

<212> DNA

<213> Artificial Sequence

<220>
<223> Synthetic Oligonucleotide
5 <400> 176
 agtccatc tcccttagcag g 21

 <210> 177
 <211> 21
 <212> DNA
10 <213> Artificial Sequence

 <220>
 <223> Synthetic Oligonucleotide
15 <400> 177
 agtccatc cctcaaggcag g 21

 <210> 178
 <211> 22
20 <212> DNA
 <213> Artificial Sequence

 <220>
 <223> Synthetic Oligonucleotide
25 <400> 178
 agtccatc tccctaaggca gg 22

 <210> 179
 <211> 23
 <212> DNA
30 <213> Artificial Sequence

 <220>
 <223> Synthetic Oligonucleotide
35 <400> 179
 agtccatc ac tccctcaaggc agg 23

 <210> 180
 <211> 24
 <212> DNA
40 <213> Artificial Sequence

45 <220>
 <223> Synthetic Oligonucleotide

 <400> 180
 agtccatc ac tccctcaaag cagg 24
50 <210> 181
 <211> 17
 <212> DNA
 <213> Artificial Sequence
55 <220>
 <223> Synthetic Oligonucleotide

1 <400> 181
 agtccctcaac tccctca 17
 5 <210> 182
 <211> 22
 <212> DNA
 <213> Artificial Sequence
 10 <220>
 <223> Synthetic Oligonucleotide
 <400> 182
 agtccctcaac tccctcagca gg 22
 15 <210> 183
 <211> 56
 <212> DNA
 <213> Artificial Sequence
 20 <220>
 <223> Synthetic Oligonucleotide
 <400> 183
 agtccctcaac tccctcaaga aagggtgtga aaatcagaaaa gagagaaaaca agcagg 56
 25 <210> 184
 <211> 62
 <212> DNA
 <213> Artificial Sequence
 30 <220>
 <223> Synthetic Oligonucleotide
 <400> 184
 35 agtccctcaac tccctcaatc tacggtccat tcccgtttcc actcaccttg cgccgcagca 60
 gg 62
 40 <210> 185
 <211> 22
 <212> DNA
 <213> Artificial Sequence
 45 <220>
 <223> Synthetic Oligonucleotide
 <400> 185
 50 agtccctcaac tccctaagca gg 22
 <210> 186
 <211> 66
 <212> DNA
 <213> Artificial Sequence
 55 <220>
 <223> Synthetic Oligonucleotide

<400> 186

5	agtccctcaac tccctcaacc aactttaaca tcctgctgg tctgtcatta ataagttgaa	60
	agcagg	66

<210> 187

<211> 64

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

<400> 187

15	agtccctcaac tccctcacag caaataaaaa agttgtttat gcatttcag ataagcaaag	60
	cagg	64

<210> 188

<211> 21

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

<400> 188

agtccctcaac tcccaaggcag g 21

<210> 189

<211> 23

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

<400> 189

actcctcatc cccctcaagg ccgg 23

<210> 190

<211> 24

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

<400> 190

actcctcatc cccctcaaag ccgg 24

<210> 191

<211> 24

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

5

<400> 191

actcctcatc cccctcagag ccgg 24

10 <210> 192

<211> 13

<212> DNA

<213> Artificial Sequence

15 <220>

<223> Synthetic Oligonucleotide

<400> 192

actcctcagc cg 13

20 <210> 193

<211> 25

<212> DNA

<213> Artificial Sequence

25 <220>

<223> Synthetic Oligonucleotide

<400> 193

actcctcatc cccctcaaaa gccgg 25

30

<210> 194

<211> 25

<212> DNA

<213> Artificial Sequence

35 <220>

<223> Synthetic Oligonucleotide

<400> 194

40 actcctcatc cccctcagca gccgg 25

<210> 195

<211> 24

<212> DNA

45 <213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

50 <400> 195

actcctcatc cccctcatag ccgg 24

<210> 196

<211> 24

55 <212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

<400> 196

actcctcatc cccctcatcc ccgg 24

5

<210> 197

<211> 18

<212> DNA

<213> Artificial Sequence

10

<220>

<223> Synthetic Oligonucleotide

<400> 197

15

actcctcatc ccagccgg 18

<210> 198

<211> 22

<212> DNA

20

<213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

25

<400> 198

actcctcatc cccctaagcc gg 22

<210> 199

<211> 25

30

<212> DNA

<213> Artificial Sequence

35

<220>

<223> Synthetic Oligonucleotide

40

<400> 199

actcctcatc cccctaata gccgg 25

<210> 200

<211> 25

<212> DNA

<213> Artificial Sequence

45

<220>

<223> Synthetic Oligonucleotide

<400> 200

actcctcacc cccctcagca gccgg 25

50

<210> 201

<211> 23

<212> DNA

<213> Artificial Sequence

55

<220>

<223> Synthetic Oligonucleotide

<400> 201

gcagatgtag tgttccaca ggg 23
 <210> 202
 <211> 22
 5 <212> DNA
 <213> Artificial Sequence

 <220>
 <223> Synthetic Oligonucleotide
 10 <400> 202
 gcagatgtag tgttcacag gg 22

 <210> 203
 15 <211> 24
 <212> DNA
 <213> Artificial Sequence

 <220>
 20 <223> Synthetic Oligonucleotide

 <400> 203
 gcagatgtag tgttccac aggg 24
 25 <210> 204
 <211> 19
 <212> DNA
 <213> Artificial Sequence

 30 <220>
 <223> Synthetic Oligonucleotide

 <400> 204
 gcagatgtag tgcacaggg 19
 35 <210> 205
 <211> 22
 <212> DNA
 <213> Artificial Sequence
 40 <220>
 <223> Synthetic Oligonucleotide

 <400> 205
 45 gcagatgtag tgttccag gg 22

 <210> 206
 <211> 16
 <212> DNA
 50 <213> Artificial Sequence

 <220>
 <223> Synthetic Oligonucleotide

 <400> 206
 55 gcagatgtac acaggg 16

 <210> 207

<211> 20
<212> DNA
<213> Artificial Sequence

5 <220>
 <223> Synthetic Oligonucleotide

 <400> 207
 gcagatgtag tgtcacaggg 20

10 <210> 208
 <211> 21
 <212> DNA
 <213> Artificial Sequence

15 <220>
 <223> Synthetic Oligonucleotide

 <400> 208
 gcagatgtag tgttcacagg g 21

20 <210> 209
 <211> 21
 <212> DNA
 <213> Artificial Sequence

25 <220>
 <223> Synthetic Oligonucleotide

 <400> 209
 gcagatgtag tgttccagg g 21

30 <210> 210
 <211> 18
 <212> DNA
 <213> Artificial Sequence

35 <220>
 <223> Synthetic Oligonucleotide

 <400> 210
 gcagatgtag tcacaggg 18

40 <210> 211
 <211> 12
 <212> DNA
 <213> Artificial Sequence

45 <220>
 <223> Synthetic Oligonucleotide

 <400> 211
 gcagatgtag gg 12

50 <210> 212
 <211> 23
 <212> DNA
 <213> Artificial Sequence

<220>
<223> Synthetic Oligonucleotide
5 <400> 212
 acatatgtat tattccaca ggg 23

<210> 213
<211> 24
<212> DNA
10 <213> Artificial Sequence

<220>
<223> Synthetic Oligonucleotide
15 <400> 213
 acatatgtat tattccac aggg 24

<210> 214
<211> 22
20 <212> DNA
 <213> Artificial Sequence

<220>
<223> Synthetic Oligonucleotide
25 <400> 214
 acatatgtat tattcacag gg 22

<210> 215
30 <211> 21
 <212> DNA
 <213> Artificial Sequence

<220>
35 <223> Synthetic Oligonucleotide

<400> 215
 acatatgtat tattccagg g 21

40 <210> 216
 <211> 13
 <212> DNA
 <213> Artificial Sequence

45 <220>
 <223> Synthetic Oligonucleotide

<400> 216
50 acatatacaca ggg 13

<210> 217
<211> 11
<212> DNA
55 <213> Artificial Sequence

<220>
55 <223> Synthetic Oligonucleotide

1 <400> 217
 acatcacagg g 11
 5 <210> 218
 <211> 18
 <212> DNA
 <213> Artificial Sequence
 10 <220>
 <223> Synthetic Oligonucleotide
 <400> 218
 acatatgtag tcacaggg 18
 15 <210> 219
 <211> 17
 <212> DNA
 <213> Artificial Sequence
 20 <220>
 <223> Synthetic Oligonucleotide
 <400> 219
 acatatgtag tatttcc 17
 25 <210> 220
 <211> 12
 <212> DNA
 <213> Artificial Sequence
 30 <220>
 <223> Synthetic Oligonucleotide
 <400> 220
 acatatgtag gg 12
 35 <210> 221
 <211> 17
 <212> DNA
 <213> Artificial Sequence
 40 <220>
 <223> Synthetic Oligonucleotide
 <400> 221
 acatatgtag cacaggg 17
 45 <210> 222
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 50 <220>
 <223> Synthetic Oligonucleotide
 <400> 222
 ccagatgtag tattccccaca ggg 23
 55

<210> 223
<211> 24
<212> DNA
<213> Artificial Sequence
5
<220>
<223> Synthetic Oligonucleotide

<400> 223
10 ccagatgtat tattccac aggg 24

<210> 224
<211> 19
<212> DNA
15 <213> Artificial Sequence

<220>
<223> Synthetic Oligonucleotide

<400> 224
20 ccagatgtatatacagg 19

<210> 225
<211> 20
25 <212> DNA
<213> Artificial Sequence

<220>
<223> Synthetic Oligonucleotide
30
<400> 225
ccagatgtat tattcacagg 20

<210> 226
35 <211> 26
<212> DNA
<213> Artificial Sequence

<220>
40 <223> Synthetic Oligonucleotide

<400> 226
ccagatgtat tattccaa acagg 26

<210> 227
45 <211> 21
<212> DNA
<213> Artificial Sequence

<220>
50 <223> Synthetic Oligonucleotide

<400> 227
ccagatgtat tattcacagg g 21
55
<210> 228
<211> 21
<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

5

<400> 228

ccagatgtat tattcccagg g 21

10 <210> 229

<211> 22

<212> DNA

<213> Artificial Sequence

15 <220>

<223> Synthetic Oligonucleotide

<400> 229

ccagatgtat tattccacag gg 22

20 <210> 230

<211> 23

<212> DNA

<213> Artificial Sequence

25 <220>

<223> Synthetic Oligonucleotide

<400> 230

ctagatgaag tgcttccaca tgg 23

30

<210> 231

<211> 24

<212> DNA

<213> Artificial Sequence

35 <220>

<223> Synthetic Oligonucleotide

<400> 231

40 ctagatgaag tgcttccac atgg 24

<210> 232

<211> 22

<212> DNA

45 <213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

50 <400> 232

ctagatgaag tgcttcacat gg 22

<210> 233

<211> 12

55 <212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

<400> 233

ctagatgaag tg

12

5

<210> 234

<211> 25

<212> DNA

<213> Artificial Sequence

10

<220>

<223> Synthetic Oligonucleotide

<400> 234

15

ctagatgaag tgcttccaca catgg 25

<210> 235

<211> 21

<212> DNA

20

<213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

25

<400> 235

ctagatgaag tgcttccatg g 21

<210> 236

<211> 22

30

<212> DNA

<213> Artificial Sequence

35

<220>

<223> Synthetic Oligonucleotide

<400> 236

40

ctagatgaag tgcttccat gg 22

<210> 237

<211> 56

<212> DNA

<213> Artificial Sequence

45

<220>

<223> Synthetic Oligonucleotide

<400> 237

taatacgact cactatagga gtcctcatct ccctcaagcg ttttagagct atgcgt

56

50

<210> 238

<211> 56

<212> DNA

<213> Artificial Sequence

55

<220>

<223> Synthetic Oligonucleotide

<400> 238

taatacgact cactataggc tccctcaagc aggccccgcg ttttagagct atgctg 56
 <210> 239
 <211> 56
 5 <212> DNA
 <213> Artificial Sequence

 <220>
 <223> Synthetic Oligonucleotide
 10 <400> 239
 taatacgact cactataggc gtgaagagct tcactgagtg ttttagagct atgctg 56

 <210> 240
 <211> 56
 15 <212> DNA
 <213> Artificial Sequence

 <220>
 20 <223> Synthetic Oligonucleotide

 <400> 240
 taatacgact cactataggc cagatgttgt gttccacag ttttagagct atgctg 56

 25 <210> 241
 <211> 55
 <212> DNA
 <213> Artificial Sequence

 30 <220>
 <223> Synthetic Oligonucleotide

 <400> 241
 gataacggac tagcattttaacttgcgtatgtttca gcatacgatctt aaaaac 55
 35 <210> 242
 <211> 81
 <212> DNA
 <213> Artificial Sequence

 40 <220>
 <223> Synthetic Oligonucleotide

 <400> 242
 45 <210> 243
 <211> 81
 <212> DNA
 <213> Artificial Sequence

 50 <220>
 <223> Synthetic Oligonucleotide

 cgacttagcc ttattttaac ttgctatttc tagctctaaa acgcttgagg gagatgagga 60

 ctcctatagt gagtcgtatt a 81

 55 <210> 243
 <211> 81
 <212> DNA
 <213> Artificial Sequence

 <220>
 <223> Synthetic Oligonucleotide

<400> 243

5 cgga~~c~~actagcc ttat~~tta~~ac ttgctattc tagctctaaa acgcggggcc tgcttgaggg 60

agc~~c~~ctata~~gt~~ gagtcgtatt a 81

<210> 244

<211> 81

<212> DNA

<213> Artificial Sequence

<220>

15 <223> Synthetic Oligonucleotide

<400> 244

20 cgga~~c~~actagcc ttat~~tta~~ac ttgctattc tagctctaaa acactcagtg aagctttca 60

cac~~c~~ctata~~gt~~ gagtcgtatt a 81

<210> 245

<211> 81

<212> DNA

<213> Artificial Sequence

<220>

30 <223> Synthetic Oligonucleotide

<400> 245

35 cgga~~c~~actagcc ttat~~tta~~ac ttgctattc tagctctaaa actgtggaaa cactacatct 60

gc~~c~~cctata~~gt~~ gagtcgtatt a 81

<210> 246

<211> 19

<212> DNA

<213> Artificial Sequence

<220>

45 <223> Synthetic Oligonucleotide

<400> 246

taatacgact cactatagg 19

50 <210> 247

<211> 53

<212> DNA

<213> Artificial Sequence

55 <220>

<223> Synthetic Oligonucleotide

<220>

5 <221> misc_feature
 <222> (7)..(10)
 <223> n is a, c, g, or t

 10 <220>
 <221> misc_feature
 <222> (13)..(13)
 <223> n is a, c, g, or t

 15 <220>
 <221> misc_feature
 <222> (34)..(37)
 <223> n is a, c, g, or t

 20 <400> 247
 aacacannnn ccngcttgag ggagatgagg actnnnnnacc tgccgagaac aca 53

 25 <210> 248
 <211> 53
 <212> DNA
 <213> Artificial Sequence

 30 <220>
 <223> Synthetic Oligonucleotide

 35 <220>
 <221> misc_feature
 <222> (7)..(10)
 <223> n is a, c, g, or t

 40 <220>
 <221> misc_feature
 <222> (13)..(13)
 <223> n is a, c, g, or t

 45 <400> 248
 tcgtctnnnn ccngcggggc ctgcttgagg gagnnnnacc tgccgagtct tct 53

 50 <210> 249
 <211> 53
 <212> DNA
 <213> Artificial Sequence

 55 <220>
 <223> Synthetic Oligonucleotide

 <220>
 <221> misc_feature
 <222> (7)..(10)
 <223> n is a, c, g, or t

 <220>
 <221> misc_feature

5 <220> (13)..(13)
 <223> n is a, c, g, or t
 <220>
 <221> misc_feature
 <222> (34)..(37)
 <223> n is a, c, g, or t
 10 <400> 249
 agagaannnn ccnactcagt gaagctttc acannnnacc tgccgagaga gaa 53
 <210> 250
 <211> 53
 <212> DNA
 15 <213> Artificial Sequence
 <220>
 <223> Synthetic Oligonucleotide
 20 <220>
 <221> misc_feature
 <222> (7)..(10)
 <223> n is a, c, g, or t
 25 <220>
 <221> misc_feature
 <222> (13)..(13)
 <223> n is a, c, g, or t
 30 <220>
 <221> misc_feature
 <222> (34)..(37)
 <223> n is a, c, g, or t
 35 <400> 250
 ttgtgtnnnn ccntgtggaa acactacatc tgcnnnnacc tgccgagttt tgt 53
 <210> 251
 <211> 29
 40 <212> DNA
 <213> Artificial Sequence
 <220>
 <223> Synthetic Oligonucleotide
 45 <400> 251
 ctgcgcgtcc tcatctccct caaggcaggc 29
 <210> 252
 <211> 29
 <212> DNA
 <213> Artificial Sequence
 50 <220>
 <223> Synthetic Oligonucleotide
 <400> 252
 agctgcctgc ttgagggaga tgaggactg 29

5 <210> 253
 <211> 29
 <212> DNA
 <213> Artificial Sequence
 10 <220>
 <223> Synthetic Oligonucleotide
 <400> 253
 ctagtctccc tcaaggcaggc cccgctggt 29
 15 <210> 254
 <211> 29
 <212> DNA
 <213> Artificial Sequence
 <220>
 <223> Synthetic Oligonucleotide
 20 <400> 254
 agctaccagc ggggcctgct tgagggaga 29
 <210> 255
 <211> 29
 25 <212> DNA
 <213> Artificial Sequence
 <220>
 <223> Synthetic Oligonucleotide
 30 <400> 255
 ctagctgtga agagcttcac tgagttagga 29
 <210> 256
 <211> 29
 <212> DNA
 <213> Artificial Sequence
 <220>
 <223> Synthetic Oligonucleotide
 40 <400> 256
 agtttcctac tcagtgaagc tcttcacag 29
 <210> 257
 <211> 29
 <212> DNA
 <213> Artificial Sequence
 50 <220>
 <223> Synthetic Oligonucleotide
 <400> 257
 ctatgcaga tgtatgttt ccacagggt 29
 55 <210> 258
 <211> 29
 <212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

5

<400> 258

agctaccctg tggaaacact acatctgca 29

10 <210> 259

<211> 26

<212> DNA

<213> Artificial Sequence

15 <220>

<223> Synthetic Oligonucleotide

<400> 259

gcgacacgga aatgttgaat actcat 26

20 <210> 260

<211> 25

<212> DNA

<213> Artificial Sequence

25 <220>

<223> Synthetic Oligonucleotide

<400> 260

ggagttaggc aactatggat gaacg 25

30

<210> 261

<211> 60

<212> DNA

<213> Artificial Sequence

35 <220>

<223> Synthetic Oligonucleotide

<400> 261

40 actgtgaaga gcttcactga taggattaa gatattgcag atgttgtt tccacagggt 60

<210> 262

<211> 60

<212> DNA

45 <213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

50 <400> 262

actgtgaaga gcttcactga taggattaa gatattgaag atgttgtt tccacagggt 60

<210> 263

<211> 60

55 <212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

<400> 263

actgtgaaga gcttcactga gtaggattaa gatattgaag atgttagtgtt tccactgggt 60

5

<210> 264

<211> 60

<212> DNA

<213> Artificial Sequence

10

<220>

<223> Synthetic Oligonucleotide

<400> 264

15 actgtgaaga gcttcactga gtaggattaa gatattgcag atggagggtt tccacagggt 60

<210> 265

<211> 60

<212> DNA

20 <213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

25

<400> 265

actgtgaaga gcttcactga gtaggattaa gatattgcag atgttagtgtt accagagggt 60

<210> 266

<211> 60

30 <212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

35

<400> 266

actgtgaaga gcttcactga gtaggattaa gatattgggg atgttagtgtt tccactgggt 60

<210> 267

40

<211> 60

<212> DNA

<213> Artificial Sequence

<220>

45

<223> Synthetic Oligonucleotide

<400> 267

tccctcaagc aggccccgct ggtgcactga agagccaccc tgtgaaaaca ctacatctgc 60

50

<210> 268

<211> 60

<212> DNA

<213> Artificial Sequence

55

<220>

<223> Synthetic Oligonucleotide

<400> 268

	tccctcaagc aggccccgct ggtgcactga agagccaccc tgtgaaaca ctacatcttc	60
	<210> 269	
	<211> 60	
5	<212> DNA	
	<213> Artificial Sequence	
	<220>	
10	<223> Synthetic Oligonucleotide	
	<400> 269	
	tccctcaagc aggccccgct ggtgcactga agagccaccc agtgaaaca ctacatcttc	60
	<210> 270	
15	<211> 60	
	<212> DNA	
	<213> Artificial Sequence	
	<220>	
20	<223> Synthetic Oligonucleotide	
	<400> 270	
	tccctcaagc aggccccgct ggtgcactga agagccaccc tggaaacc ctccatctgc	60
25	<210> 271	
	<211> 60	
	<212> DNA	
	<213> Artificial Sequence	
30	<220>	
	<223> Synthetic Oligonucleotide	
	<400> 271	
	tccctcaagc aggccccgct ggtgcactga agagccaccc tctggtaaca ctacatctgc	60
35	<210> 272	
	<211> 60	
	<212> DNA	
	<213> Artificial Sequence	
40	<220>	
	<223> Synthetic Oligonucleotide	
	<400> 272	
45	tccctcaagc aggccccgct ggtgcactga agagccaccc agtgaaaca ctacatcccc	60
	<210> 273	
	<211> 62	
	<212> DNA	
50	<213> Artificial Sequence	
	<220>	
	<223> Synthetic Oligonucleotide	
55	<400> 273	
	aatatacg cgaccaccga gatctacact cttccctac acgacgctct tccgatctaa	60
	ca 62	

5 <210> 274
 <211> 49
 <212> DNA
 <213> Artificial Sequence
 10 <220>
 <223> Synthetic Oligonucleotide
 <400> 274
 tgtagatcg gaagagcgtc gtgtaggaa agagtgtaga tctcggtgg 49

15 <210> 275
 <211> 62
 <212> DNA
 <213> Artificial Sequence
 20 <220>
 <223> Synthetic Oligonucleotide
 <400> 275
 aatgatacgg cgaccaccga gatctacact cttccctac acgacgctct tccgatctt 60
 25 ca 62
 <210> 276
 <211> 49
 <212> DNA
 <213> Artificial Sequence
 30 <220>
 <223> Synthetic Oligonucleotide
 <400> 276
 tgaaagatcg gaagagcgtc gtgtaggaa agagtgtaga tctcggtgg 49

35 <210> 277
 <211> 58
 <212> DNA
 <213> Artificial Sequence
 40 <220>
 <223> Synthetic Oligonucleotide
 <400> 277
 aatgatacgg cgaccaccga gatctacact cttccctac acgacgctct tccgatct 58

45 <210> 278
 <211> 45
 <212> DNA
 <213> Artificial Sequence
 50 <220>
 <223> Synthetic Oligonucleotide
 <400> 278
 agatcggaaag agcgtcgttgtggaaagag tgttagatctc ggtgg 45

5 <210> 279
 <211> 15
 <212> DNA
 <213> Artificial Sequence

10 <220>
 <223> Synthetic Oligonucleotide

 <400> 279
 gacggcatac gagat 15

15 <210> 280
 <211> 28
 <212> DNA
 <213> Artificial Sequence

 <220>
 <223> Synthetic Oligonucleotide

20 <400> 280
 aacaatctcg tatgccgtct tctgcttg 28

 <210> 281
 <211> 28
25 <212> DNA
 <213> Artificial Sequence

 <220>
 <223> Synthetic Oligonucleotide

30 <400> 281
 tcttatctcg tatgccgtct tctgcttg 28

 <210> 282
 <211> 28
35 <212> DNA
 <213> Artificial Sequence

 <220>
40 <223> Synthetic Oligonucleotide

 <400> 282
 agagatctcg tatgccgtct tctgcttg 28

45 <210> 283
 <211> 28
 <212> DNA
 <213> Artificial Sequence

50 <220>
 <223> Synthetic Oligonucleotide

 <400> 283
 ttgtatctcg tatgccgtct tctgcttg 28
55 <210> 284
 <211> 40
 <212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

5

<400> 284

caagcagaag acggcatacg agattgtgtt ctccggcagg 40

10 <210> 285

<211> 40

<212> DNA

<213> Artificial Sequence

15 <220>

<223> Synthetic Oligonucleotide

<400> 285

caagcagaag acggcatacg agatagaaga ctccggcagg 40

20 <210> 286

<211> 40

<212> DNA

<213> Artificial Sequence

25 <220>

<223> Synthetic Oligonucleotide

<400> 286

caagcagaag acggcatacg agatttcctt ctccggcagg 40

30

<210> 287

<211> 40

<212> DNA

<213> Artificial Sequence

35 <220>

<223> Synthetic Oligonucleotide

<400> 287

40 caagcagaag acggcatacg agatacacaa ctccggcagg 40

<210> 288

<211> 20

<212> DNA

45 <213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

50 <400> 288

aatgatacgg cgaccaccga 20

<210> 289

<211> 82

55 <212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide
 <220>
 <221> misc_feature
 5 <222> (59)..(62)
 <223> n is a, c, g, or t

 <400> 289

 10 aatgatacgg cgaccaccga gatctacact ctttccctac acgacgctct tccgatctnn 60

 nnacacctacct gccgagaaca ca 82

 15 <210> 290
 <211> 82
 <212> DNA
 <213> Artificial Sequence

 20 <220>
 <223> Synthetic Oligonucleotide

 <220>
 <221> misc_feature
 25 <222> (59)..(62)
 <223> n is a, c, g, or t

 <400> 290

 30 aatgatacgg cgaccaccga gatctacact ctttccctac acgacgctct tccgatctnn 60

 nnacacctacct gccgagtctt ct 82

 35 <210> 291
 <211> 82
 <212> DNA
 <213> Artificial Sequence

 40 <220>
 <223> Synthetic Oligonucleotide

 <220>
 <221> misc_feature
 45 <222> (59)..(62)
 <223> n is a, c, g, or t

 <400> 291

 50 aatgatacgg cgaccaccga gatctacact ctttccctac acgacgctct tccgatctnn 60

 nnacacctacct gccgagagag aa 82

 55 <210> 292
 <211> 82
 <212> DNA
 <213> Artificial Sequence

<220>
 <223> Synthetic Oligonucleotide
 5 <220>
 <221> misc_feature
 <222> (59)..(62)
 <223> n is a, c, g, or t

 <400> 292
 10

 aatgataacgg cgaccaccga gatctacact ctttccctac acgacgctct tccgatctnn 60

 nnacacctacct gcccagttgt gt 82
 15

 <210> 293
 <211> 24
 <212> DNA
 <213> Artificial Sequence
 20

 <220>
 <223> Synthetic Oligonucleotide

 <400> 293
 25 caaggcagaag acggcatacg agat 24

 <210> 294
 <211> 53
 <212> DNA
 <213> Artificial Sequence
 30

 <220>
 <223> Synthetic Oligonucleotide

 <400> 294
 35 acactcttc cctacacgcac gctctccga tctcaagtct agcaaggcagg cca 53

 <210> 295
 <211> 53
 <212> DNA
 <213> Artificial Sequence

 <220>
 <223> Synthetic Oligonucleotide
 45

 <400> 295
 acactcttc cctacacgcac gctctccga tctcaggcac tgagtggaa agt 53

 <210> 296
 <211> 53
 <212> DNA
 <213> Artificial Sequence

 <220>
 <223> Synthetic Oligonucleotide

 <400> 296
 55 acactcttc cctacacgcac gctctccga tcttaacccc aagtcaagcaa gca 53

5 <210> 297
 <211> 53
 <212> DNA
 <213> Artificial Sequence
 10 <220>
 <223> Synthetic Oligonucleotide
 <400> 297
 acactcttc cctacacgac gctctccga tccttgctgg tcaataccct ggc 53
 15 <210> 298
 <211> 53
 <212> DNA
 <213> Artificial Sequence
 <220>
 <223> Synthetic Oligonucleotide
 20 <400> 298
 acactcttc cctacacgac gctctccga tcctgagttac ccctgaaatg ggc 53
 <210> 299
 <211> 53
 25 <212> DNA
 <213> Artificial Sequence
 <220>
 <223> Synthetic Oligonucleotide
 30 <400> 299
 acactcttc cctacacgac gctctccga tcctcgctac caatcaggc ttt 53
 <210> 300
 <211> 53
 35 <212> DNA
 <213> Artificial Sequence
 <220>
 <223> Synthetic Oligonucleotide
 40 <400> 300
 acactcttc cctacacgac gctctccga tctccattgc cacttgttt cat 53
 <210> 301
 <211> 53
 <212> DNA
 <213> Artificial Sequence
 50 <220>
 <223> Synthetic Oligonucleotide
 <400> 301
 acactcttc cctacacgac gctctccga tctccatccc ccacaactt gct 53
 55 <210> 302
 <211> 53
 <212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

5

<400> 302

acactcttc cctacacgac gctctccga tctgtgtaca tccagtgcac cca 53

10 <210> 303

<211> 57

<212> DNA

<213> Artificial Sequence

15 <220>

<223> Synthetic Oligonucleotide

<400> 303

acactcttc cctacacgac gctctccga tcitcgaaa ggacttgaa tacttgt 57

20 <210> 304

<211> 53

<212> DNA

<213> Artificial Sequence

25 <220>

<223> Synthetic Oligonucleotide

<400> 304

acactcttc cctacacgac gctctccga tctcgccca agacctcatt cac 53

30

<210> 305

<211> 53

<212> DNA

<213> Artificial Sequence

35 <220>

<223> Synthetic Oligonucleotide

<400> 305

40 acactcttc cctacacgac gctctccga tctgtcctct ctggggcaga agt 53

<210> 306

<211> 55

<212> DNA

45 <213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

50 <400> 306

acactcttc cctacacgac gctctccga tctagctgag tcatgaggtg tctcc 55

<210> 307

<211> 53

<212> DNA

55 <213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

<400> 307

acactcttc cttacacgac gctctccga tctctgccag ctctcacac cat 53

5

<210> 308

<211> 53

<212> DNA

<213> Artificial Sequence

10

<220>

<223> Synthetic Oligonucleotide

<400> 308

15 acactcttc cttacacgac gctctccga tctctgaagg acaaaggcgg gaa 53

<210> 309

<211> 53

<212> DNA

20 <213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

25

<400> 309

acactcttc cttacacgac gctctccga tctaagggtgc taaagggtcc acg 53

<210> 310

<211> 53

30 <212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

35

<400> 310

acactcttc cttacacgac gctctccga tctgaccatt ggtgagccca gag 53

<210> 311

40

<211> 53

<212> DNA

<213> Artificial Sequence

<220>

45

<223> Synthetic Oligonucleotide

<400> 311

acactcttc cttacacgac gctctccga tccttttgc ggcaactgtc cac 53

50

<210> 312

<211> 53

<212> DNA

<213> Artificial Sequence

55

<220>

<223> Synthetic Oligonucleotide

<400> 312

acactttc cctacacgac gcttccga tctgcaagcc ttcttcctc aga 53
 <210> 313
 <211> 54
 5 <212> DNA
 <213> Artificial Sequence
 <220>
 <223> Synthetic Oligonucleotide
 10 <400> 313
 acactttc cctacacgac gcttccga tctacacaaa ctcccttag accc 54
 <210> 314
 <211> 53
 15 <212> DNA
 <213> Artificial Sequence
 <220>
 20 <223> Synthetic Oligonucleotide
 <400> 314
 acactttc cctacacgac gcttccga tcttgagttt gcccgttgt tca 53
 25 <210> 315
 <211> 55
 <212> DNA
 <213> Artificial Sequence
 30 <220>
 <223> Synthetic Oligonucleotide
 <400> 315
 acactttc cctacacgac gcttccga tcttgaagag cttcaactgag tagga 55
 35 <210> 316
 <211> 54
 <212> DNA
 <213> Artificial Sequence
 40 <220>
 <223> Synthetic Oligonucleotide
 <400> 316
 45 acactttc cctacacgac gcttccga tcctccctt acagccaatt tcgt 54
 <210> 317
 <211> 58
 <212> DNA
 50 <213> Artificial Sequence
 <220>
 <223> Synthetic Oligonucleotide
 <400> 317
 acactttc cctacacgac gcttccga tcctgctgtt gaaatgcaat taagagg 58
 55 <210> 318

<211> 53
 <212> DNA
 <213> Artificial Sequence

5 <220>
 <223> Synthetic Oligonucleotide

<400> 318
 acactcttc cctacacgac gctctccga tctggccct gcaaggccagt atg 53

10 <210> 319
 <211> 55
 <212> DNA
 <213> Artificial Sequence

15 <220>
 <223> Synthetic Oligonucleotide

<400> 319
 20 acactcttc cctacacgac gctctccga tctatcaaag ccttgtatca cagtt 55

<210> 320
 <211> 54
 <212> DNA

25 <213> Artificial Sequence

<220>
 <223> Synthetic Oligonucleotide

<400> 320
 30 acactcttc cctacacgac gctctccga tctcccaaatt aatgcaggag ccaa 54

<210> 321
 <211> 55

35 <212> DNA
 <213> Artificial Sequence

<220>
 <223> Synthetic Oligonucleotide

<400> 321
 40 acactcttc cctacacgac gctctccga tctctgcctt tagtgggaca gactt 55

<210> 322
 <211> 54

45 <212> DNA
 <213> Artificial Sequence

<220>
 <223> Synthetic Oligonucleotide

<400> 322
 50 acactcttc cctacacgac gctctccga tctagtaacc cttagtagccc tcca 54

<210> 323
 <211> 54

55 <212> DNA
 <213> Artificial Sequence

5 <220>
 <223> Synthetic Oligonucleotide
 <400> 323
 acactcttc cctacacgac gctttccga ttcattgca gtgagccgag attg 54
 10 <210> 324
 <211> 54
 <212> DNA
 <213> Artificial Sequence
 <220>
 <223> Synthetic Oligonucleotide
 15 <400> 324
 acactcttc cctacacgac gctttccga ttggcaaa gttcacttcc atgt 54
 <210> 325
 <211> 53
 20 <212> DNA
 <213> Artificial Sequence
 <220>
 <223> Synthetic Oligonucleotide
 25 <400> 325
 acactcttc cctacacgac gctttccga tcgtctcg tgatgtgc cac 53
 <210> 326
 <211> 55
 <212> DNA
 <213> Artificial Sequence
 <220>
 35 <223> Synthetic Oligonucleotide
 <400> 326
 acactcttc cctacacgac gctttccga tcgtgttag gattgtgaac cagca 55
 40 <210> 327
 <211> 53
 <212> DNA
 <213> Artificial Sequence
 45 <220>
 <223> Synthetic Oligonucleotide
 <400> 327
 acactcttc cctacacgac gctttccga tttccagc ccaggattt tct 53
 50 <210> 328
 <211> 53
 <212> DNA
 <213> Artificial Sequence
 55 <220>
 <223> Synthetic Oligonucleotide

5 <400> 328
 acactcttc cctacacgac gctctccga tctagggtgc tttgtcaca gtc 53
 <210> 329
 <211> 53
 <212> DNA
 <213> Artificial Sequence
 10 <220>
 <223> Synthetic Oligonucleotide
 <400> 329
 acactcttc cctacacgac gctctccga tctcctggct tggatgtt gaa 53
 15 <210> 330
 <211> 53
 <212> DNA
 <213> Artificial Sequence
 20 <220>
 <223> Synthetic Oligonucleotide
 <400> 330
 acactcttc cctacacgac gctctccga tcttgccca aggtcataact gct 53
 25 <210> 331
 <211> 56
 <212> DNA
 <213> Artificial Sequence
 30 <220>
 <223> Synthetic Oligonucleotide
 <400> 331
 acactcttc cctacacgac gctctccga tctaccact agtagccat aatcca 56
 35 <210> 332
 <211> 53
 <212> DNA
 <213> Artificial Sequence
 40 <220>
 <223> Synthetic Oligonucleotide
 <400> 332
 acactcttc cctacacgac gctctccga tctcggtcat gtgcgttga aga 53
 45 <210> 333
 <211> 56
 <212> DNA
 <213> Artificial Sequence
 50 <220>
 <223> Synthetic Oligonucleotide
 <400> 333
 acactcttc cctacacgac gctctccga tcttggccc atattgctt atgctg 56
 55

5 <210> 334
 <211> 53
 <212> DNA
 <213> Artificial Sequence
 10 <220>
 <223> Synthetic Oligonucleotide
 <400> 334
 acactcttc cctacacgac gctctccga tctattaggg gttggctgca tga 53

15 <210> 335
 <211> 53
 <212> DNA
 <213> Artificial Sequence
 20 <220>
 <223> Synthetic Oligonucleotide
 <400> 335
 acactcttc cctacacgac gctctccga tctccaagac gtgttgcatg ctg 53

25 <210> 336
 <211> 57
 <212> DNA
 <213> Artificial Sequence
 30 <220>
 <223> Synthetic Oligonucleotide
 <400> 336
 acactcttc cctacacgac gctctccga tcttggagg tgataaattc cctaaat 57

35 <210> 337
 <211> 53
 <212> DNA
 <213> Artificial Sequence
 40 <220>
 <223> Synthetic Oligonucleotide
 <400> 337
 acactcttc cctacacgac gctctccga tctccagaga caaagggtgg gag 53

45 <210> 338
 <211> 56
 <212> DNA
 <213> Artificial Sequence
 50 <220>
 <223> Synthetic Oligonucleotide
 <400> 338
 acactcttc cctacacgac gctctccga tcttcataca gaagagcaaa gtacca 56

55 <210> 339
 <211> 53
 <212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

5

<400> 339

acactcttc cctacacgac gctctccga tctcaaagag gggtatcggg agc 53

10 <210> 340

<211> 57

<212> DNA

<213> Artificial Sequence

15 <220>

<223> Synthetic Oligonucleotide

<400> 340

acactcttc cctacacgac gctctccga tctaaatgga agaaccaagt agatgaa 57

20 <210> 341

<211> 55

<212> DNA

<213> Artificial Sequence

25 <220>

<223> Synthetic Oligonucleotide

<400> 341

acactcttc cctacacgac gctctccga tccttttgt tgacagatgg ccaca 55

30

<210> 342

<211> 58

<212> DNA

<213> Artificial Sequence

35 <220>

<223> Synthetic Oligonucleotide

<400> 342

40 acactcttc cctacacgac gctctccga tcctcttact tgtgtgattt tagaaca 58

<210> 343

<211> 53

<212> DNA

45 <213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

50 <400> 343

acactcttc cctacacgac gctctccga tctgatgggt catgcagagg gct 53

<210> 344

<211> 53

55 <212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

<400> 344

acactcttc cttacacgac gctctccga tctgctggc ttccctgagc tgt 53

5

<210> 345

<211> 55

<212> DNA

<213> Artificial Sequence

10

<220>

<223> Synthetic Oligonucleotide

<400> 345

15

acactcttc cttacacgac gctctccga tctctccatc agataccgt accca 55

<210> 346

<211> 55

<212> DNA

20

<213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

25

<400> 346

acactcttc cttacacgac gctctccga tctggaaaa cactctctct ctgct 55

<210> 347

<211> 53

30

<212> DNA

<213> Artificial Sequence

35

<220>

<223> Synthetic Oligonucleotide

40

<400> 347

acactcttc cttacacgac gctctccga tctggaggcc acgacacaca ata 53

<210> 348

<211> 54

<212> DNA

<213> Artificial Sequence

45

<220>

<223> Synthetic Oligonucleotide

<400> 348

gtgactggg ttcatcacgtg tgctctccg atctcacagg gtggcttcc agtg 54

50

<210> 349

<211> 54

<212> DNA

<213> Artificial Sequence

55

<220>

<223> Synthetic Oligonucleotide

<400> 349

gtgactggag ttcagacgtg tgcttccg atctgcaca tgttccaca gggt 54
 <210> 350
 <211> 54
 5 <212> DNA
 <213> Artificial Sequence

 <220>
 <223> Synthetic Oligonucleotide
 10 <400> 350
 gtgactggag ttcagacgtg tgcttccg atctagtgtt tccaggagcg gtt 54

 <210> 351
 15 <211> 54
 <212> DNA
 <213> Artificial Sequence

 <220>
 20 <223> Synthetic Oligonucleotide

 <400> 351
 gtgactggag ttcagacgtg tgcttccg atctaaggct caggcacaac tctg 54

 25 <210> 352
 <211> 53
 <212> DNA
 <213> Artificial Sequence

 30 <220>
 <223> Synthetic Oligonucleotide

 <400> 352
 gtgactggag ttcagacgtg tgcttccg atcttagggg aggggcaaag aca 53
 35 <210> 353
 <211> 54
 <212> DNA
 <213> Artificial Sequence

 40 <220>
 <223> Synthetic Oligonucleotide

 <400> 353
 45 gtgactggag ttcagacgtg tgcttccg atctggAAC agtggtatgc tggt 54

 <210> 354
 <211> 55
 <212> DNA
 50 <213> Artificial Sequence

 <220>
 <223> Synthetic Oligonucleotide

 <400> 354
 55 gtgactggag ttcagacgtg tgcttccg atctagtgtg gacactgaca aggaa 55

 <210> 355

<211> 54
 <212> DNA
 <213> Artificial Sequence
 5 <220>
 <223> Synthetic Oligonucleotide

 <400> 355
 gtgactggag ttcagacgtg tgctctccg atcttcactg cctgggtgt tttag 54
 10 <210> 356
 <211> 54
 <212> DNA
 <213> Artificial Sequence
 15 <220>
 <223> Synthetic Oligonucleotide

 <400> 356
 20 gtgactggag ttcagacgtg tgctctccg atcttacccc agcctccagc ttta 54

 <210> 357
 <211> 54
 <212> DNA
 25 <213> Artificial Sequence

 <220>
 <223> Synthetic Oligonucleotide

 30 <400> 357
 gtgactggag ttcagacgtg tgctctccg atcttgacta ctggggagcg atga 54

 <210> 358
 <211> 56
 35 <212> DNA
 <213> Artificial Sequence

 <220>
 <223> Synthetic Oligonucleotide
 40 <400> 358
 gtgactggag ttcagacgtg tgctctccg atcttaggctg ttatgcagga aaggaa 56

 <210> 359
 <211> 54
 <212> DNA
 45 <213> Artificial Sequence

 <220>
 <223> Synthetic Oligonucleotide

 50 <400> 359
 gtgactggag ttcagacgtg tgctctccg atctgcgggt gaggtggatg gaag 54

 <210> 360
 <211> 54
 <212> DNA
 55 <213> Artificial Sequence

5 <220>
 <223> Synthetic Oligonucleotide
 <400> 360
 gtgactggag ttcagacgtg tgcttccg atctggcagc atccttaca tcct 54
 <210> 361
 <211> 57
 <212> DNA
 10 <213> Artificial Sequence
 <220>
 <223> Synthetic Oligonucleotide
 15 <400> 361
 gtgactggag ttcagacgtg tgcttccg atctagaaaa agctccccca gaaagga 57
 <210> 362
 <211> 54
 20 <212> DNA
 <213> Artificial Sequence
 <220>
 <223> Synthetic Oligonucleotide
 25 <400> 362
 gtgactggag ttcagacgtg tgcttccg atctctgcac caaccctcac gtcc 54
 <210> 363
 <211> 54
 <212> DNA
 30 <213> Artificial Sequence
 <220>
 <223> Synthetic Oligonucleotide
 35 <400> 363
 gtgactggag ttcagacgtg tgcttccg atctctggag agggcatagt tggc 54
 <210> 364
 <211> 54
 <212> DNA
 40 <213> Artificial Sequence
 <220>
 <223> Synthetic Oligonucleotide
 <400> 364
 45 gtgactggag ttcagacgtg tgcttccg atcttgaaag gcttttgtg ggtt 54
 <210> 365
 <211> 54
 <212> DNA
 50 <213> Artificial Sequence
 <220>
 <223> Synthetic Oligonucleotide

<400> 365
 gtgactggag ttcagacgtg tgctctccg atctttccta gcgggaactg gaaa 54
 <210> 366
 <211> 54
 <212> DNA
 <213> Artificial Sequence
 5
 <220>
 10 <223> Synthetic Oligonucleotide

 <400> 366
 gtgactggag ttcagacgtg tgctctccg atctaggcta atggggtagg ggat 54
 15 <210> 367
 <211> 54
 <212> DNA
 <213> Artificial Sequence
 20 <220>
 <223> Synthetic Oligonucleotide

 <400> 367
 gtgactggag ttcagacgtg tgctctccg atcttgtcca tggggctga ggtg 54
 25 <210> 368
 <211> 54
 <212> DNA
 <213> Artificial Sequence
 30 <220>
 <223> Synthetic Oligonucleotide

 <400> 368
 gtgactggag ttcagacgtg tgctctccg atctcaggcc aaccttgaca actt 54
 35 <210> 369
 <211> 54
 <212> DNA
 <213> Artificial Sequence
 40 <220>
 <223> Synthetic Oligonucleotide

 <400> 369
 gtgactggag ttcagacgtg tgctctccg atctaggcagg ccaaagatgt ctcc 54
 45 <210> 370
 <211> 57
 <212> DNA
 <213> Artificial Sequence
 50 <220>
 <223> Synthetic Oligonucleotide
 55 <220>
 <223> Synthetic Oligonucleotide
 <400> 370
 gtgactggag ttcagacgtg tgctctccg atcttctgct ctggaggta ttgtcc 57

5 <210> 371
 <211> 56
 <212> DNA
 <213> Artificial Sequence
 10 <220>
 <223> Synthetic Oligonucleotide
 <400> 371
 gtgactggag ttcagacgtg tgctctccg atctggacc aatttgcac tcatgg 56

15 <210> 372
 <211> 54
 <212> DNA
 <213> Artificial Sequence
 20 <220>
 <223> Synthetic Oligonucleotide
 <400> 372
 gtgactggag ttcagacgtg tgctctccg atctggagg ctgtaaacgt cctg 54

25 <210> 373
 <211> 59
 <212> DNA
 <213> Artificial Sequence
 30 <220>
 <223> Synthetic Oligonucleotide
 <400> 373
 gtgactggag ttcagacgtg tgctctccg atcttgctat gatttgctga attactcct 59

35 <210> 374
 <211> 55
 <212> DNA
 <213> Artificial Sequence
 40 <220>
 <223> Synthetic Oligonucleotide
 <400> 374
 gtgactggag ttcagacgtg tgctctccg atctgcaatt ttgcagacca ccac 55

45 <210> 375
 <211> 54
 <212> DNA
 <213> Artificial Sequence
 50 <220>
 <223> Synthetic Oligonucleotide
 <400> 375
 gtgactggag ttcagacgtg tgctctccg atctggcagc ttgcaacctt ctg 54

55 <210> 376
 <211> 54
 <212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

5

<400> 376

gtgactggag ttcagacgtg tgctctccg atcttcatga gagttcccc aaca 54

10 <210> 377

<211> 58

<212> DNA

<213> Artificial Sequence

15 <220>

<223> Synthetic Oligonucleotide

<400> 377

gtgactggag ttcagacgtg tgctctccg atcttacttga gggggaaaaaa gtttctta 58

20 <210> 378

<211> 54

<212> DNA

<213> Artificial Sequence

25 <220>

<223> Synthetic Oligonucleotide

<400> 378

gtgactggag ttcagacgtg tgctctccg atcttggtcc ctgttgtca ttgg 54

30

<210> 379

<211> 54

<212> DNA

<213> Artificial Sequence

35 <220>

<223> Synthetic Oligonucleotide

<400> 379

40 gtgactggag ttcagacgtg tgctctccg atctaaggca gtgactgtct ggga 54

<210> 380

<211> 54

<212> DNA

45 <213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

50 <400> 380

gtgactggag ttcagacgtg tgctctccg atctcatggg tgggacacgt agtt 54

<210> 381

<211> 55

55 <212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

<400> 381

gtgactggag ttcagacgtg tgctttccg atctggctt cctggacacc ctatc 55

5

<210> 382

<211> 53

<212> DNA

<213> Artificial Sequence

10

<220>

<223> Synthetic Oligonucleotide

<400> 382

15

gtgactggag ttcagacgtg tgctttccg atcttagagcg agggagcgat gta 53

<210> 383

<211> 55

<212> DNA

20

<213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

25

<400> 383

gtgactggag ttcagacgtg tgctttccg atctttgtgg accactgctt agtgc 55

<210> 384

<211> 53

30

<212> DNA

<213> Artificial Sequence

35

<220>

<223> Synthetic Oligonucleotide

<400> 384

gtgactggag ttcagacgtg tgctttccg atctcaacta ccctgaggcc acc 53

40

<210> 385

<211> 54

<212> DNA

<213> Artificial Sequence

45

<220>

<223> Synthetic Oligonucleotide

<400> 385

gtgactggag ttcagacgtg tgctttccg atctggtcag cactcctcag cttt 54

50

<210> 386

<211> 54

<212> DNA

<213> Artificial Sequence

55

<220>

<223> Synthetic Oligonucleotide

<400> 386

gtgactggag ttcagacgtg tgctttccg atcttgagg atgcatgcc catt 54
 <210> 387
 <211> 54
 5 <212> DNA
 <213> Artificial Sequence

 <220>
 <223> Synthetic Oligonucleotide
 10 <400> 387
 gtgactggag ttcagacgtg tgctttccg atctcccagc ctcttgacc cttc 54

 <210> 388
 15 <211> 53
 <212> DNA
 <213> Artificial Sequence

 <220>
 20 <223> Synthetic Oligonucleotide

 <400> 388
 gtgactggag ttcagacgtg tgctttccg atctcccaca ccaggctgta agg 53

 25 <210> 389
 <211> 58
 <212> DNA
 <213> Artificial Sequence

 30 <220>
 <223> Synthetic Oligonucleotide

 <400> 389
 gtgactggag ttcagacgtg tgctttccg atcttagata tatgggtgtg tctgtacg 58
 35 <210> 390
 <211> 54
 <212> DNA
 <213> Artificial Sequence

 40 <220>
 <223> Synthetic Oligonucleotide

 <400> 390
 45 gtgactggag ttcagacgtg tgctttccg atctttccaa agtggctgaa ccat 54

 <210> 391
 <211> 54
 <212> DNA
 50 <213> Artificial Sequence

 <220>
 <223> Synthetic Oligonucleotide

 <400> 391
 55 gtgactggag ttcagacgtg tgctttccg atctcccaca gggctgtatgt ttca 54

 <210> 392

<211> 57

<212> DNA

<213> Artificial Sequence

5 <220>

<223> Synthetic Oligonucleotide

<400> 392

gtgactggag ttcagacgtg tgctctccg atctttgtaa tgcaacctct gtcatgc 57

10 <210> 393

<211> 54

<212> DNA

<213> Artificial Sequence

15 <220>

<223> Synthetic Oligonucleotide

<400> 393

20 gtgactggag ttcagacgtg tgctctccg atctccagct ccagcaatcc atga 54

<210> 394

<211> 55

<212> DNA

25 <213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

30 <400> 394

gtgactggag ttcagacgtg tgctctccg atctttggg aaagatagcc ctgga 55

<210> 395

<211> 54

35 <212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

40 <400> 395

gtgactggag ttcagacgtg tgctctccg atctcaatga aacagcgggg aggt 54

<210> 396

45 <211> 54

<212> DNA

<213> Artificial Sequence

<220>

50 <223> Synthetic Oligonucleotide

<400> 396

gtgactggag ttcagacgtg tgctctccg atctacaatc acgtgtccct cact

54

55 <210> 397

<211> 54

<212> DNA

<213> Artificial Sequence

5 <220>
 <223> Synthetic Oligonucleotide
 <400> 397
 gtgactggag ttcagacgtg tgcttccg attcagatc cctccggc aatg 54
 10 <210> 398
 <211> 54
 <212> DNA
 <213> Artificial Sequence
 <220>
 <223> Synthetic Oligonucleotide
 15 <400> 398
 gtgactggag ttcagacgtg tgcttccg atctgtcagg aggcaaggag gaac 54
 <210> 399
 <211> 58
 20 <212> DNA
 <213> Artificial Sequence
 <220>
 <223> Synthetic Oligonucleotide
 25 <400> 399
 gtgactggag ttcagacgtg tgcttccg atctacttcc ttcctttga gaccaagt 58
 <210> 400
 <211> 54
 <212> DNA
 <213> Artificial Sequence
 <220>
 <223> Synthetic Oligonucleotide
 35 <400> 400
 gtgactggag ttcagacgtg tgcttccg atctgcggca gattcctgggt gatt 54
 <210> 401
 <211> 54
 <212> DNA
 <213> Artificial Sequence
 45 <220>
 <223> Synthetic Oligonucleotide
 <400> 401
 gtgactggag ttcagacgtg tgcttccg atctggtcac catcagcaca gtca 54
 50 <210> 402
 <211> 56
 <212> DNA
 <213> Artificial Sequence
 55 <220>
 <223> Synthetic Oligonucleotide

1 <400> 402
 caagcagaag acggcatacg agatatatca gtgtgactgg agttcagacg tgtgct 56
 5 <210> 403
 <211> 56
 <212> DNA
 <213> **Artificial Sequence**
 10 <220>
 <223> Synthetic Oligonucleotide
 15 <400> 403
 caagcagaag acggcatacg agatttcac cggtgactgg agttcagacg tgtgct 56
 20 <210> 404
 <211> 56
 <212> DNA
 <213> **Artificial Sequence**
 25 <220>
 <223> Synthetic Oligonucleotide
 30 <400> 404
 caagcagaag acggcatacg agatccactc atgtgactgg agttcagacg tgtgct 56
 35 <210> 405
 <211> 56
 <212> DNA
 <213> **Artificial Sequence**
 40 <220>
 <223> Synthetic Oligonucleotide
 45 <400> 405
 caagcagaag acggcatacg agattacgta cggtgactgg agttcagacg tgtgct 56
 50 <210> 406
 <211> 56
 <212> DNA
 <213> **Artificial Sequence**
 55 <220>
 <223> Synthetic Oligonucleotide
 55 <400> 406
 caagcagaag acggcatacg agatcgaaac tcgtgactgg agttcagacg tgtgct 56
 60 <210> 407
 <211> 56
 <212> DNA
 <213> Artificial Sequence
 65 <220>
 <223> Synthetic Oligonucleotide
 65 <400> 407
 caagcagaag acggcatacg agatatcagt atgtgactgg agttcagacg tgtgct 56

<210> 408
 <211> 57
 <212> DNA
 <213> Artificial Sequence

5

<220>
 <223> Synthetic Oligonucleotide

<400> 408

aatgatacgg cgaccaccga gatctacaca ttactcgaca ctcttccct acacgac 57

<210> 409

<211> 57

<212> DNA

15 <213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

20 <400> 409

aatgatacgg cgaccaccga gatctacact ccggagaaca ctcttccct acacgac 57

<210> 410

<211> 57

25 <212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic Oligonucleotide

30

<400> 410

aatgatacgg cgaccaccga gatctacacc gctcattaca ctcttccct acacgac 57

35 Claims

1. A method for identifying a target site of a nuclease, the method comprising:

40 (a) providing a nuclease that cuts a double-stranded nucleic acid target site, wherein cutting of the target site results in cut nucleic acid strands comprising a 5' phosphate moiety;

(b) contacting the nuclease of step (a) with a library of candidate nucleic acid molecules, wherein each nucleic acid molecule comprises a concatemer of a sequence comprising a candidate nuclease target site and a constant insert sequence of at least 10 and not more than 80 nucleotides, under conditions suitable for the nuclease to cut a candidate nucleic acid molecule comprising a target site of the nuclease; and

45 (c) identifying nuclease target sites cut by the nuclease in step (b) by determining the sequence of an uncut nuclease target site on the nucleic acid strand that was cut by the nuclease in step (b), wherein the determining of step (c) comprises (i) ligating a first nucleic acid adapter to the 5' end of a nucleic acid strand that was cut by the nuclease in step (b) via 5'-phosphate-dependent ligation, and (ii) amplifying a fragment of the concatemer cut by the nuclease that comprises an uncut target site via a PCR reaction using a PCR primer that hybridizes 50 with the adapter and a PCR primer that hybridizes with the constant insert sequence.

2. The method of claim 1, wherein the nucleic acid adapter is provided in double-stranded form.

3. The method of claim 1 or claim 2, wherein the 5'-phosphate-dependent ligation is a blunt end ligation.

55 4. The method of any one of claims 1-3, wherein the method comprises filling in the 5'-overhang before ligating the first nucleic acid adapter to the nucleic acid strand that was cut by the nuclease.

5. The method of any one of claims 1-4, wherein the determining of step (c) further comprises sequencing the fragment obtained by the PCR reaction of step c(ii); and/or

wherein the library of candidate nucleic acid molecules comprises at least 10^8 , at least 10^9 , at least 10^{10} , at least 10^{11} , or at least 10^{12} different candidate nuclease cleavage sites; and/or
 wherein the nuclease is a therapeutic nuclease which cuts a specific nuclease target site in a gene associated with a disease, optionally wherein the method further comprises determining a maximum concentration of the therapeutic nuclease at which the nuclease cuts the specific nuclease target site, and does not cut more than 10, more than 5, more than 4, more than 3, more than 2, more than 1, or no additional nuclease target sites.

10. 6. The method of any one of claims 1-5, wherein the nuclease is an RNA-programmable nuclease that forms a complex with an RNA molecule, and wherein the nuclease:RNA complex specifically binds a nucleic acid sequence complementary to the sequence of the RNA molecule, optionally wherein the RNA molecule is a single-guide RNA (sgRNA), and optionally wherein the nuclease is a Cas9 nuclease.

15. 7. The method of any preceding claim, wherein the nuclease is a Cas9 nuclease in complex with a sgRNA.

20. 8. The method of any preceding claim, wherein the target site comprises a [sgRNA-complementary sequence]-[protospacer adjacent motif (PAM)] structure, and the nuclease cuts the target site within the sgRNA-complementary sequence.

9. 9. The method of any preceding claim, wherein the constant insert sequence comprises not more than 70, not more than 60, or not more than 55 nucleotides.

25. 10. A method of selecting a nuclease that specifically cuts a consensus target site from a plurality of nucleases, the method comprising:

(a) providing a plurality of candidate nucleases that cut the same consensus sequence;
 30 (b) for each of the candidate nucleases of step (a), identifying a nuclease target site cleaved by the candidate nuclease that differs from the consensus target site using the method of any one of claims 1-9; and
 (c) selecting a nuclease based on the nuclease target site(s) identified in step (b) which cleaves the consensus target site with the highest specificity.

35. 11. The method of claim 10, wherein the nuclease target site comprises a [sgRNA-complementary sequence]-[protospacer adjacent motif (PAM)] structure, and the nuclease cuts the target site within the sgRNA-complementary sequence, optionally wherein the sgRNA-complementary sequence comprises 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, or 30 nucleotides.

40. 12. The method of claim 10 or claim 11, wherein the nuclease that cleaves the consensus target site with the highest specificity is one in which

(i) the nuclease cleaves the lowest number of target sites that differ from the consensus site, or
 (ii) the nuclease cleaves the lowest number of target sites that are different from the consensus site in the context of a target genome.

45. 13. The method of any one of claims 10-12, wherein the nuclease that cleaves the consensus target site with the highest specificity is one which:

(i) does not cleave any target site other than the consensus target site, or
 (ii) does not cleave any target site other than the consensus target site within the genome of a subject at a therapeutically effective concentration of the nuclease.

55. 14. The method of any one of claims 10-13, wherein the nuclease selected in step (c) cleaves a genome *in vitro*, optionally wherein the genome is a vertebrate, mammalian, human, non-human primate, rodent, mouse, rat, hamster, goat, sheep, cattle, dog, cat, reptile, amphibian, fish, nematode, insect, or fly genome, and/or wherein the genome is optionally within an *in vitro* living cell.

15. 15. The method of any one of claims 10-14, wherein the consensus target site is within an allele that is associated with

a disease or disorder.

16. The method of claim 15, wherein the disease is HIV/AIDS, and the allele is optionally a CCR5 allele; or wherein the disease is a proliferative disease, and the allele is optionally a VEGFA allele.

5

Patentansprüche

1. Verfahren zur Identifizierung einer Zielstelle einer Nuklease, wobei das Verfahren Folgendes umfasst:

10

- (a) Bereitstellen einer Nuklease, die eine Doppelstrang-Nukleinsäure-Zielstelle schneidet, wobei Schneiden der Zielstelle geschnittene Nukleinsäurestränge ergibt, die eine 5'-Phosphatgruppierung umfassen;
- (b) In-Kontakt-Bringen der Nuklease aus Schritt (a) mit einer Bibliothek von Kandidaten-Nukleinsäuremolekülen, wobei jedes Nukleinsäuremolekül ein Concatemer einer Sequenz umfasst, die eine Kandidaten-Nuklease-Zielstelle und eine konstante Insertionssequenz von wenigstens 10 und nicht mehr als 80 Nukleotiden umfasst, unter geeigneten Bedingungen, so dass die Nuklease ein Kandidaten-Nukleinsäuremolekül, das eine Zielstelle der Nuklease umfasst, schneidet; und
- (c) Identifizieren von in Schritt (b) von der Nuklease geschnittenen Nuklease-Zielstellen durch Bestimmen der Sequenz einer ungeschnittenen Nuklease-Zielstelle auf dem Nukleinsäurestrang, der von der Nuklease in Schritt (b) geschnitten wurde, wobei das Bestimmen unter Schritt (c) (i) Ligieren eines ersten Nukleinsäureadapters an das 5'-Ende eines Nukleinsäurestrangs, der von der Nuklease in Schritt (b) geschnitten wurde, über 5'-phosphatabhängige Ligation und (ii) Amplifizieren eines Fragments des von der Nuklease geschnittenen Concatemers, das eine ungeschnittene Zielstelle umfasst, über eine PCR-Reaktion unter Verwendung eines PCR-Primers, der mit dem Adapter hybridisiert, und eines PCR-Primers, der mit der konstanten Insertionssequenz hybridisiert, umfasst.

15

2. Verfahren nach Anspruch 1, wobei der Nukleinsäureadapter in Doppelstrangform bereitgestellt wird.

20

3. Verfahren nach Anspruch 1 oder Anspruch 2, wobei es sich bei der 5'-phosphat-abhängigen Ligation um eine Stumpfe-Enden-Ligation handelt.
4. Verfahren nach einem der Ansprüche 1-3, wobei das Verfahren Auffüllen des 5'-Überhangs vor Ligieren des ersten Nukleinsäureadapters an den Nukleinsäurestrang, der von der Nuklease geschnitten wurde, umfasst.

25

5. Verfahren nach einem der Ansprüche 1-4, wobei das Bestimmen unter Schritt (c) ferner Sequenzieren des über die PCR-Reaktion unter Schritt c(ii) erhaltenen Fragments umfasst; und/oder

30

wobei die Bibliothek von Kandidaten-Nukleinsäuremolekülen wenigstens 10^8 , wenigstens 10^9 , wenigstens 10^{10} , wenigstens 10^{11} oder wenigstens 10^{12} unterschiedliche Kandidaten-Nuklease-Spaltstellen umfasst; und/oder wobei es sich bei der Nuklease um eine therapeutische Nuklease handelt, die eine spezifische Nuklease-Zielstelle in einem mit einer Krankheit assoziierten Gen schneidet, gegebenenfalls wobei das Verfahren ferner Bestimmen einer Maximalkonzentration der therapeutischen Nuklease umfasst, bei der die Nuklease die spezifische Nuklease-Zielstelle schneidet und nicht mehr als 10, mehr als 5, mehr als 4, mehr als 3, mehr als 2, mehr als 1 oder keine zusätzlichen Nuklease-Zielstellen schneidet.

35

6. Verfahren nach einem der Ansprüche 1-5, wobei es sich bei der Nuklease um eine RNA-programmierbare Nuklease handelt, die einen Komplex mit einem RNA-Molekül bildet, und wobei der Nuklease:RNA-Komplex eine Nukleinsäuresequenz, die komplementär zur Sequenz des RNA-Moleküls ist, spezifisch bindet, gegebenenfalls wobei es sich bei dem RNA-Molekül um eine Single-guide-RNA (sgRNA) handelt und gegebenenfalls wobei es sich bei der Nuklease um eine Cas9-Nuklease handelt.

40

7. Verfahren nach einem vorhergehenden Anspruch, wobei es sich bei der Nuklease um eine Cas9-Nuklease im Komplex mit einer sgRNA handelt.

45

8. Verfahren nach einem vorhergehenden Anspruch, wobei die Zielstelle eine [sgRNA-Komplementärsequenz]-[PAM(Protospacer Adjacent Motif)]-Struktur umfasst und die Nuklease die Zielstelle innerhalb der sgRNA-Komplementärsequenz schneidet.

9. Verfahren nach einem vorhergehenden Anspruch, wobei die konstante Insertionssequenz nicht mehr als 70, nicht mehr als 60 oder nicht mehr als 55 Nukleotide umfasst.
- 5 10. Verfahren zum Auswählen einer Nuklease, die eine Konsensus-Zielstelle spezifisch schneidet, aus mehreren Nukreasen, wobei das Verfahren Folgendes umfasst:
- (a) Bereitstellen mehrerer Kandidaten-Nukreasen, die die gleiche Konsensussequenz schneiden;
 - (b) für jede der Kandidaten-Nukreasen aus Schritt (a) Identifizieren einer von der Kandidaten-Nuklease gespaltenen Nuklease-Zielstelle, die von der Konsensus-Zielstelle verschieden ist, unter Verwendung des Verfahrens nach einem der Ansprüche 1-9; und
 - (c) Auswählen einer Nuklease anhand der in Schritt (b) identifizierten Nuklease-Zielstelle(n), die die Konsensus-Zielstelle mit der höchsten Spezifität spaltet.
- 10 11. Verfahren nach Anspruch 10, wobei die Nuklease-Zielstelle eine [sgRNA-Komplementärsequenz]-[PAM(Proto-spacer Adjacent Motif)]-Struktur umfasst und die Nuklease die Zielstelle innerhalb der sgRNA-Komplementärsequenz schneidet, gegebenenfalls wobei die sgRNA-Komplementärsequenz 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29 oder 30 Nukleotide umfasst.
- 15 12. Verfahren nach Anspruch 10 oder Anspruch 11, wobei es sich bei der Nuklease, die die Konsensus-Zielstelle mit der höchsten Spezifität spaltet, um eine handelt, bei der
- (i) die Nuklease die geringste Anzahl von Zielstellen, die von der Konsensusstelle verschieden sind, spaltet oder
 - (ii) die Nuklease die geringste Anzahl von Zielstellen, die sich von der Konsensusstelle im Kontext eines Zielgenoms unterscheiden, spaltet.
- 20 25 13. Verfahren nach einem der Ansprüche 10-12, wobei es sich bei der Nuklease, die die Konsensus-Zielstelle mit der höchsten Spezifität spaltet, um eine handelt, die:
- (i) keine andere Zielstelle als die Konsensus-Zielstelle spaltet oder
 - (ii) keine andere Zielstelle als die Konsensus-Zielstelle innerhalb des Genoms eines Individuums bei einer therapeutisch wirksamen Konzentration der Nuklease spaltet.
- 30 35 14. Verfahren nach einem der Ansprüche 10-13, wobei die in Schritt (c) ausgewählte Nuklease ein Genom *in vitro* spaltet, gegebenenfalls wobei es sich bei dem Genom um das Genom eines Wirbeltiers, Säugers, Menschen, nichtmenschlichen Primaten, Nagers, einer Maus, Ratte, eines Hamsters, einer Ziege, eines Schafs, Rinds, Hunds, einer Katze, eines Reptils, Amphibiums, Fischs, Nematoden, Insekts oder einer Fliege handelt und/oder wobei sich das Genom gegebenenfalls in einer In-vitro-Lebendzelle befindet.
- 40 45 15. Verfahren nach einem der Ansprüche 10-14, wobei sich die Konsensus-Zielstelle in einem Allel befindet, das mit einer Krankheit oder Störung assoziiert ist.
16. Verfahren nach Anspruch 15, wobei es sich bei der Krankheit um HIV/AIDS und bei dem Allel gegebenenfalls um ein CCR5-Allel handelt; oder wobei es sich bei der Krankheit um eine Proliferationskrankheit und bei dem Allel gegebenenfalls um ein VEGFA-Allel handelt.

Revendications

1. Procédé d'identification d'un site cible d'une nucléase, le procédé comprenant :
- 50
- (a) la fourniture d'une nucléase qui coupe un site cible d'acide nucléique double brin, dans lequel une coupure du site cible conduit à des brins d'acide nucléique coupés comprenant une fraction de phosphate en 5' ;
 - (b) la mise en contact de la nucléase de l'étape (a) avec une banque de molécules d'acide nucléique candidates, dans lequel chaque molécule d'acide nucléique comprend un concatémère d'une séquence comprenant un site cible de nucléase candidat et une séquence d'insert constante d'au moins 10 et pas plus de 80 nucléotides, dans des conditions adaptées pour la nucléase pour couper une molécule d'acide nucléique candidate comprenant un site cible de la nucléase ; et
 - (c) l'identification de sites cibles de nucléase coupés par la nucléase dans l'étape (b) par détermination de la

5 séquence d'un site cible de nucléase non coupé sur le brin d'acide nucléique qui a été coupé par la nucléase dans l'étape (b), dans lequel la détermination de l'étape (c) comprend (i) la ligature d'un premier adaptateur d'acide nucléique à l'extrémité 5' d'un brin d'acide nucléique qui a été coupé par la nucléase dans l'étape (b) par ligature dépendante du 5'-phosphate, et (ii) l'amplification d'un fragment du concatémère coupé par la nucléase qui comprend un site cible non coupé par l'intermédiaire d'une réaction PCR en utilisant une amorce de PCR qui s'hybride avec l'adaptateur et une amorce de PCR qui s'hybride avec la séquence d'insert constante.

2. Procédé selon la revendication 1, dans lequel l'adaptateur d'acide nucléique est fourni sous forme de double brin.
- 10 3. Procédé selon la revendication 1 ou la revendication 2, dans lequel la ligature dépendante du 5'-phosphate est une ligature à extrémités franches.
- 15 4. Procédé selon l'une quelconque des revendications 1 à 3, où le procédé comprend le remplissage de l'extrémité saillante en 5' avant la ligature du premier adaptateur d'acide nucléique au brin d'acide nucléique qui a été coupé par la nucléase.
5. Procédé selon l'une quelconque des revendications 1 à 4, dans lequel la détermination de l'étape (c) comprend en outre le séquençage du fragment obtenu par la réaction PCR de l'étape c(ii) ; et/ou
 - 20 dans lequel la banque de molécules d'acide nucléique candidates comprend au moins 10^8 , au moins 10^9 , au moins 10^{10} , au moins 10^{11} , ou au moins 10^{12} sites de clivage de nucléase candidats différents ; et/ou dans lequel la nucléase est une nucléase thérapeutique qui coupe un site cible de nucléase spécifique dans un gène associé à une maladie, facultativement où le procédé comprend en outre la détermination d'une concentration maximale de la nucléase thérapeutique à laquelle la nucléase coupe le site cible de nucléase spécifique, et ne coupe pas plus de 10, plus de 5, plus de 4, plus de 3, plus de 2, plus de 1, ou aucun, sites cibles de nucléase supplémentaires.
- 30 6. Procédé selon l'une quelconque des revendications 1 à 5, dans lequel la nucléase est une nucléase programmable par ARN qui forme un complexe avec une molécule d'ARN, et dans lequel le complexe nucléase:ARN se lie spécifiquement à une séquence d'acide nucléique complémentaire de la séquence de la molécule d'ARN, facultativement dans lequel la molécule d'ARN est un ARN guide unique (ARNsg), et facultativement dans lequel la nucléase est une nucléase Cas9.
- 35 7. Procédé selon l'une quelconque des revendications précédentes, dans lequel la nucléase est une nucléase Cas9 complexée à un ARNsg.
- 40 8. Procédé selon l'une quelconque des revendications précédentes, dans lequel le site cible comprend une structure [séquence complémentaire d'ARNsg]-[motif adjacent de protoespaceur (PAM)], et la nucléase coupe le site cible dans la séquence complémentaire d'ARNsg.
9. Procédé selon l'une quelconque des revendications précédentes, dans lequel la séquence d'insert constante ne comprend pas plus de 70, pas plus de 60, ou pas plus de 55 nucléotides.
- 45 10. Procédé de sélection d'une nucléase qui coupe spécifiquement un site cible consensus parmi une pluralité de nucléases, le procédé comprenant :
 - (a) la fourniture d'une pluralité de nucléases candidates qui coupent la même séquence consensus ;
 - 50 (b) pour chacune des nucléases candidates de l'étape (a), l'identification d'un site cible de nucléase clivé par la nucléase candidate qui diffère du site cible consensus au moyen du procédé selon l'une quelconque des revendications 1 à 9 ; et
 - (c) la sélection d'une nucléase sur la base du ou des sites cibles de nucléase identifiés dans l'étape (b) qui clive le site cible consensus avec la plus haute spécificité.
- 55 11. Procédé selon la revendication 10, dans lequel le site cible de nucléase comprend une structure [séquence complémentaire d'ARNsg]-[motif adjacent de protoespaceur (PAM)], et la nucléase coupe le site cible dans la séquence complémentaire d'ARNsg, facultativement dans lequel la séquence complémentaire d'ARNsg comprend 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, ou 30 nucléotides.

12. Procédé selon la revendication 10 ou la revendication 11, dans lequel la nucléase qui clive le site cible consensus avec la plus haute spécificité est une nucléase **caractérisée en ce que**

- 5 (i) la nucléase clive le plus petit nombre de sites cibles qui diffèrent du site consensus, ou
(ii) la nucléase clive le plus petit nombre de sites cibles qui sont différents du site consensus dans le contexte d'un génome cible.

13. Procédé selon l'une quelconque des revendications 10 à 12, dans lequel la nucléase qui clive le site cible consensus avec la plus haute spécificité est une nucléase qui :

- 10 (i) ne clive aucun site cible autre que le site cible consensus, ou
(ii) ne clive aucun site cible autre que le site cible consensus dans le génome d'un sujet à une concentration thérapeutiquement efficace de la nucléase.

15 14. Procédé selon l'une quelconque des revendications 10 à 13, dans lequel la nucléase sélectionnée dans l'étape (c) clive un génome *in vitro*, facultativement dans lequel le génome est un génome de vertébré, mammifère, humain, primate non humain, rongeur, souris, rat, hamster, chèvre, mouton, bovin, chien, chat, reptile, amphibiens, poisson, nématode, insecte ou mouche, et/ou dans lequel le génome est facultativement dans une cellule vivante *in vitro*.

20 15. Procédé selon l'une quelconque des revendications 10 à 14, dans lequel le site cible consensus est dans un allèle qui est associé à une maladie ou un trouble.

25 16. Procédé selon la revendication 15, dans lequel la maladie est le VIH/SIDA, et l'allèle est facultativement un allèle de CCR5 ; ou dans lequel la maladie est une maladie proliférative, et l'allèle est facultativement un allèle de VEGFA.

30

35

40

45

50

55

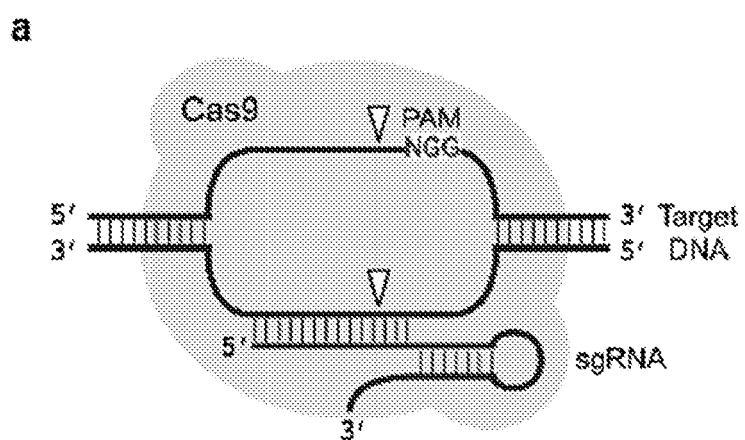
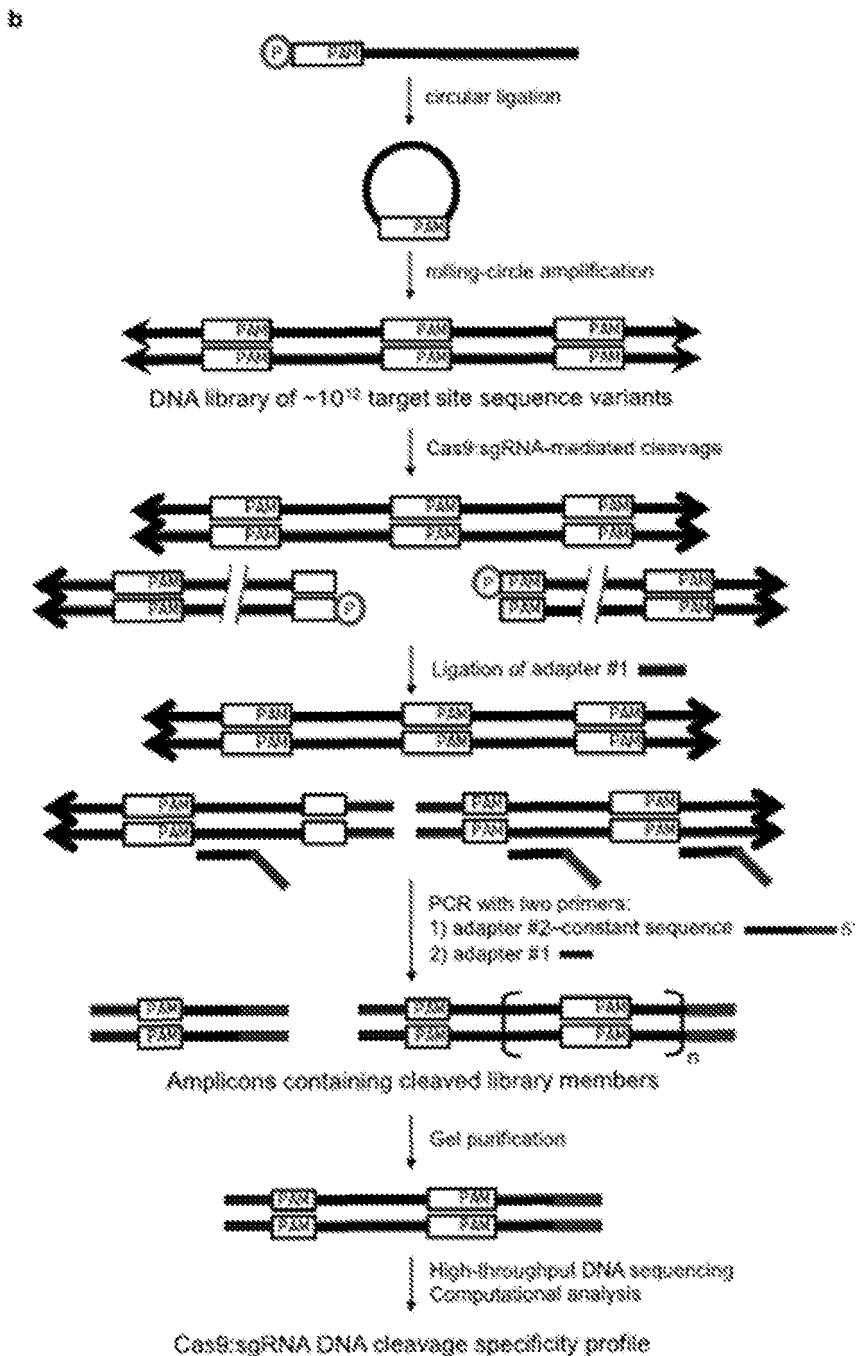
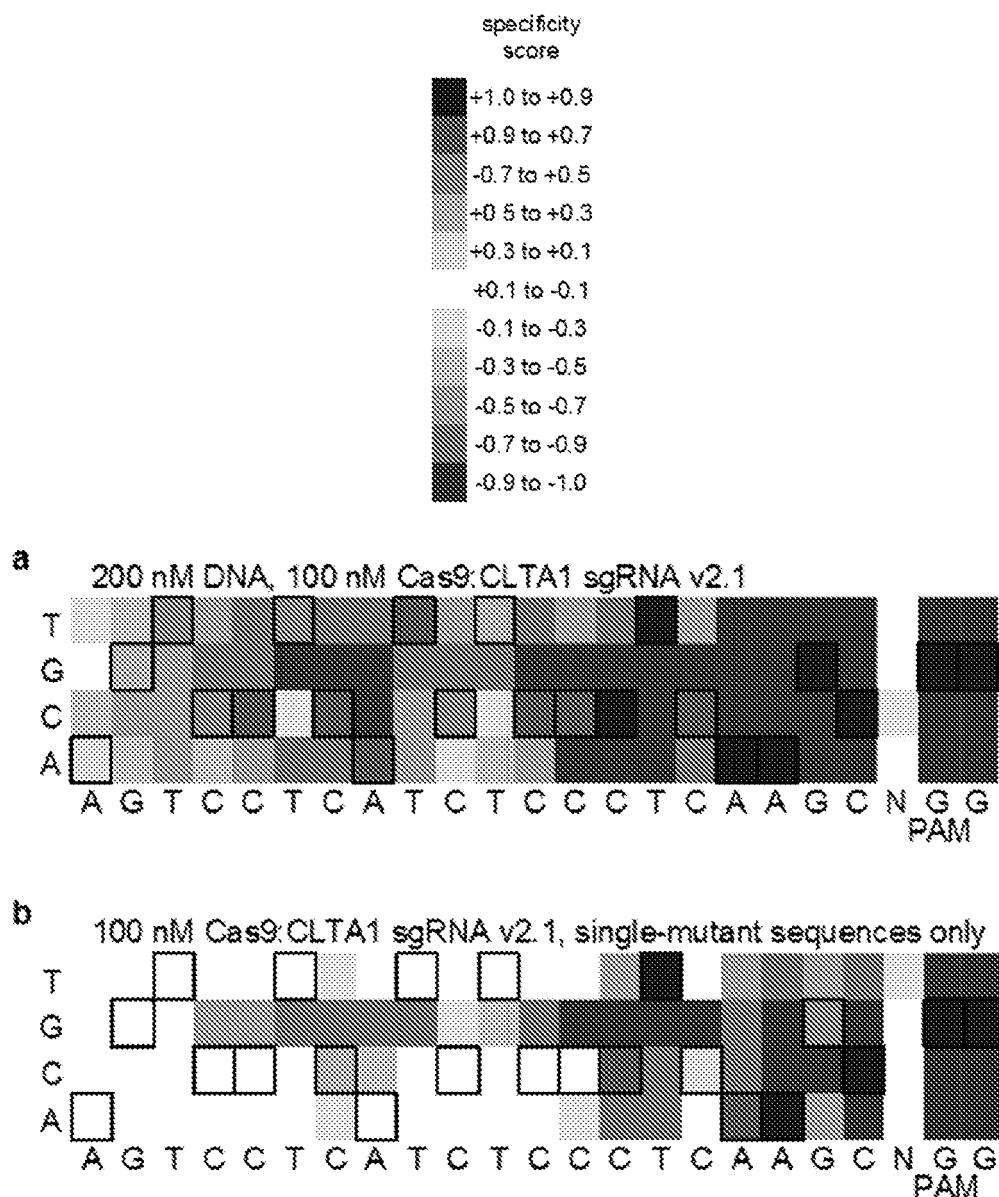
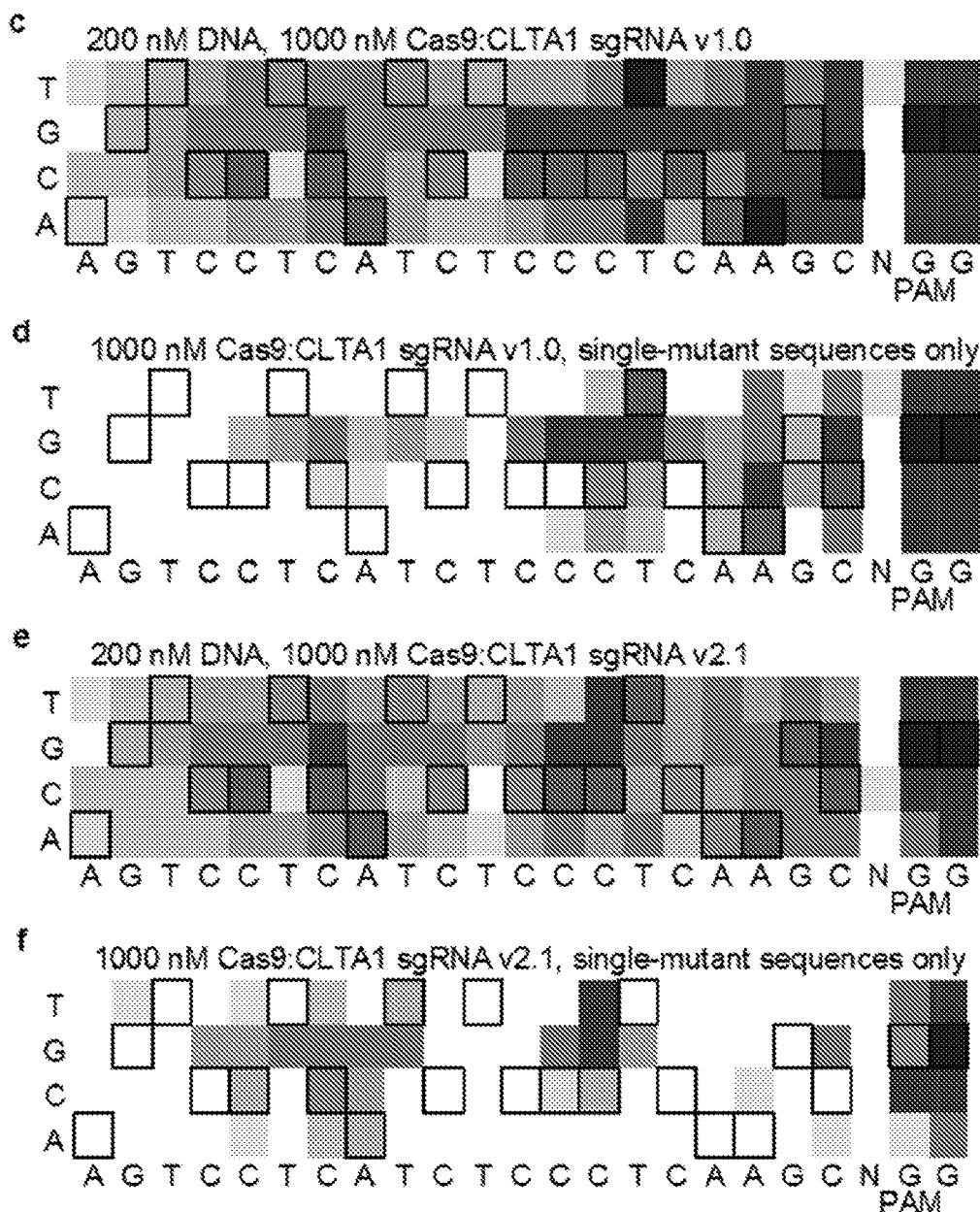
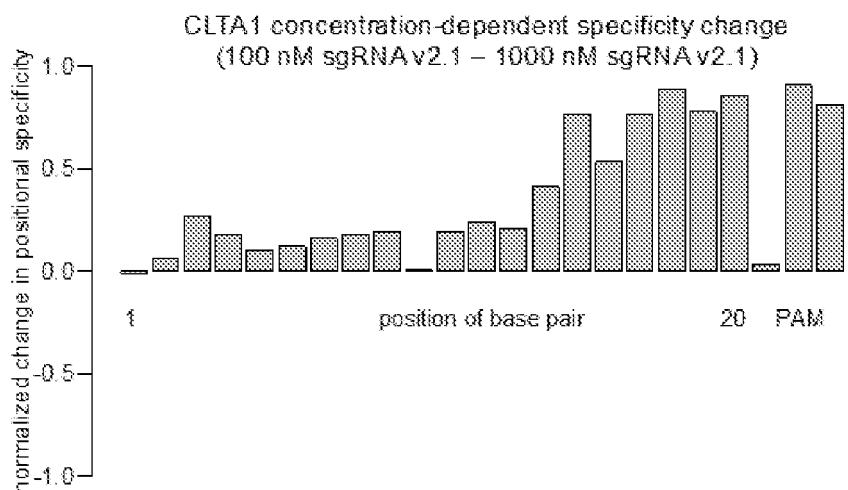
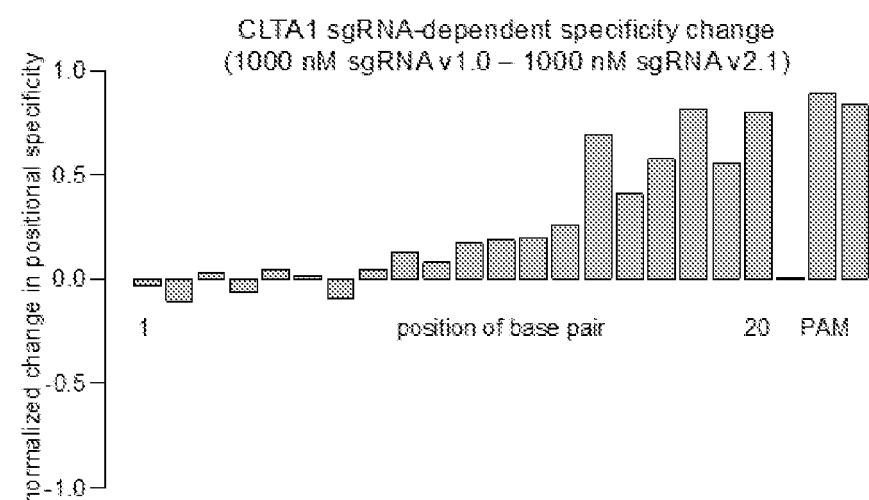


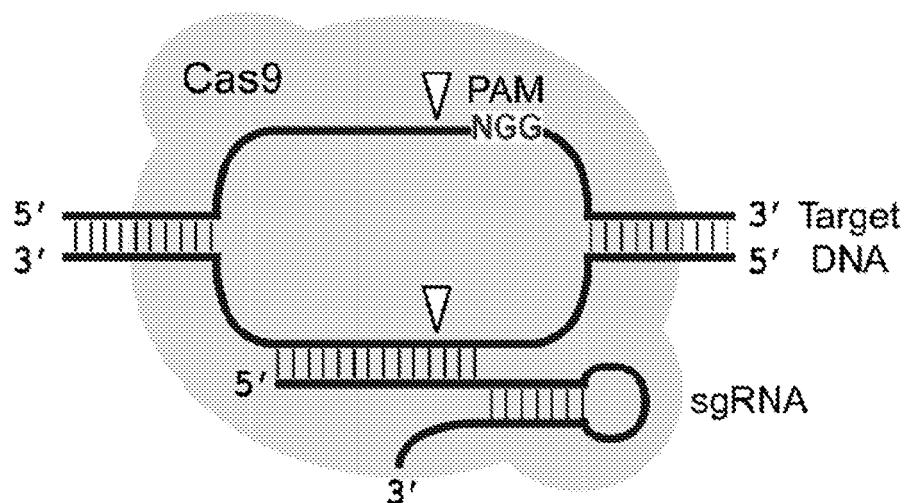
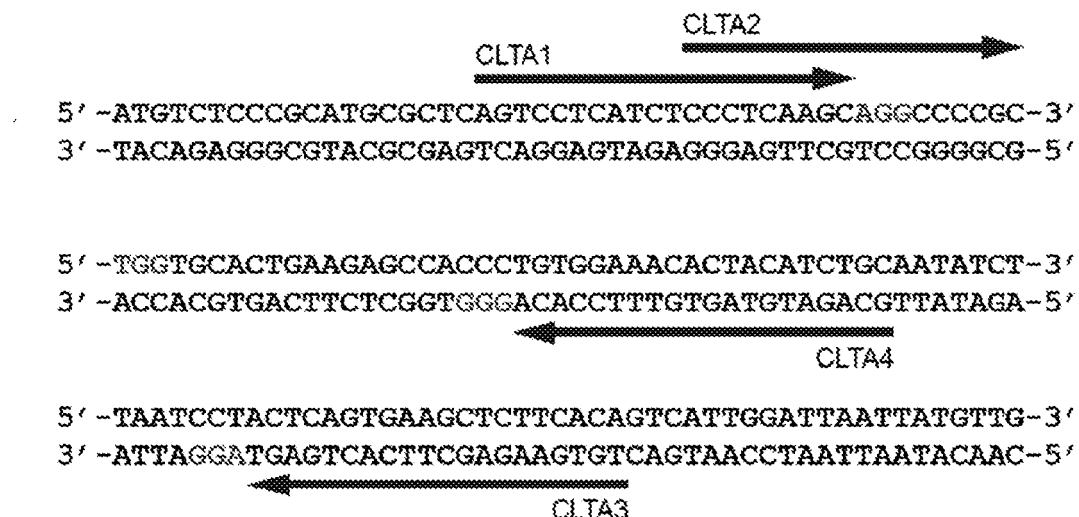
FIG. 1A

**FIG. 1B**

**FIG. 2A-B**

**FIG. 2C-F**

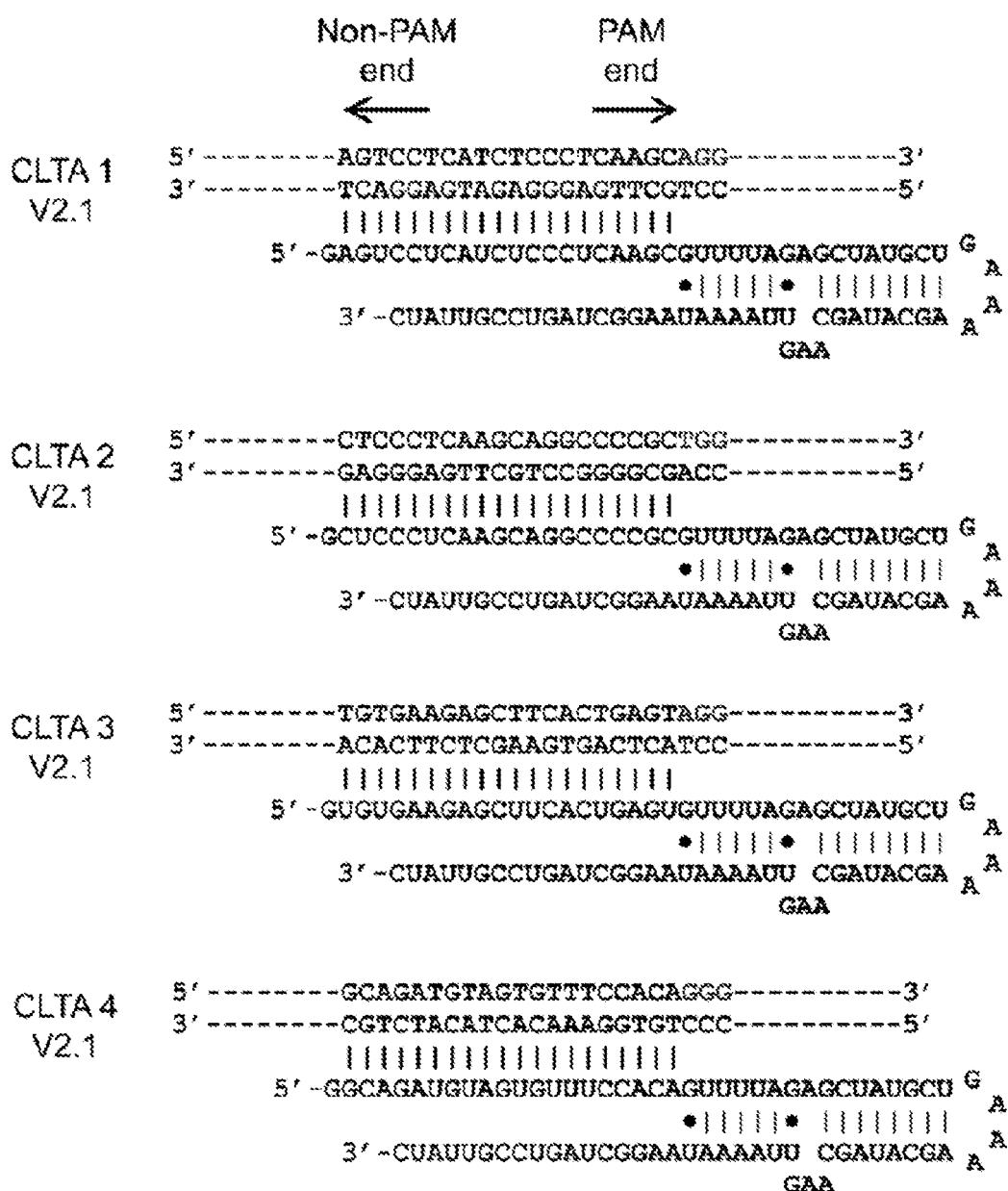
g**h****FIG. 2G-H**

a**b****FIG. 3A-B**

C

	Non-PAM end	PAM end
	←	→
CLTA 1 V1.0	5' -----AGTCCTCATCTCCCTCAAGCAGG-----3' 3' -----TCAGGGAGTAGAGGGAGTTCGTCC-----5' 5' -GAGUCCUCAUCCCCUCAAGCGUUUUAGACCUA G • • A 3' -GCCUGAUCGGAAUAAAAUU CGAU A GAA	
CLTA 2 V1.0	5' -----CTCCCTCAAGCAGGCCCCCGCTGC-----3' 3' -----GAGGGAGTTCGTCCGGGGCGACCC-----5' 5' -GCTUCCCUCUAGCAGGCCCCCGCGUUUUAGACCUA G • • A 3' -GCCUGAUCGGAAUAAAAUU CGAU A GAA	
CLTA 3 V1.0	5' -----TGTGAAGAGCTTCACTGACTGAGG-----3' 3' -----ACACTTCTCGAAGTGACTCATCC-----5' 5' -GUGUGAAGAGCUUCACUGAGUGUUUUAGACCUA G • • A 3' -GCCUGAUCGGAAUAAAAUU CGAU A GAA	
CLTA 4 V1.0	5' -----GCAGATGTAGTGTTCACAGGG-----3' 3' -----CGTCTACATCACAAAGGTGTCCC-----5' 5' -GGCAGAUGUAUGUUUCCACAGUUUUAGACCUA G • • A 3' -GCCUGAUCGGAAUAAAAUU CGAU A GAA	

FIG. 3C

d**FIG. 3D**

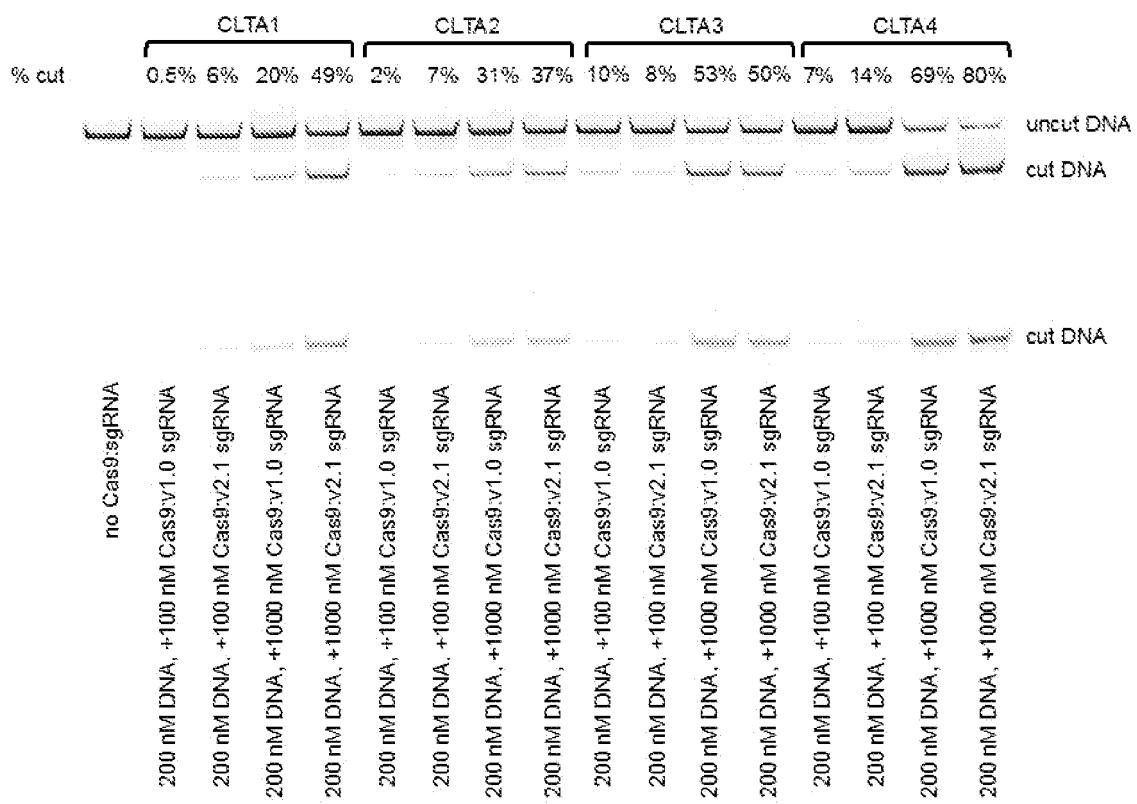
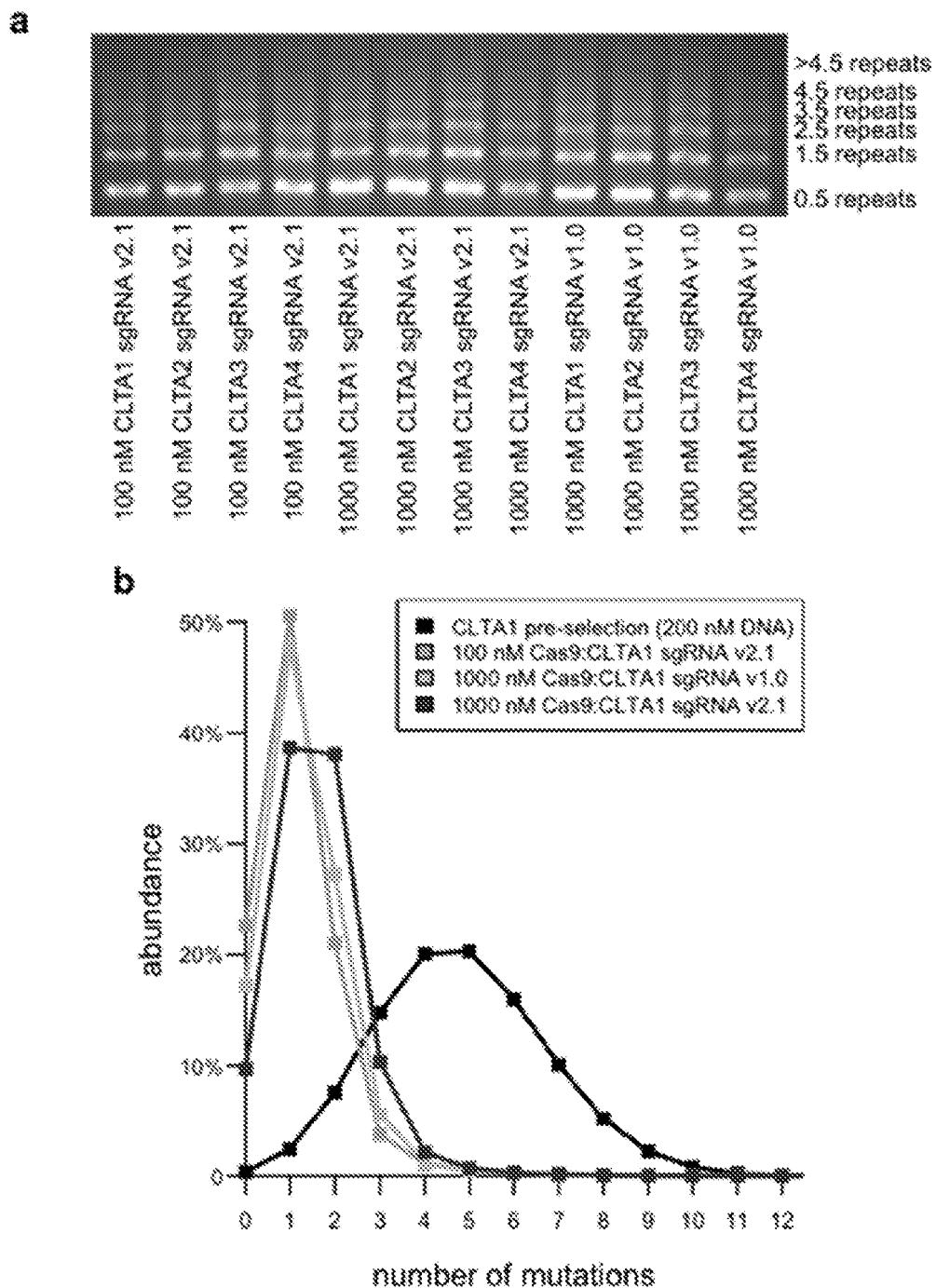
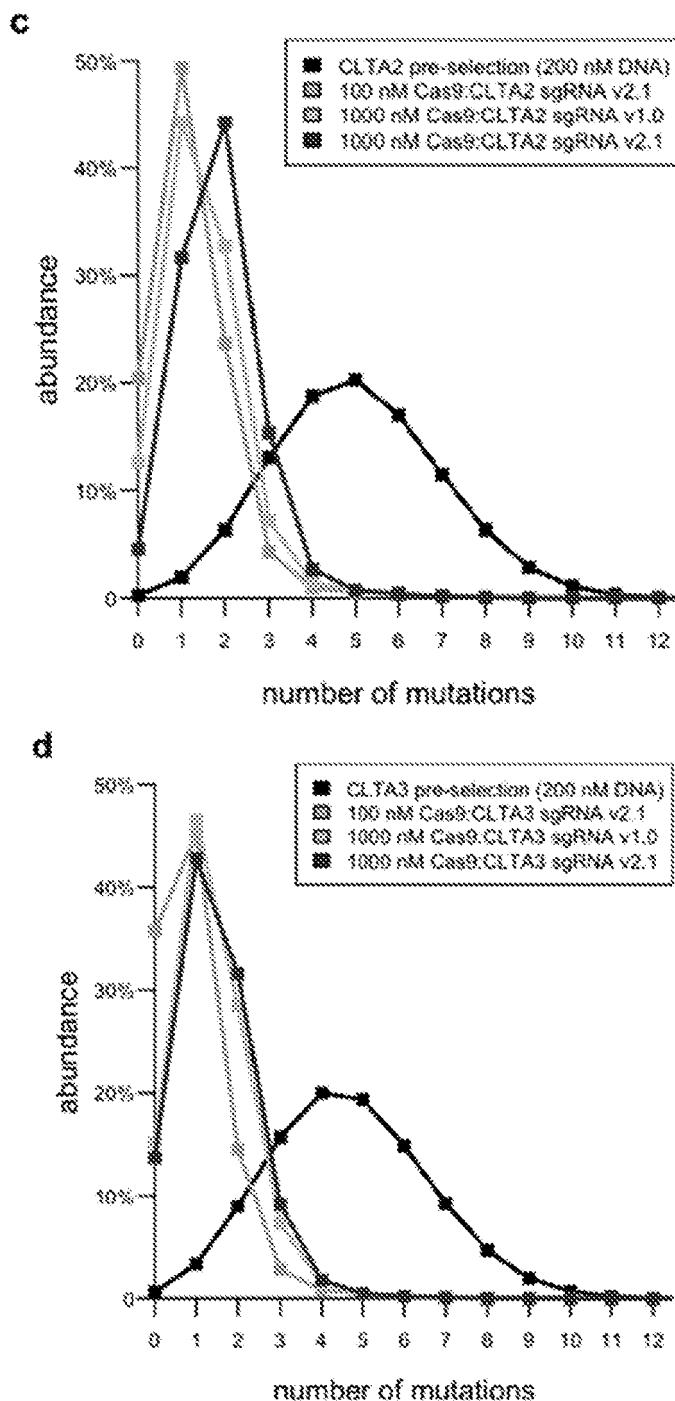


FIG. 4

**FIG. 5A-B**

**FIG. 5C-D**

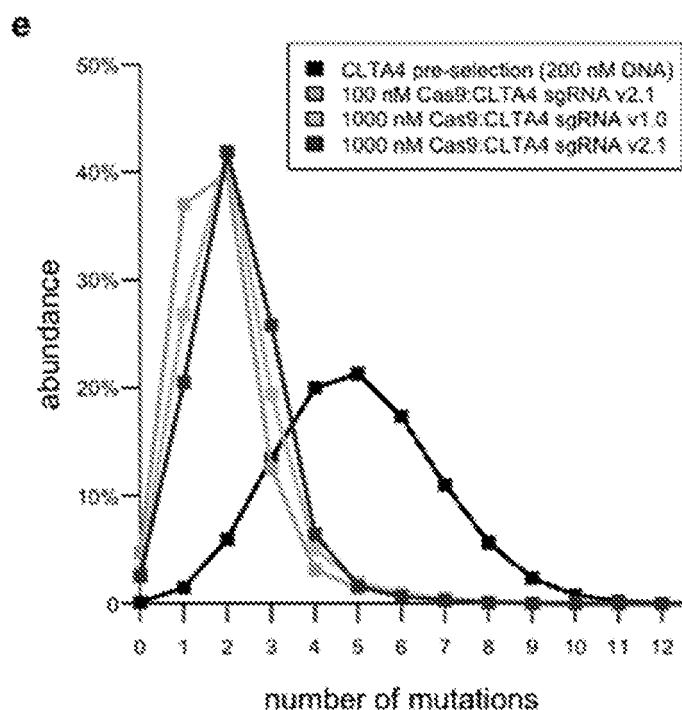
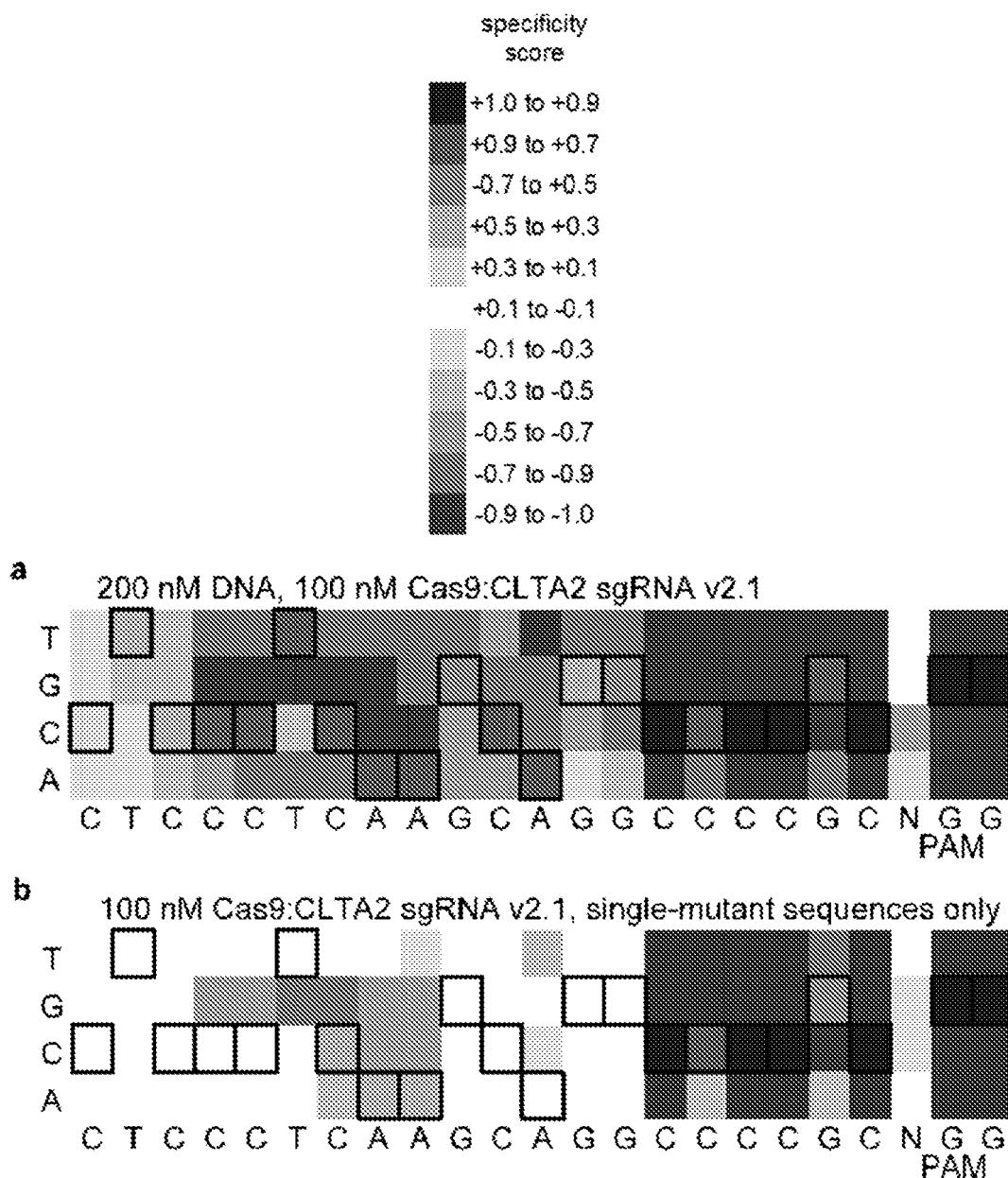


FIG. 5E

**FIG. 6A-B**

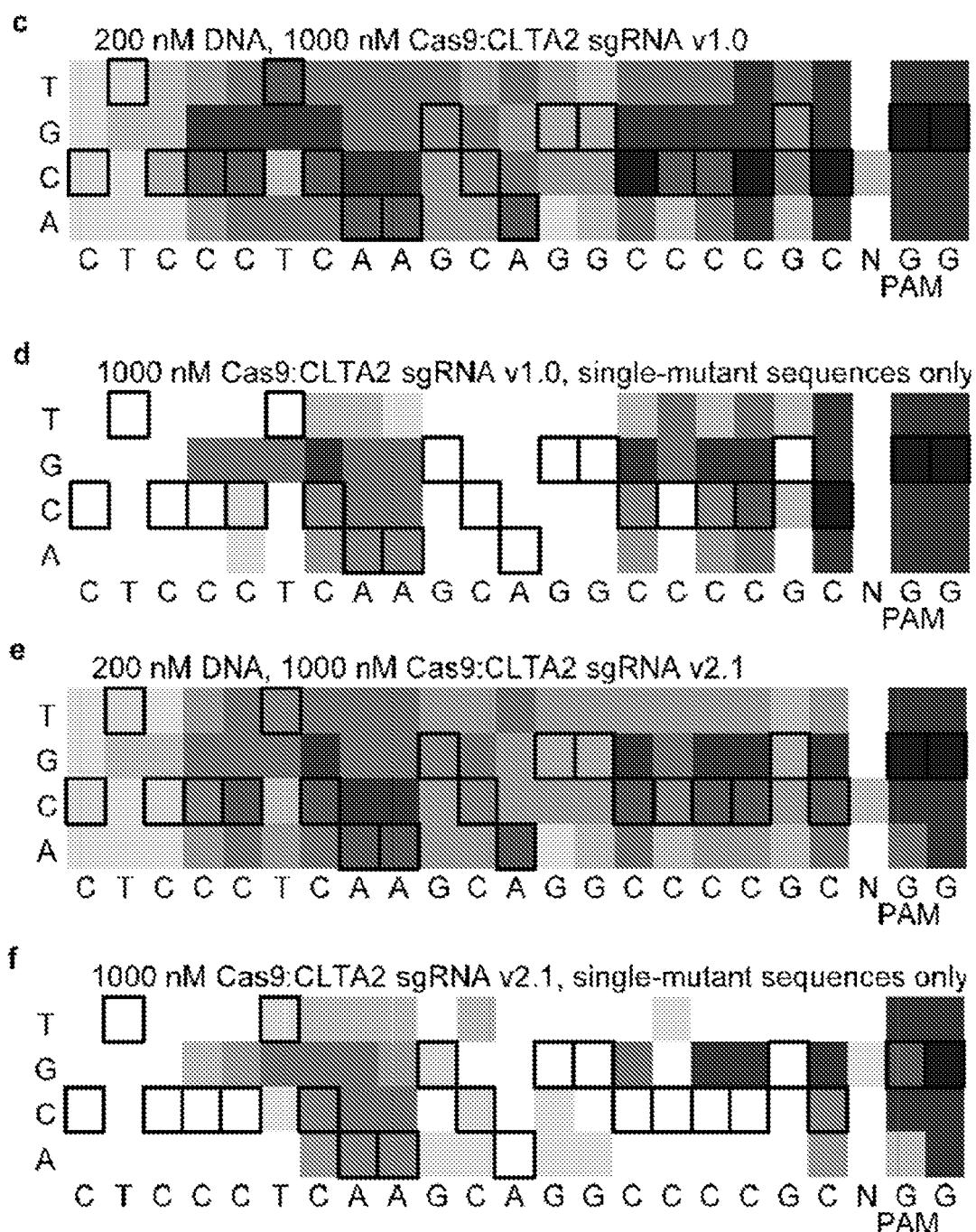
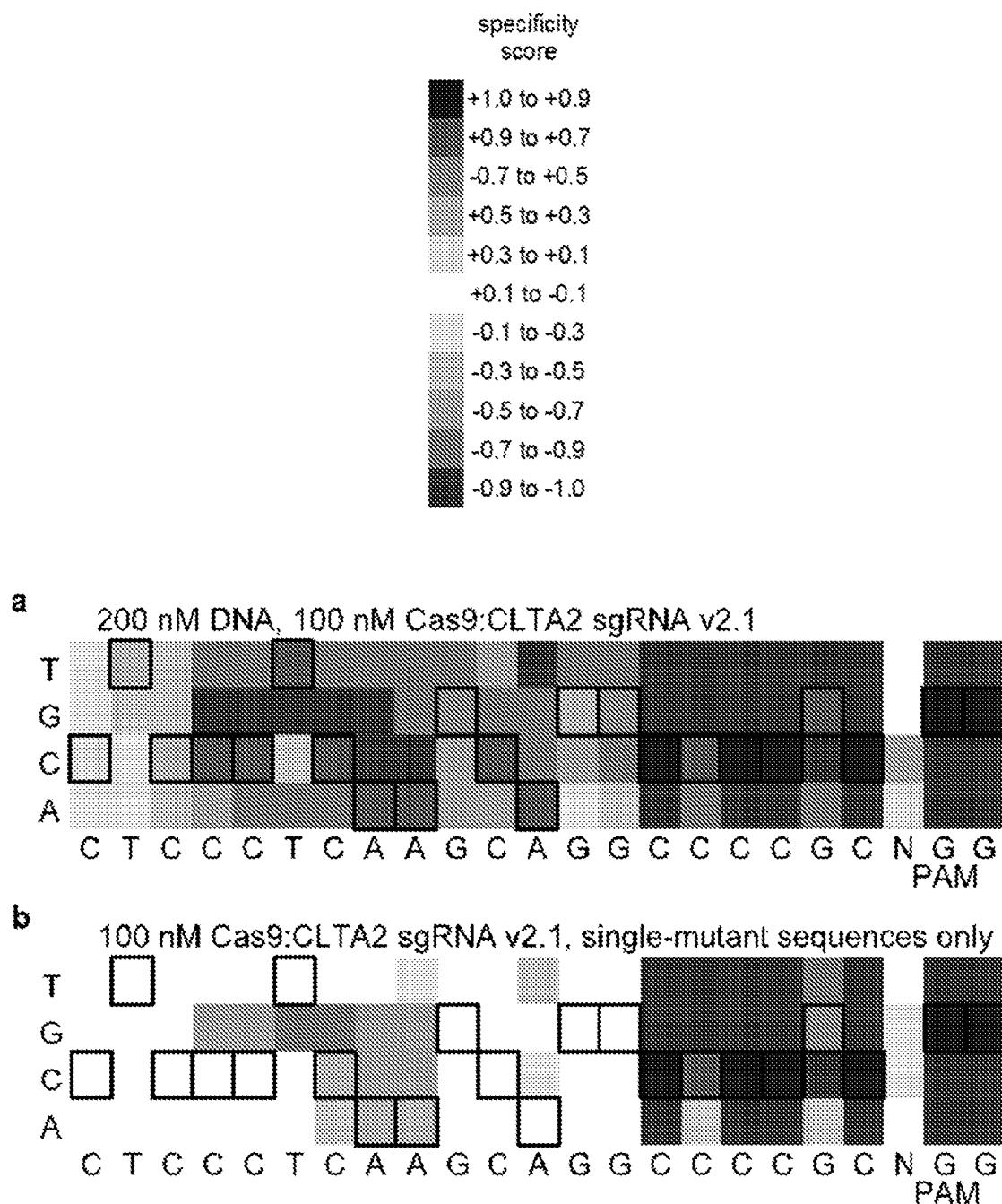
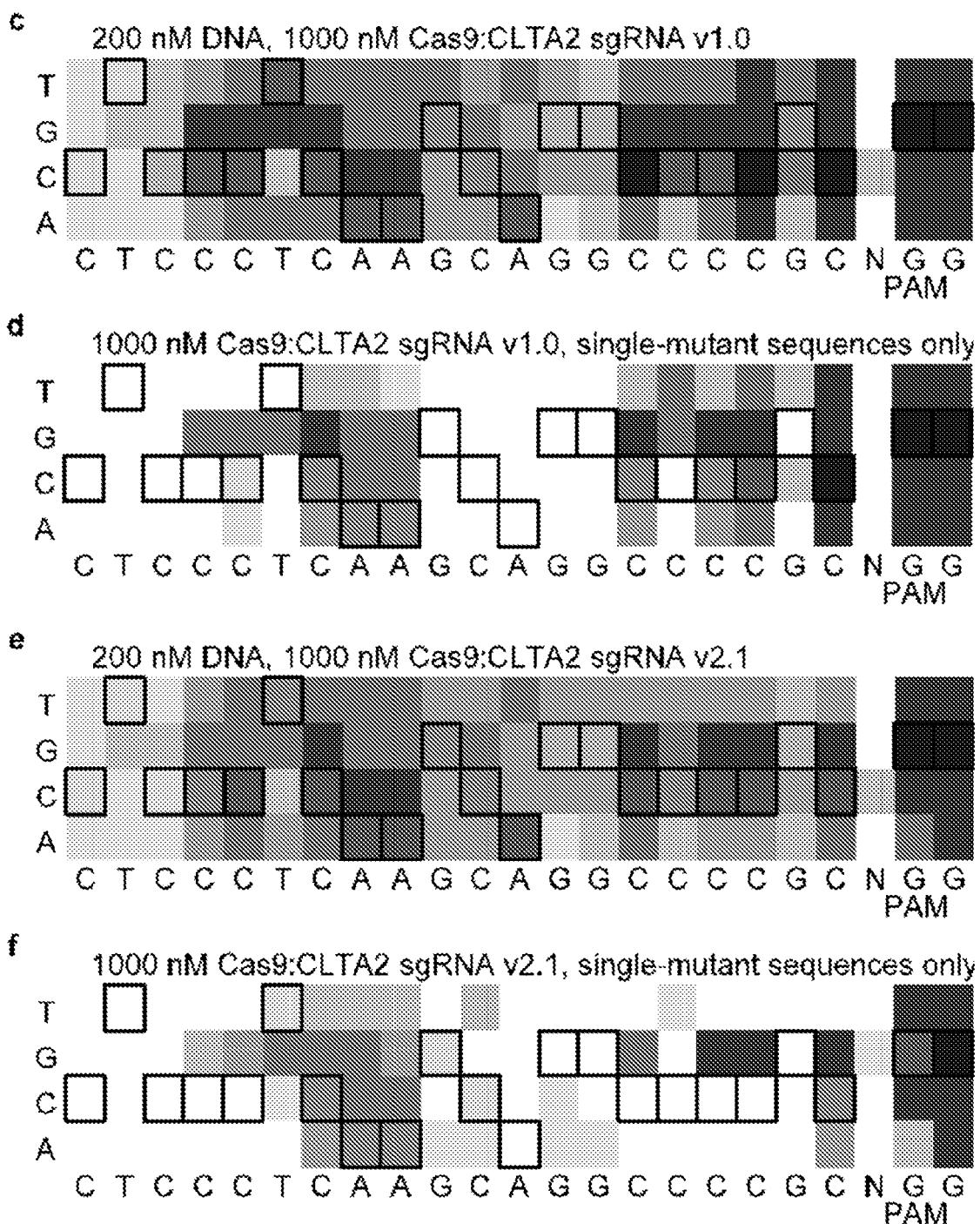


FIG. 6C-F

**FIG. 7A-B**

**FIG. 7-C-F**

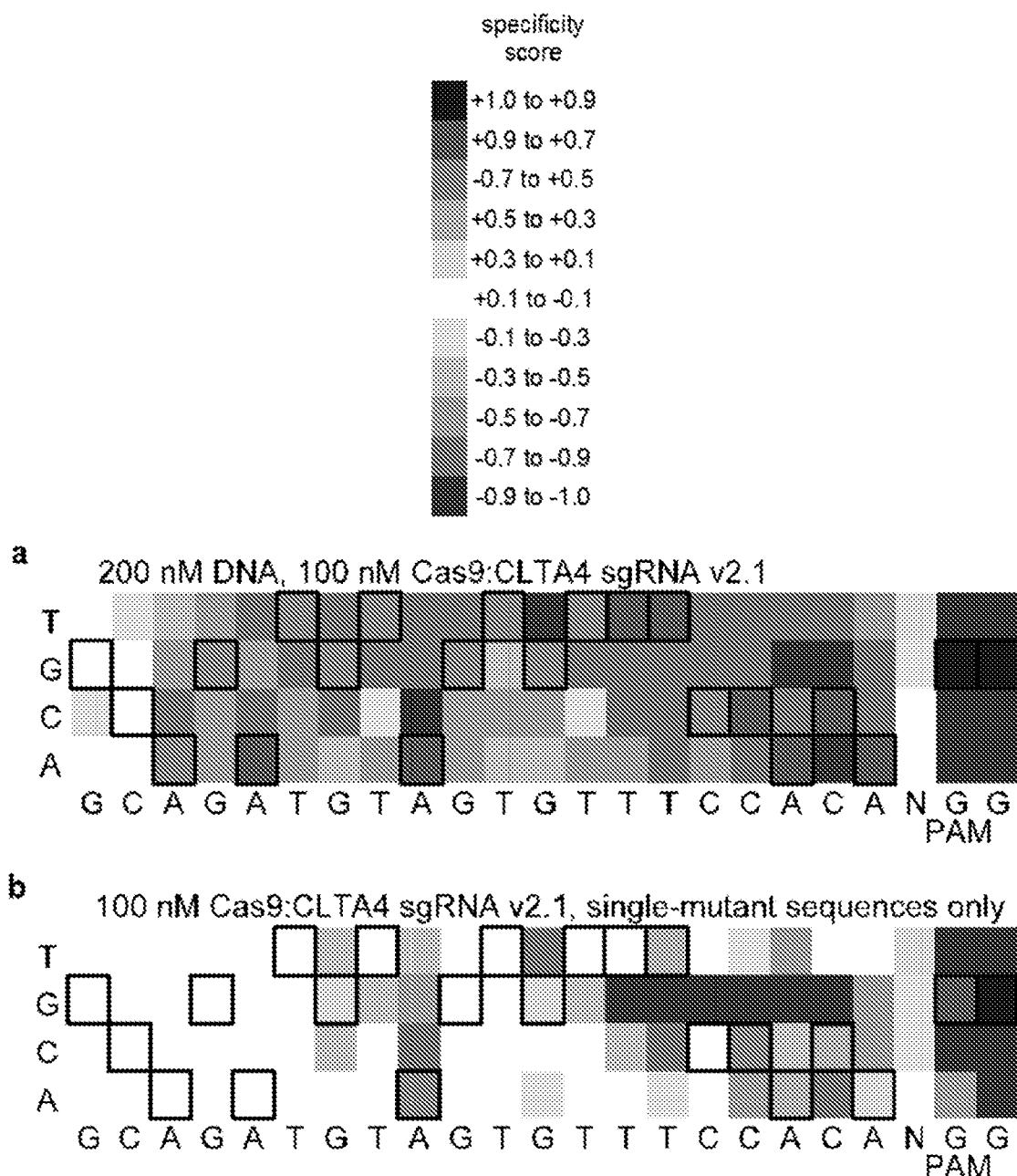
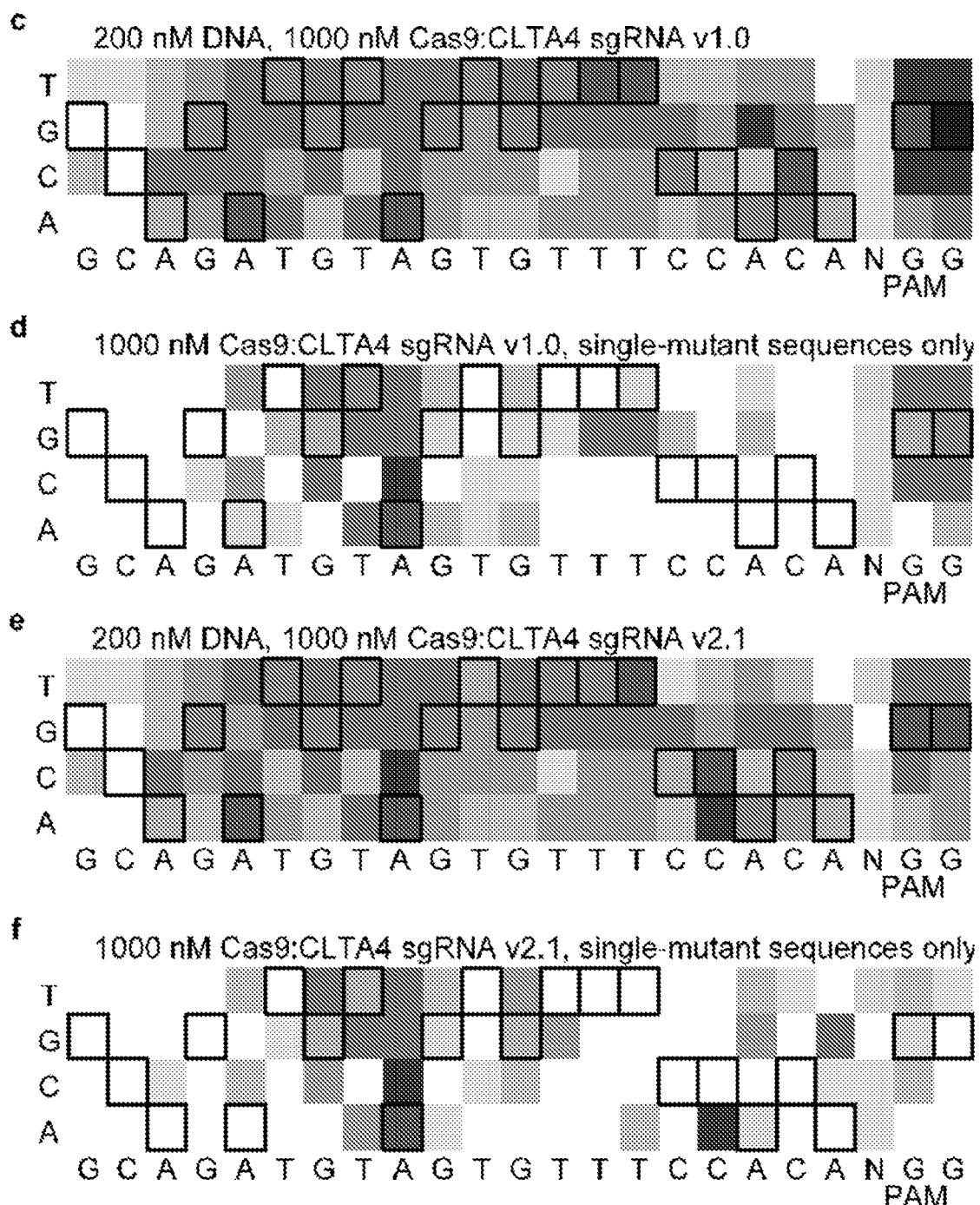


FIG. 8A-B

**FIG. 8C-F**

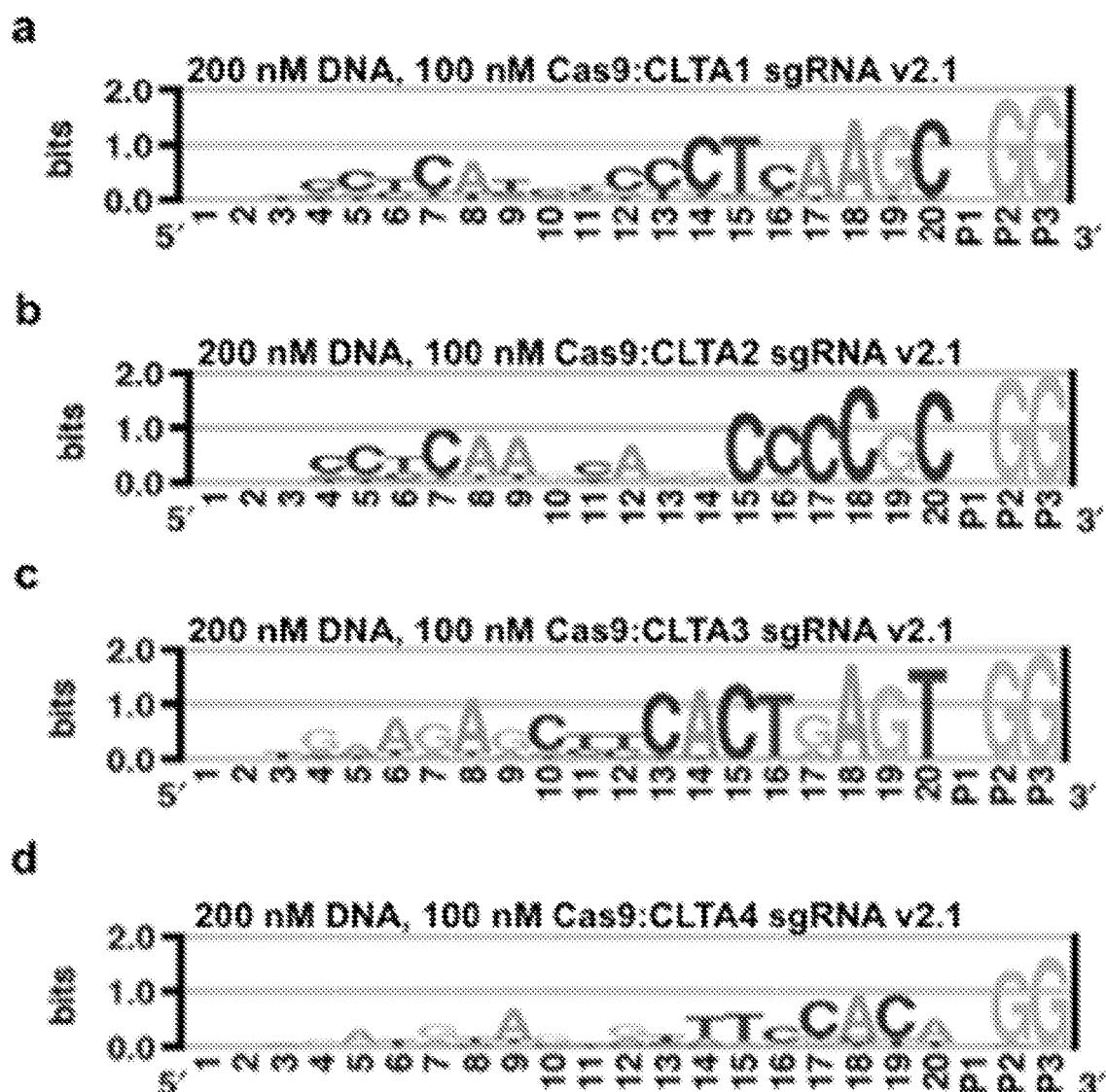
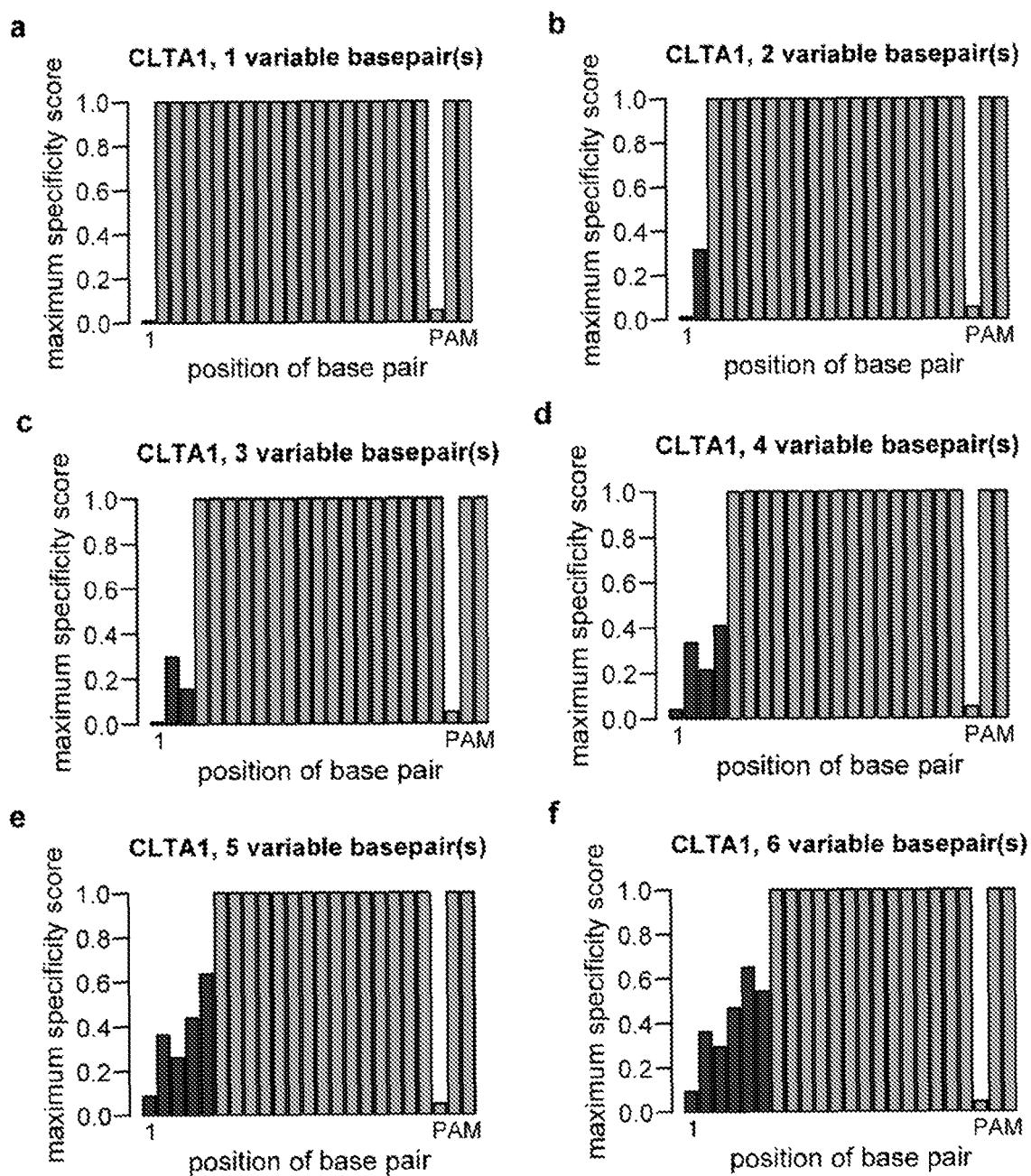
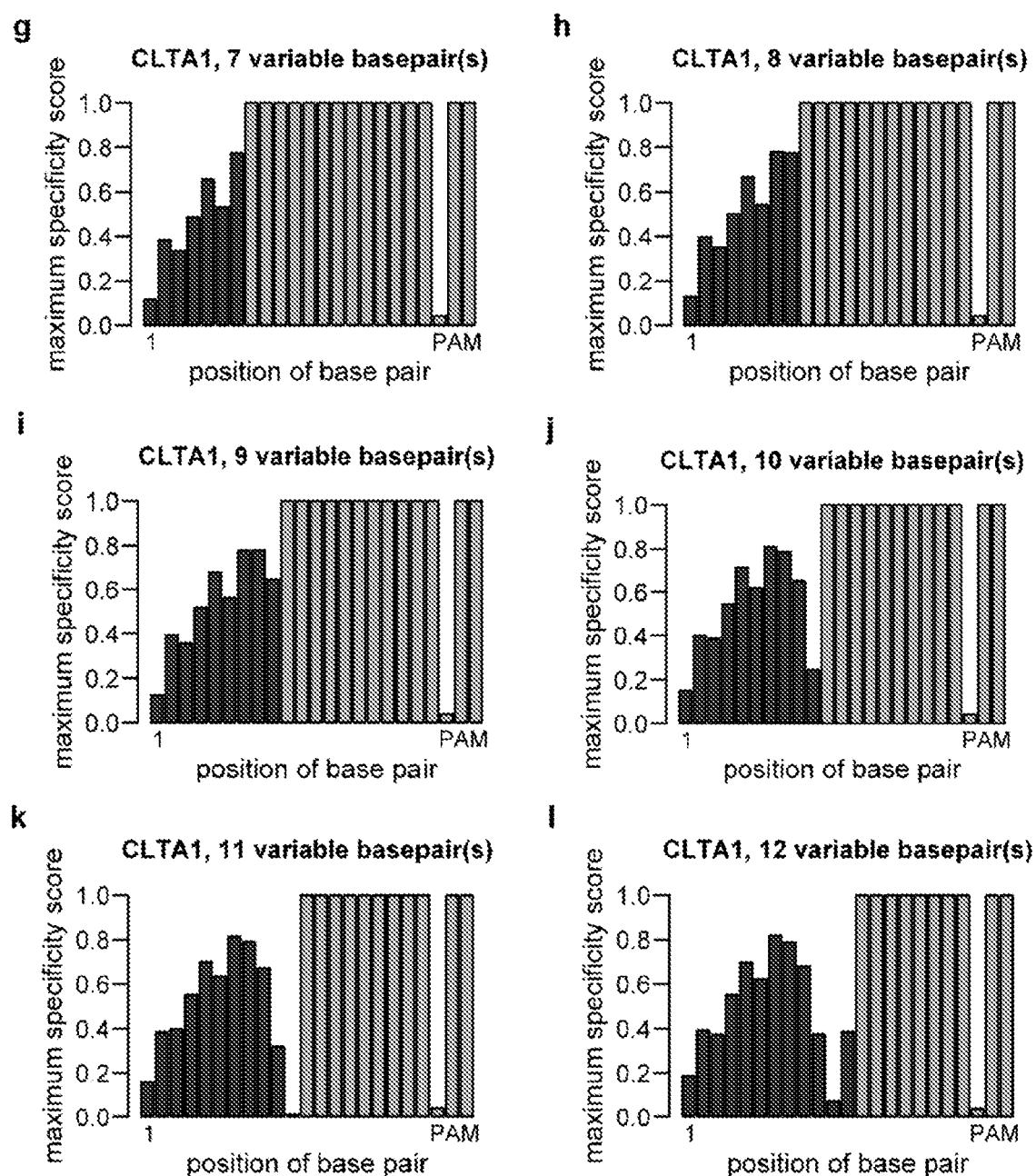
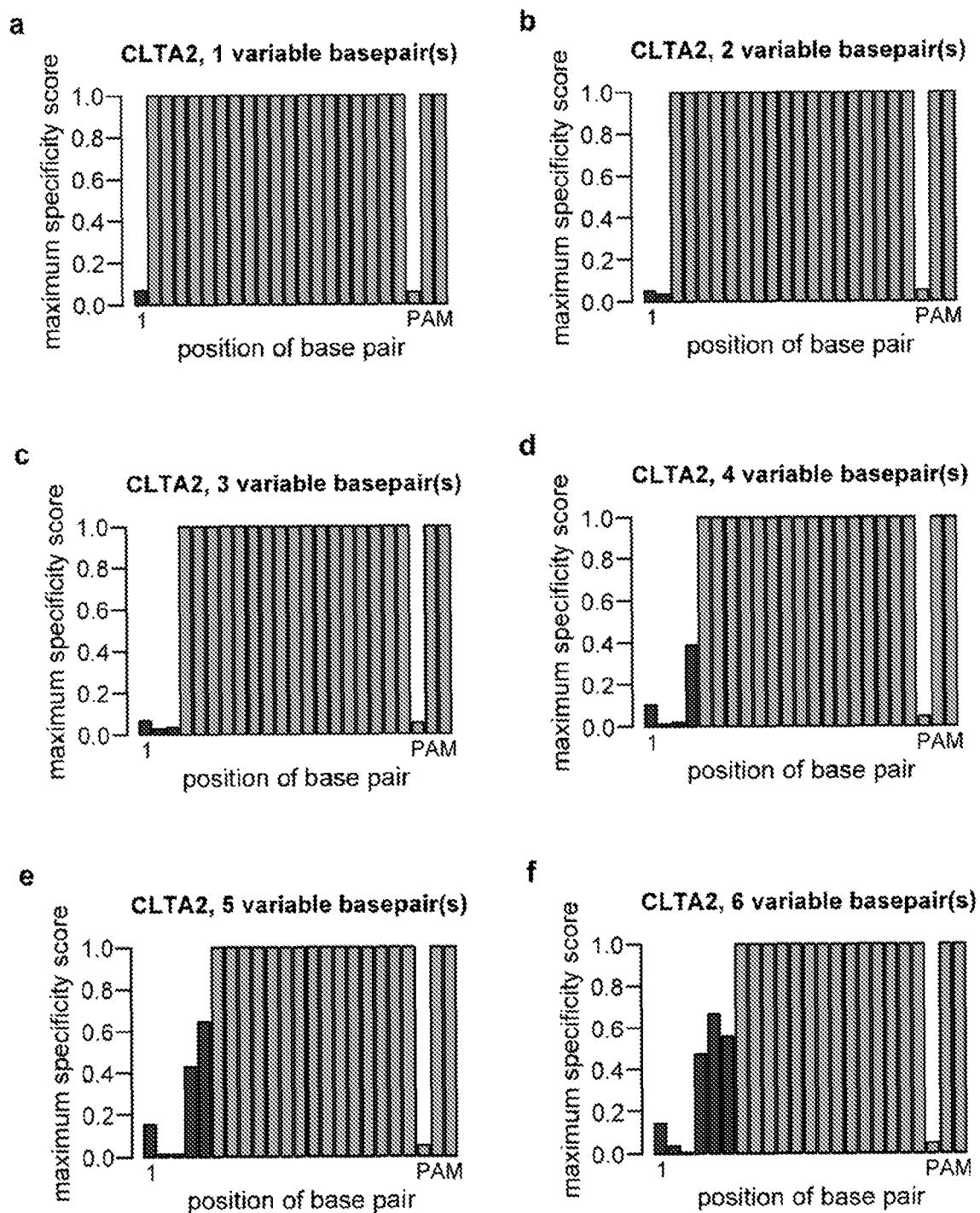
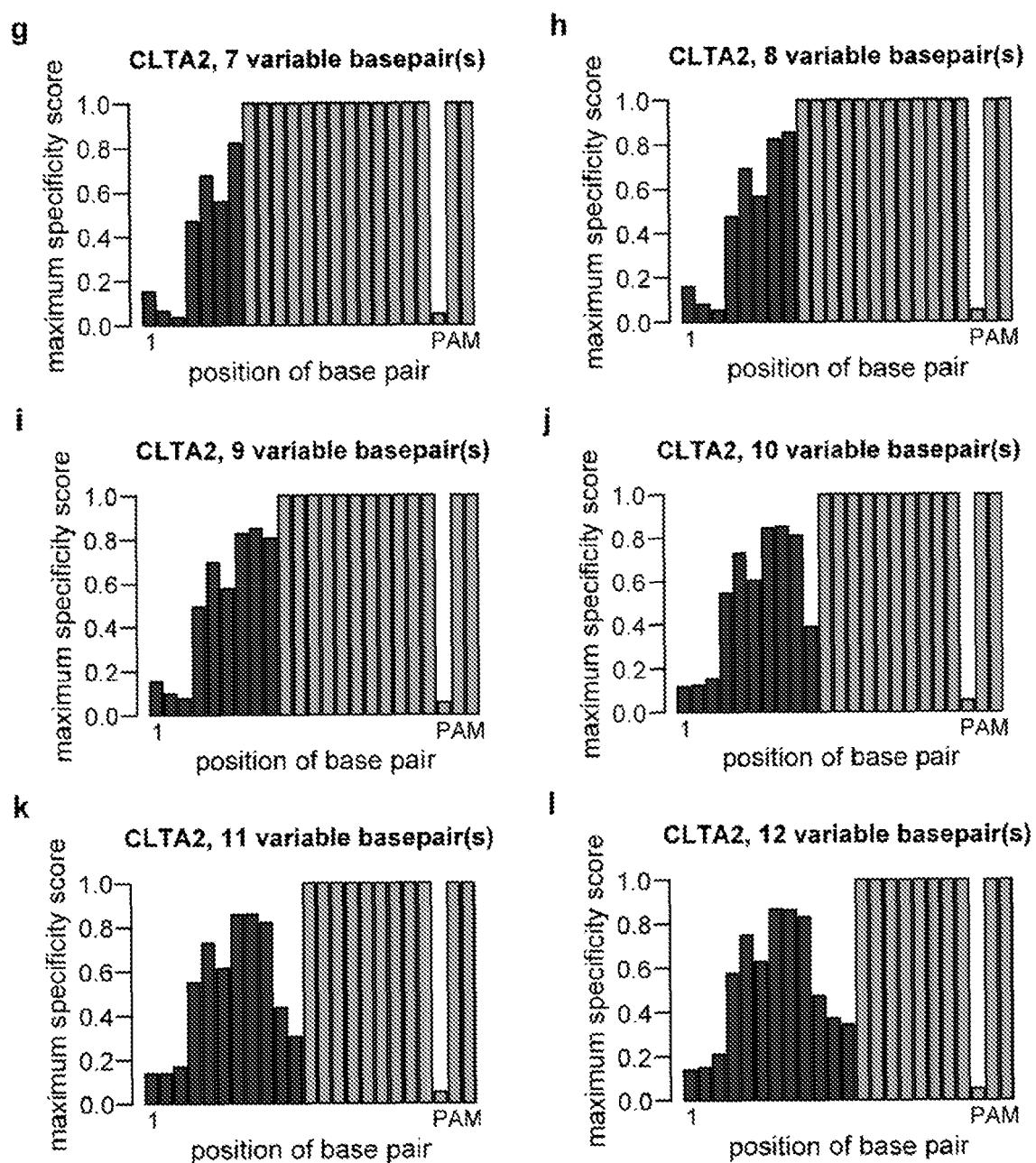


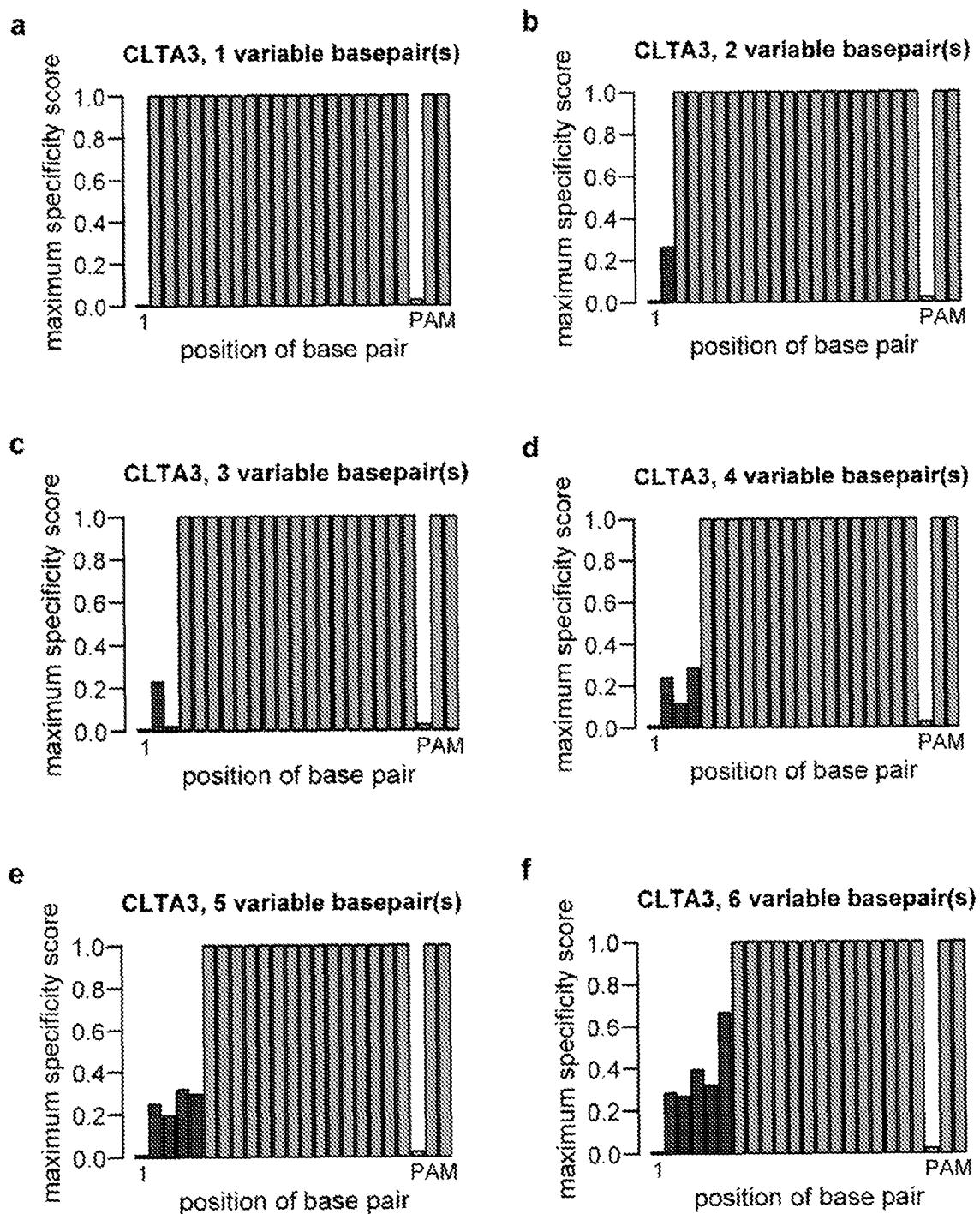
FIG. 9

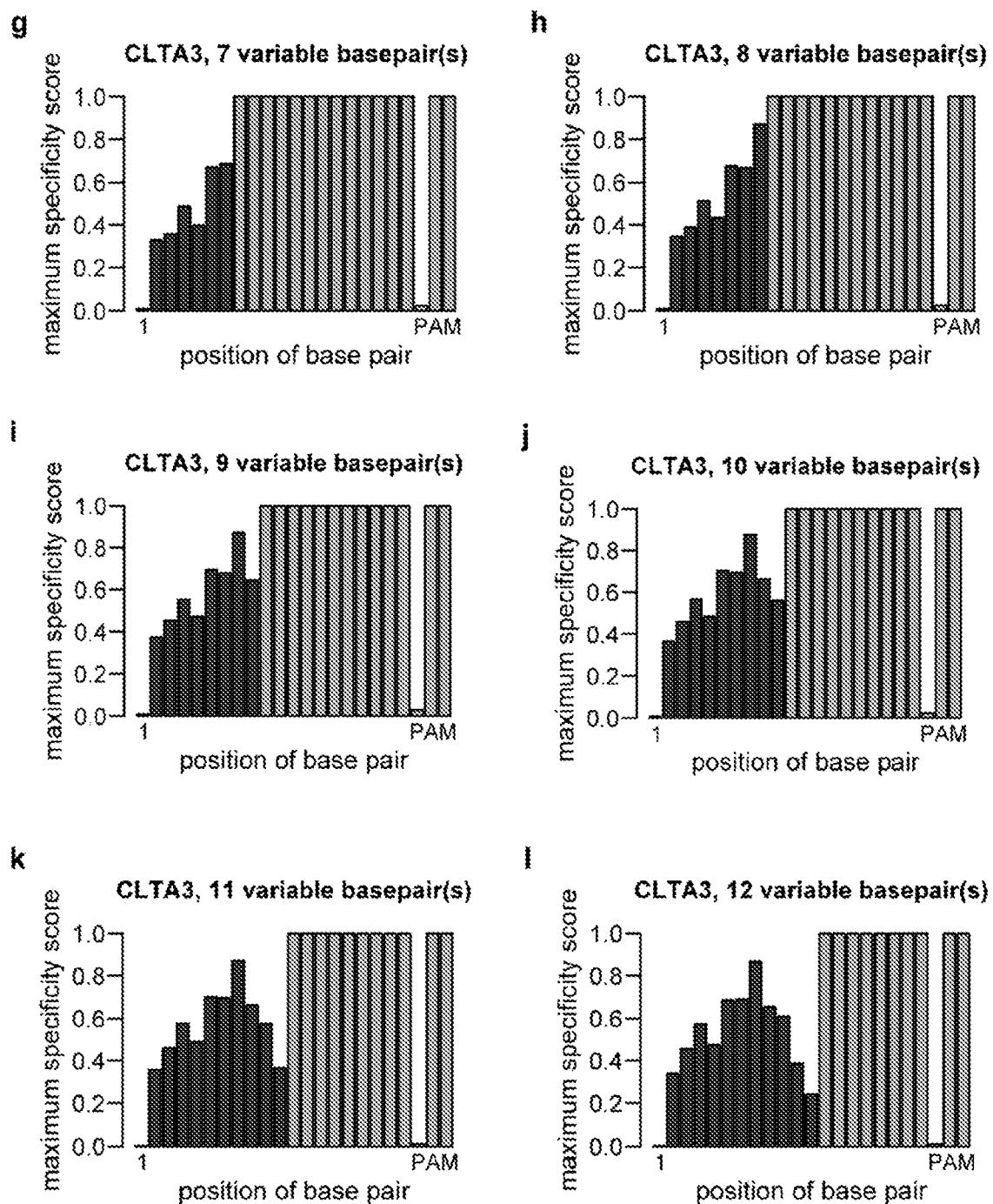
**FIG. 10A-F**

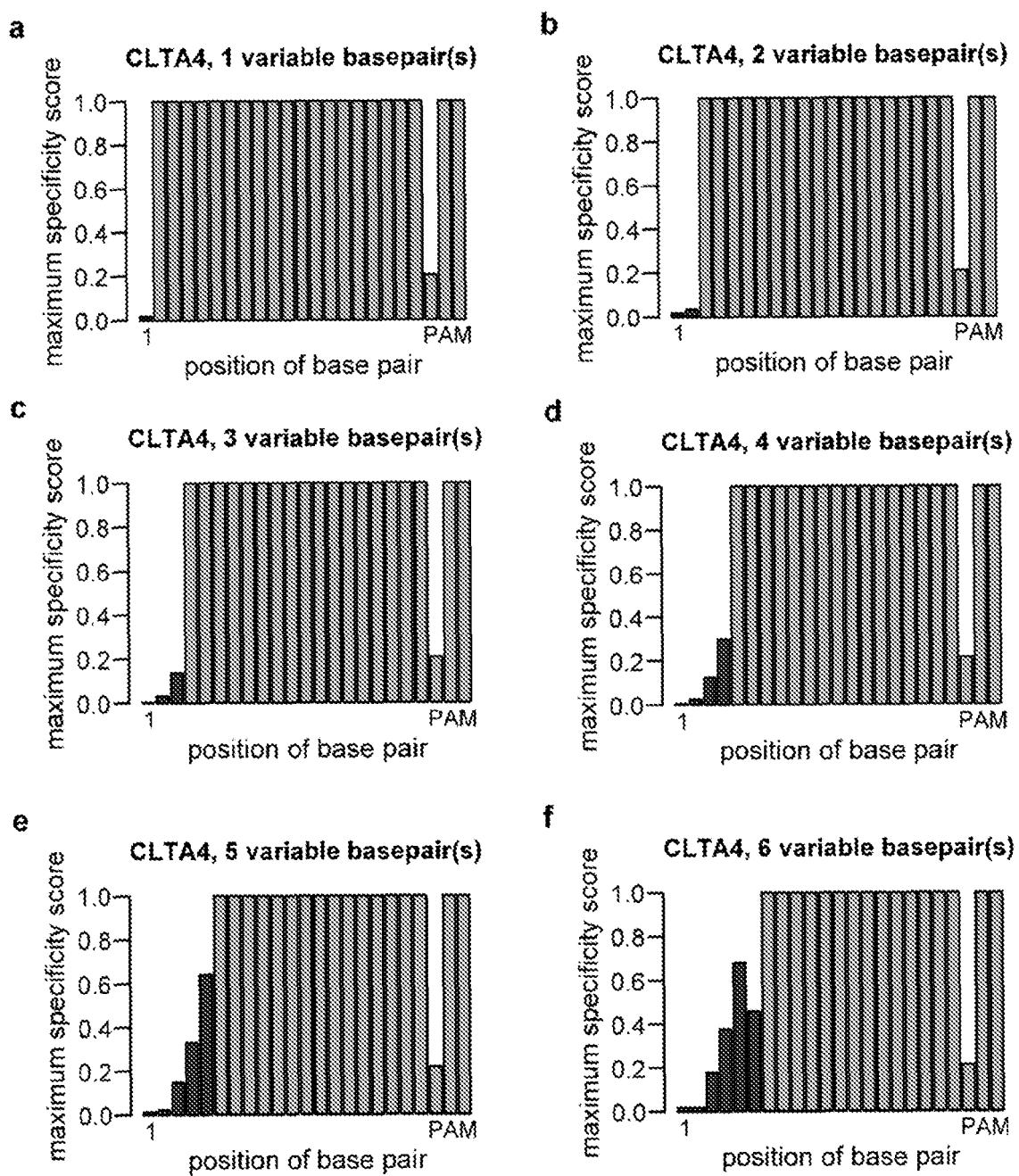
**FIG. 10G-L**

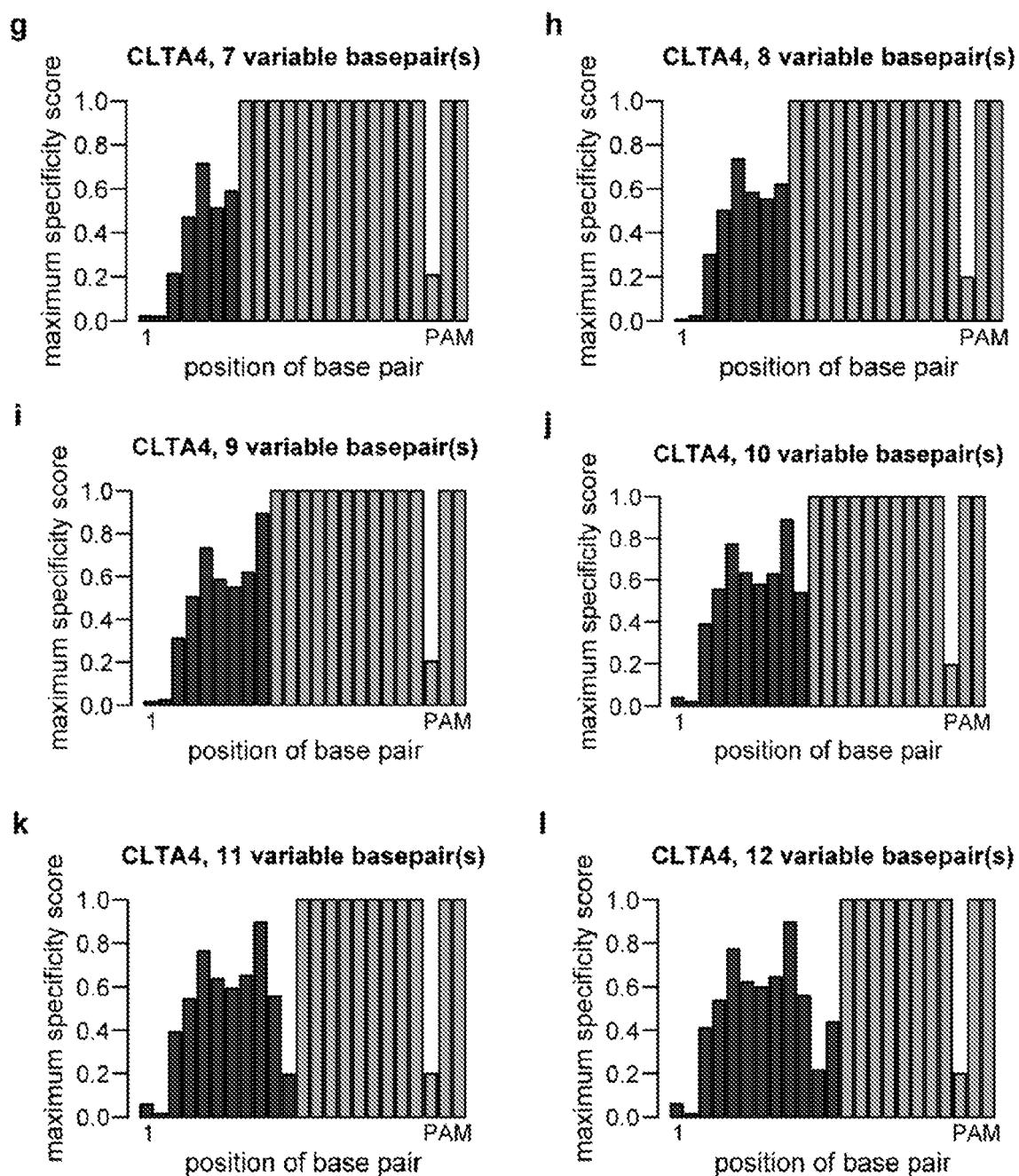
**FIG. 11A-F**

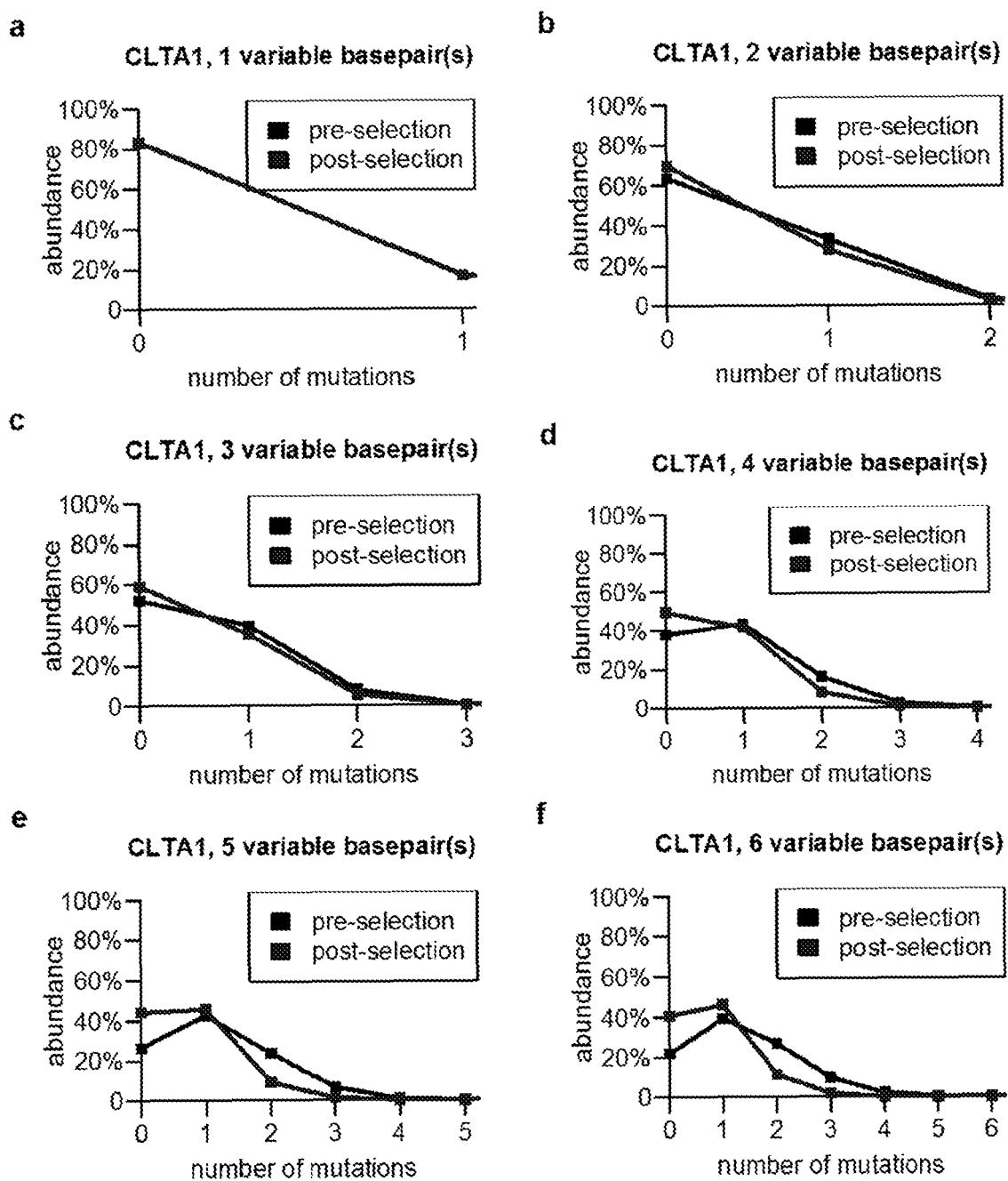
**FIG. 11G-L**

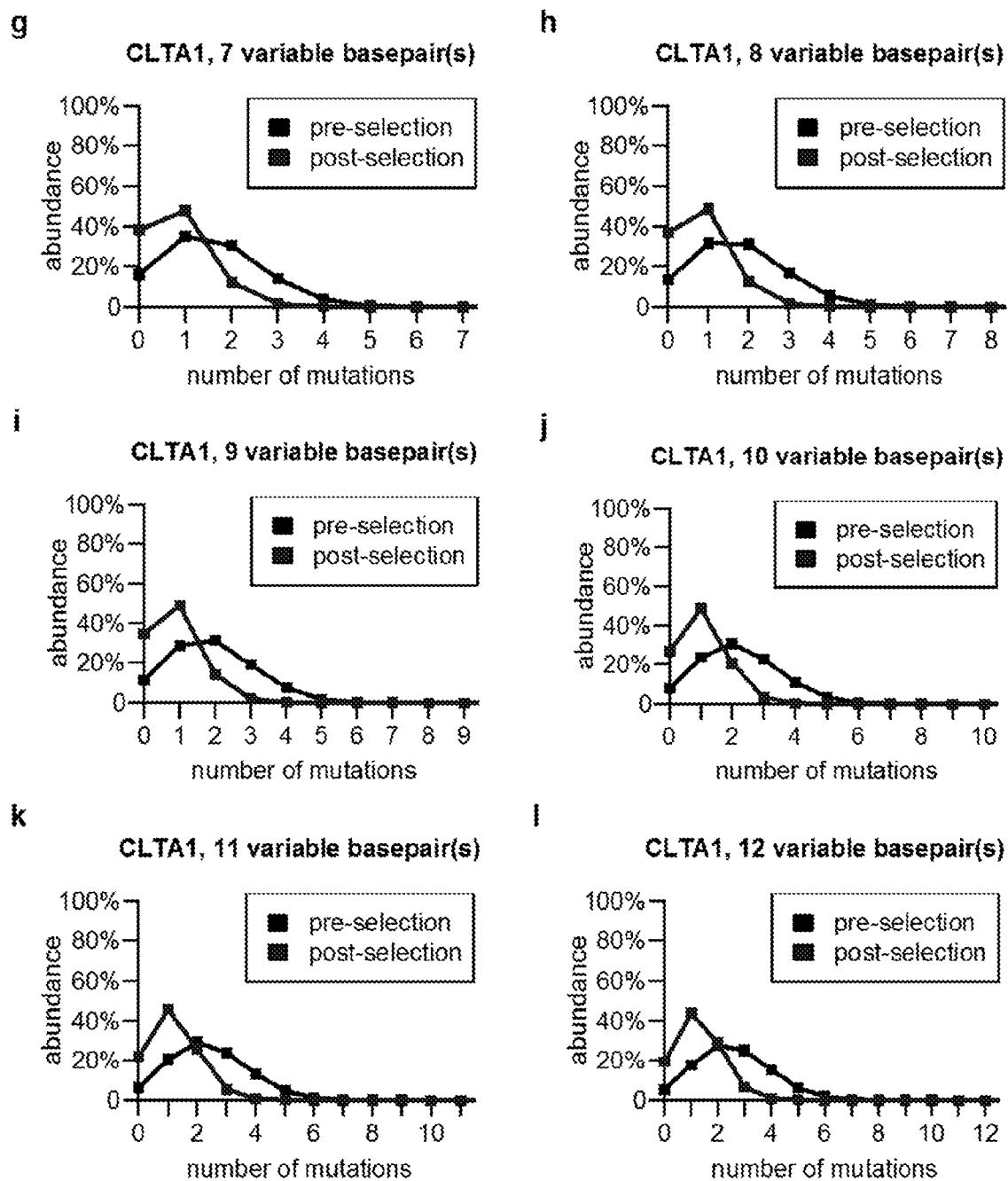
**FIG. 12A-F**

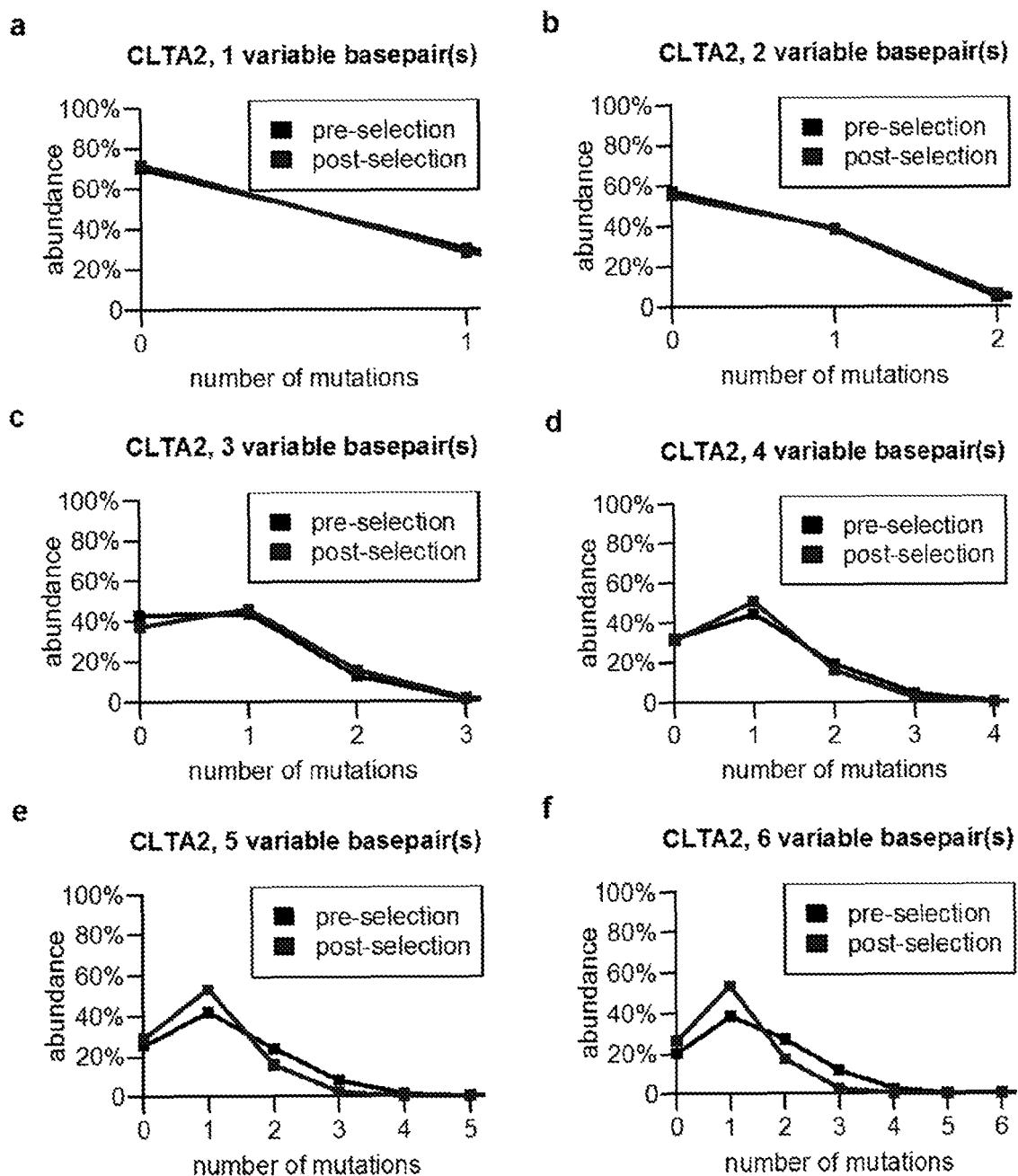
**FIG. 12G-L**

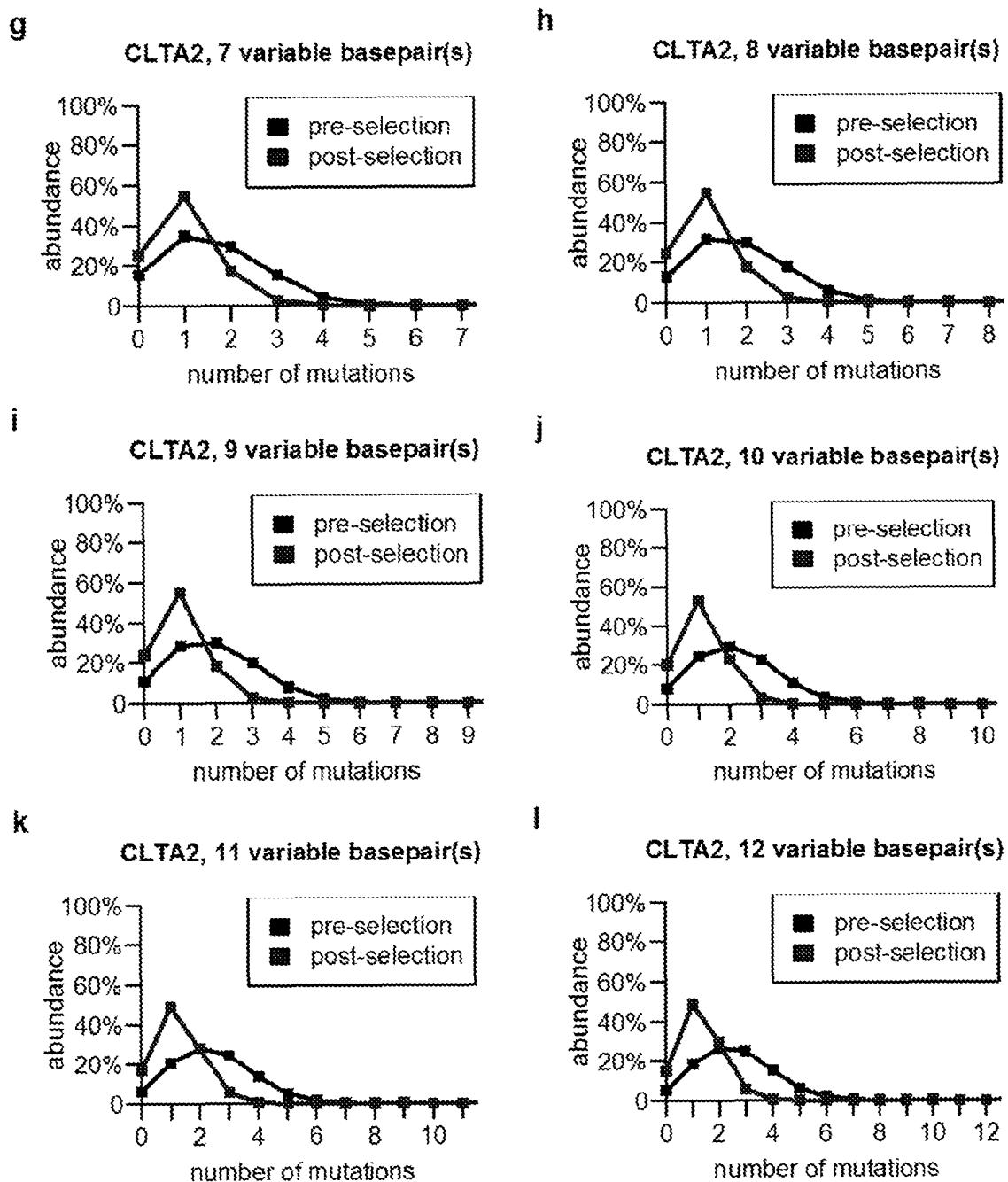
**FIG. 13A-F**

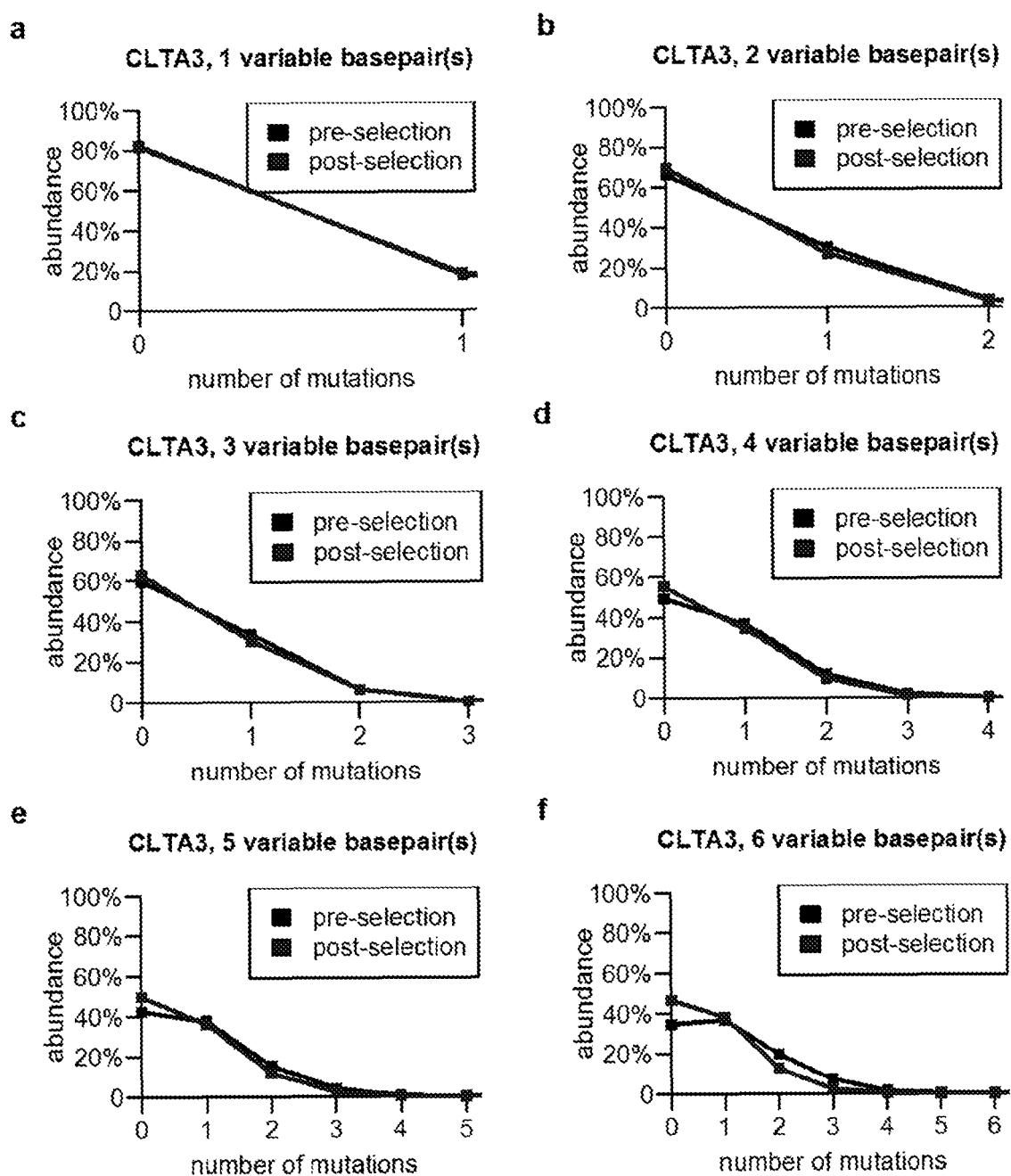
**FIG. 13G-L**

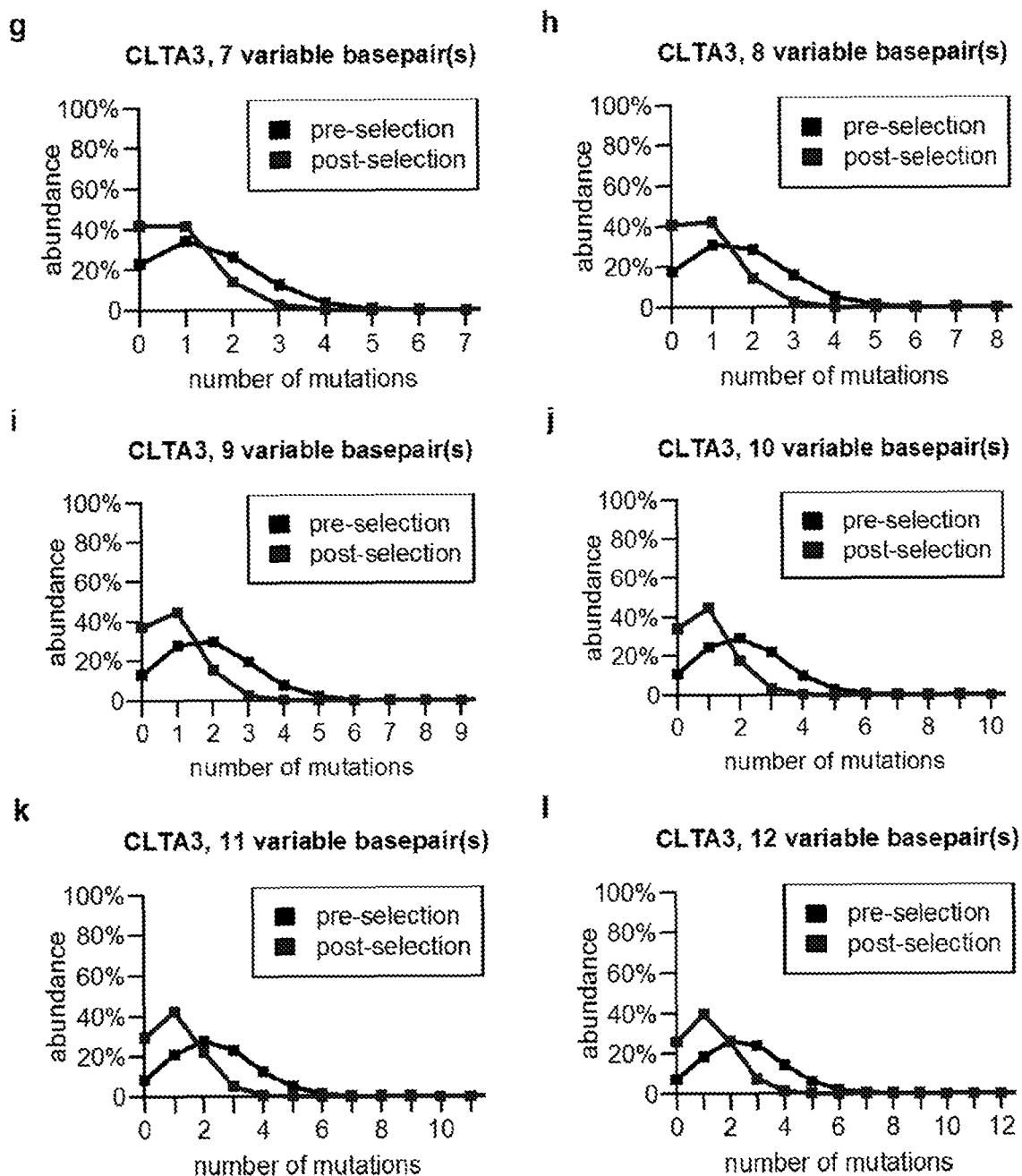
**FIG. 14A-F**

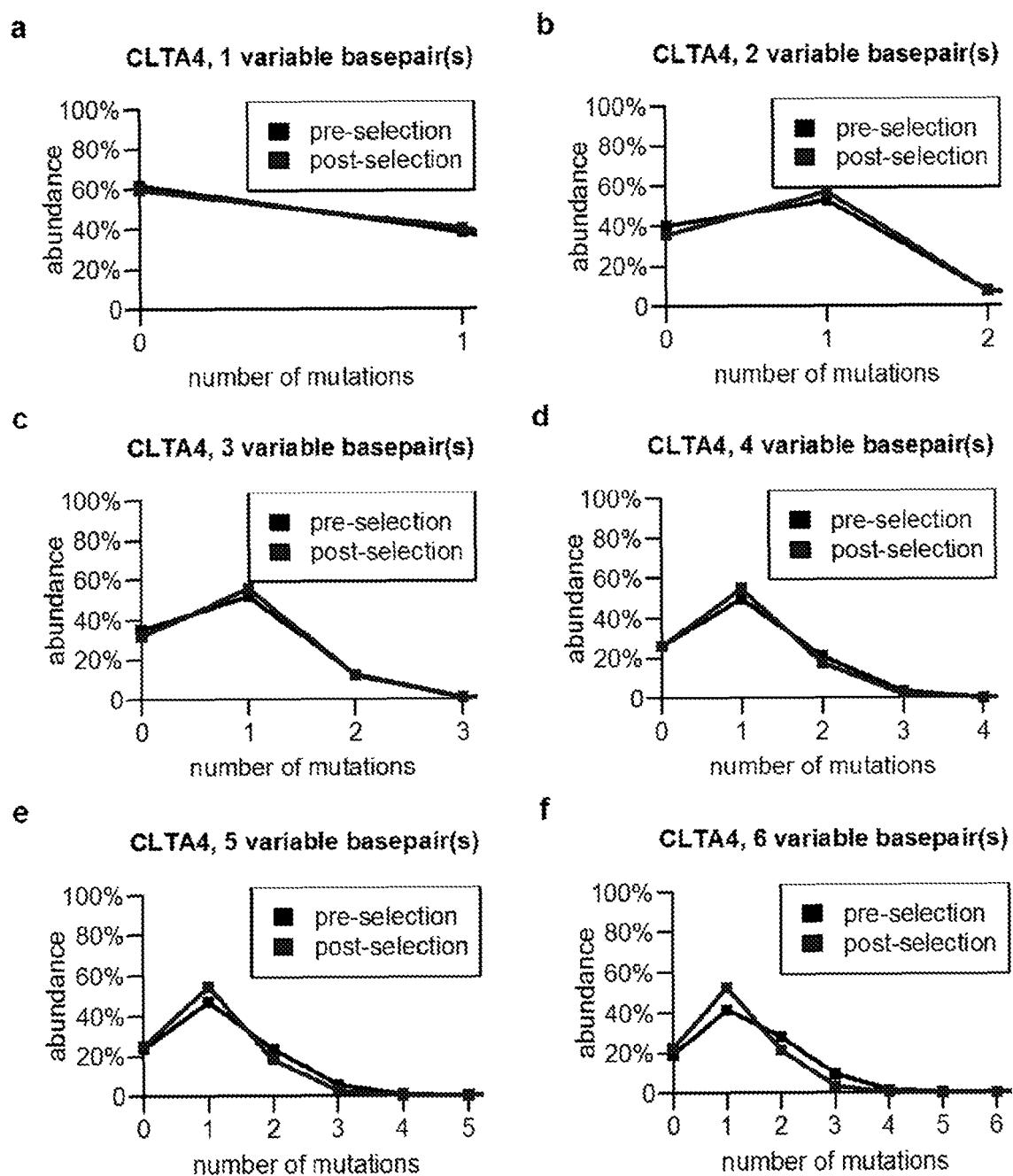
**FIG. 14G-L**

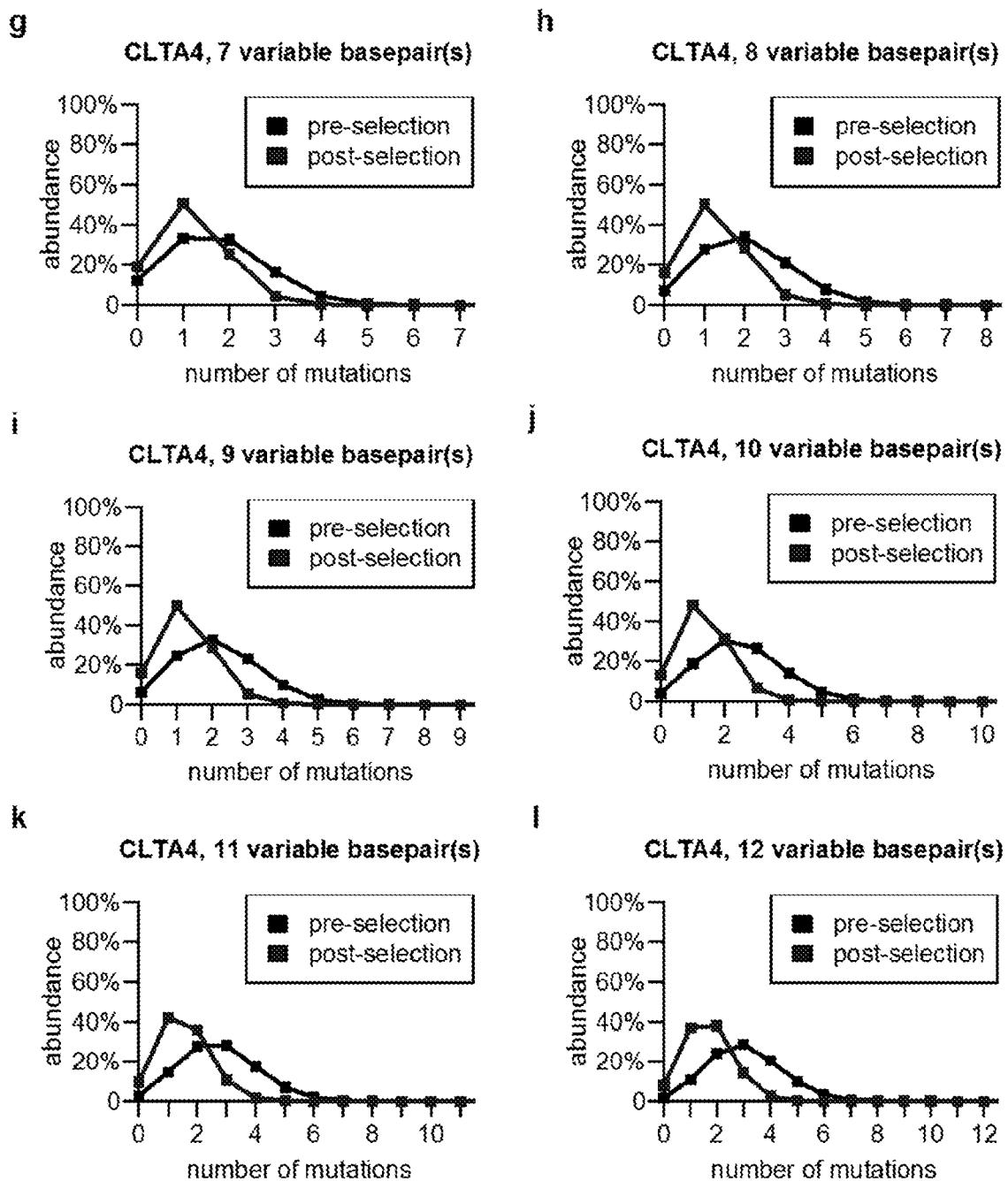
**FIG. 15A-F**

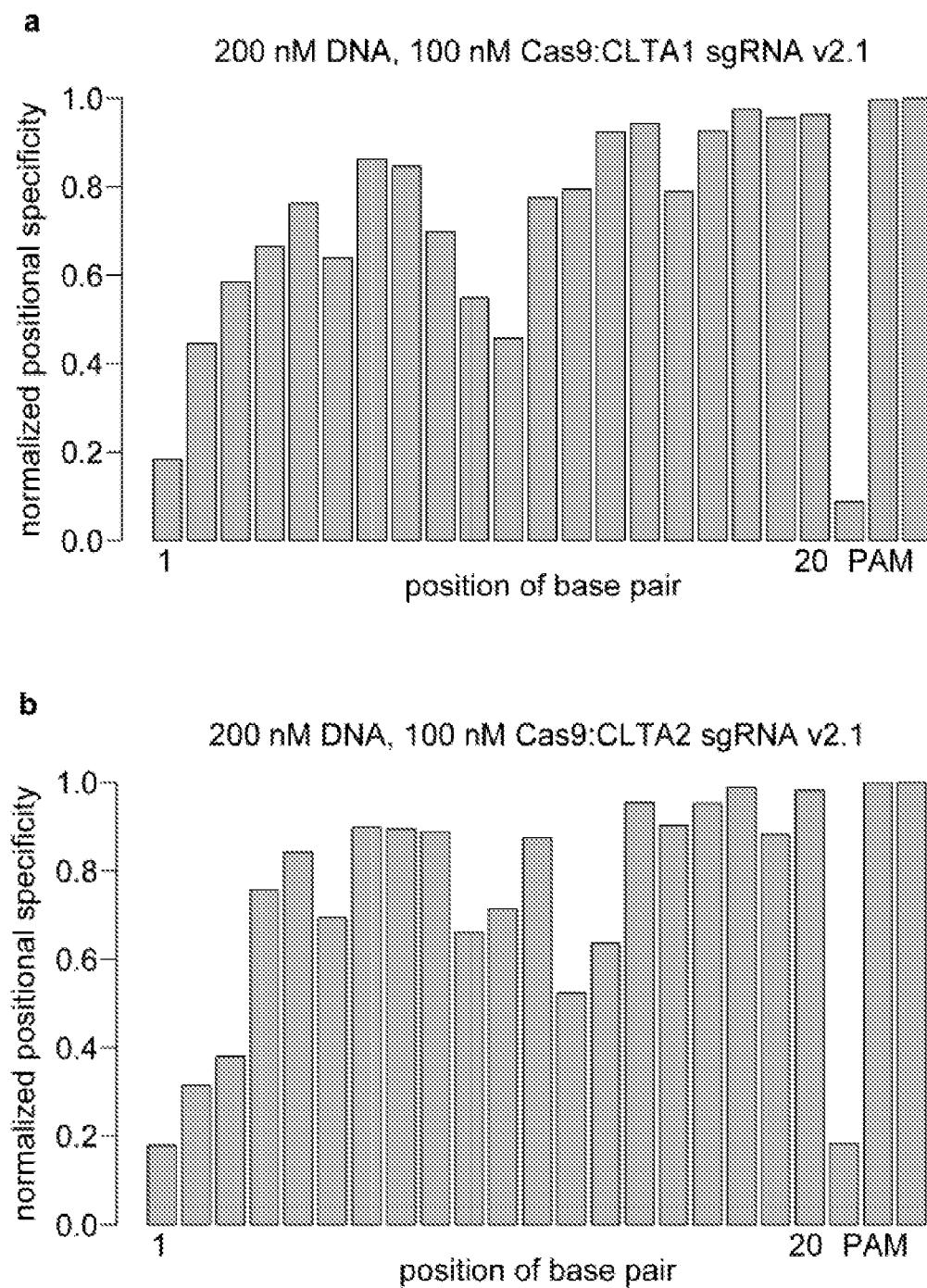
**FIG. 15G-L**

**FIG. 16A-F**

**FIG. 16G-L**

**FIG. 17A-F**

**FIG. 17G-L**

**FIG. 18A-B**

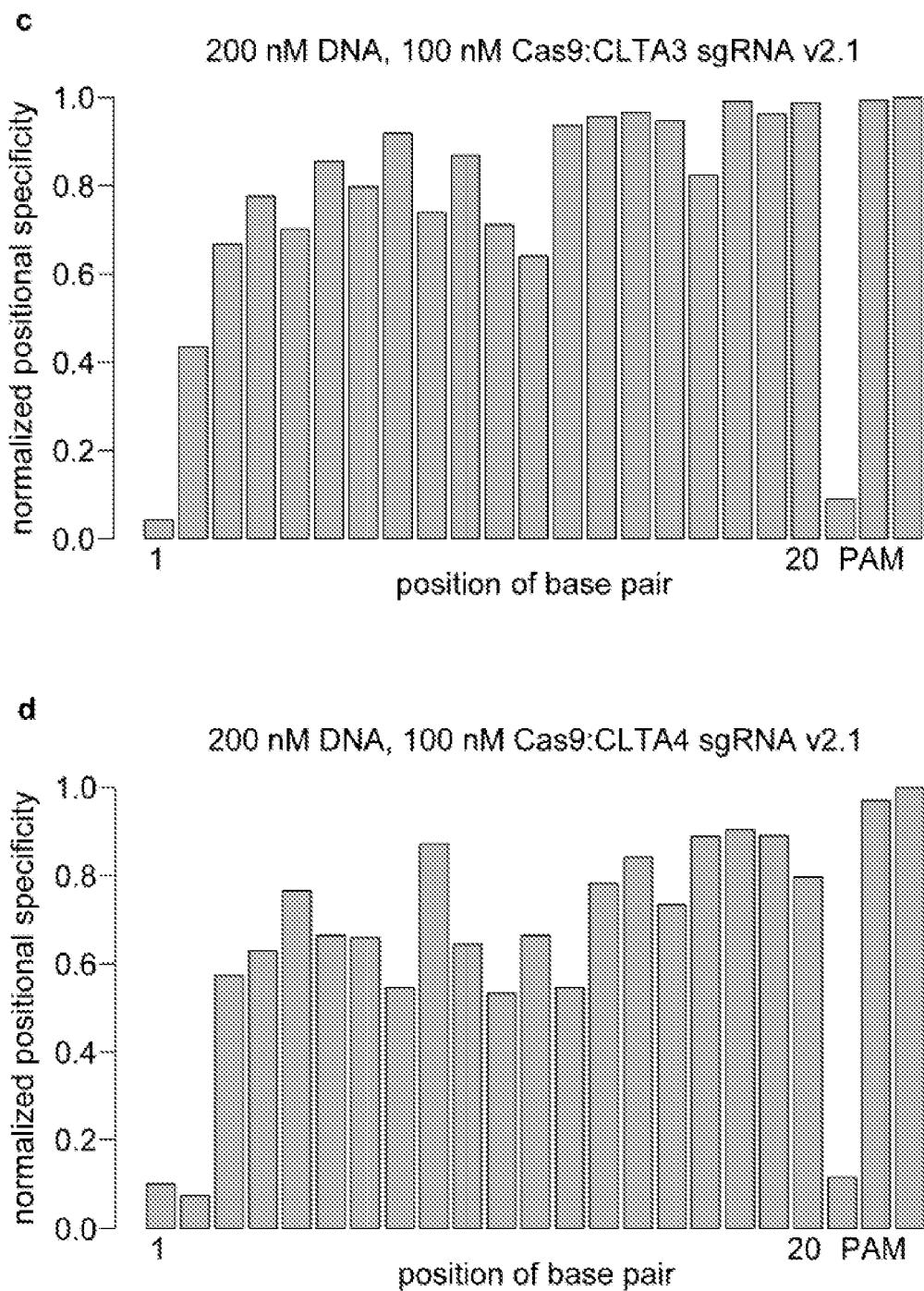


FIG. 18C-D

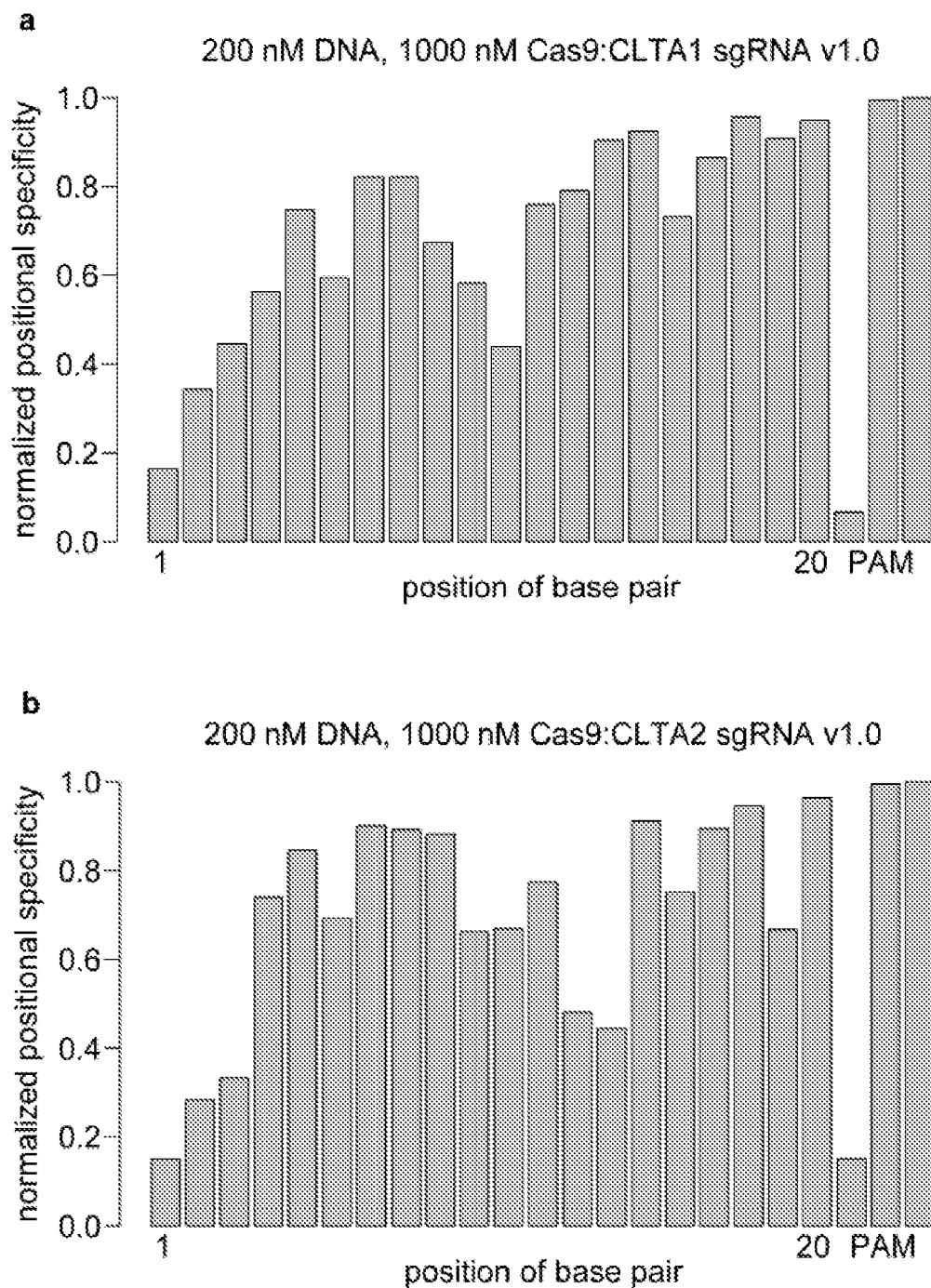


FIG. 19A-B

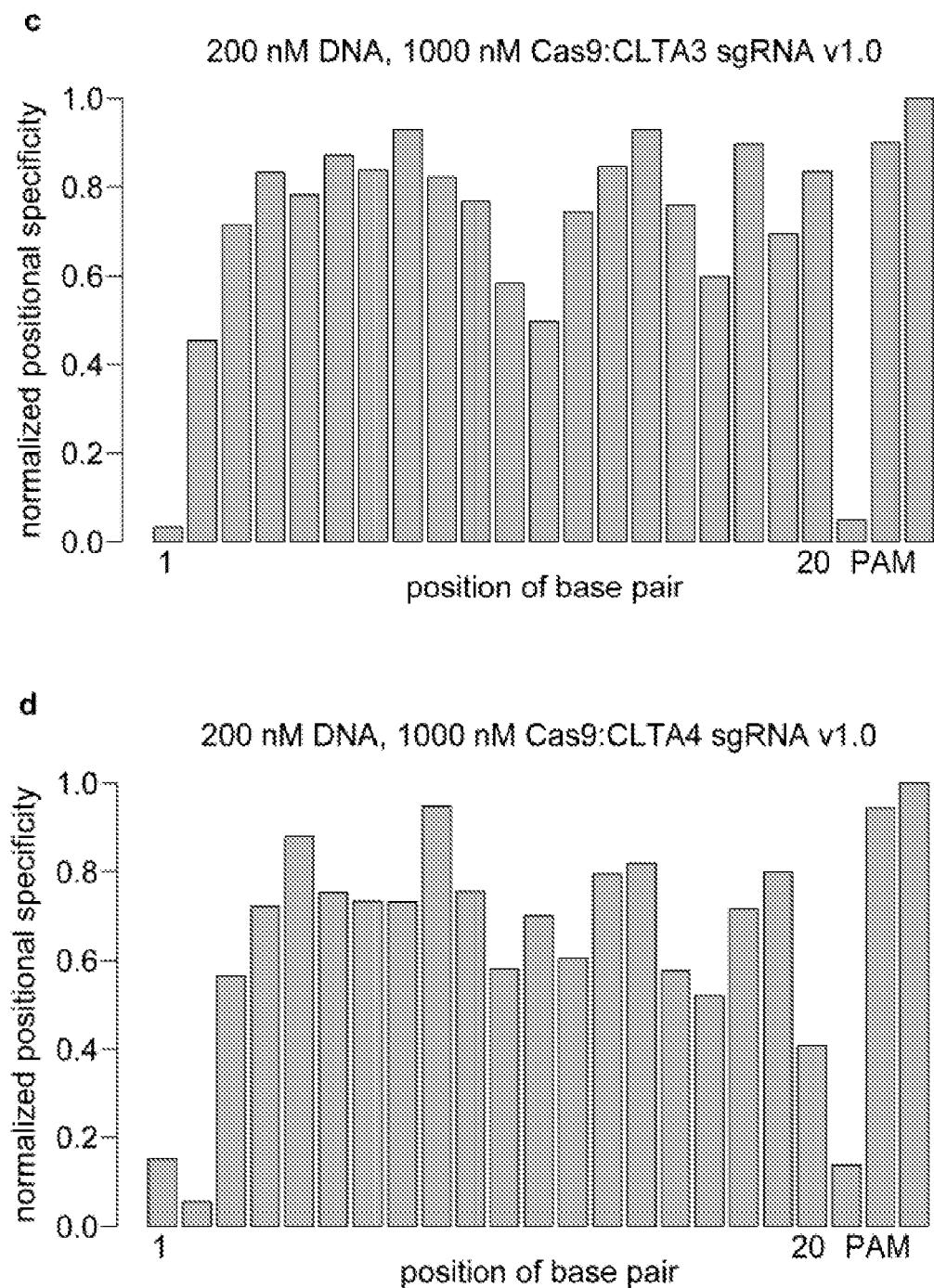
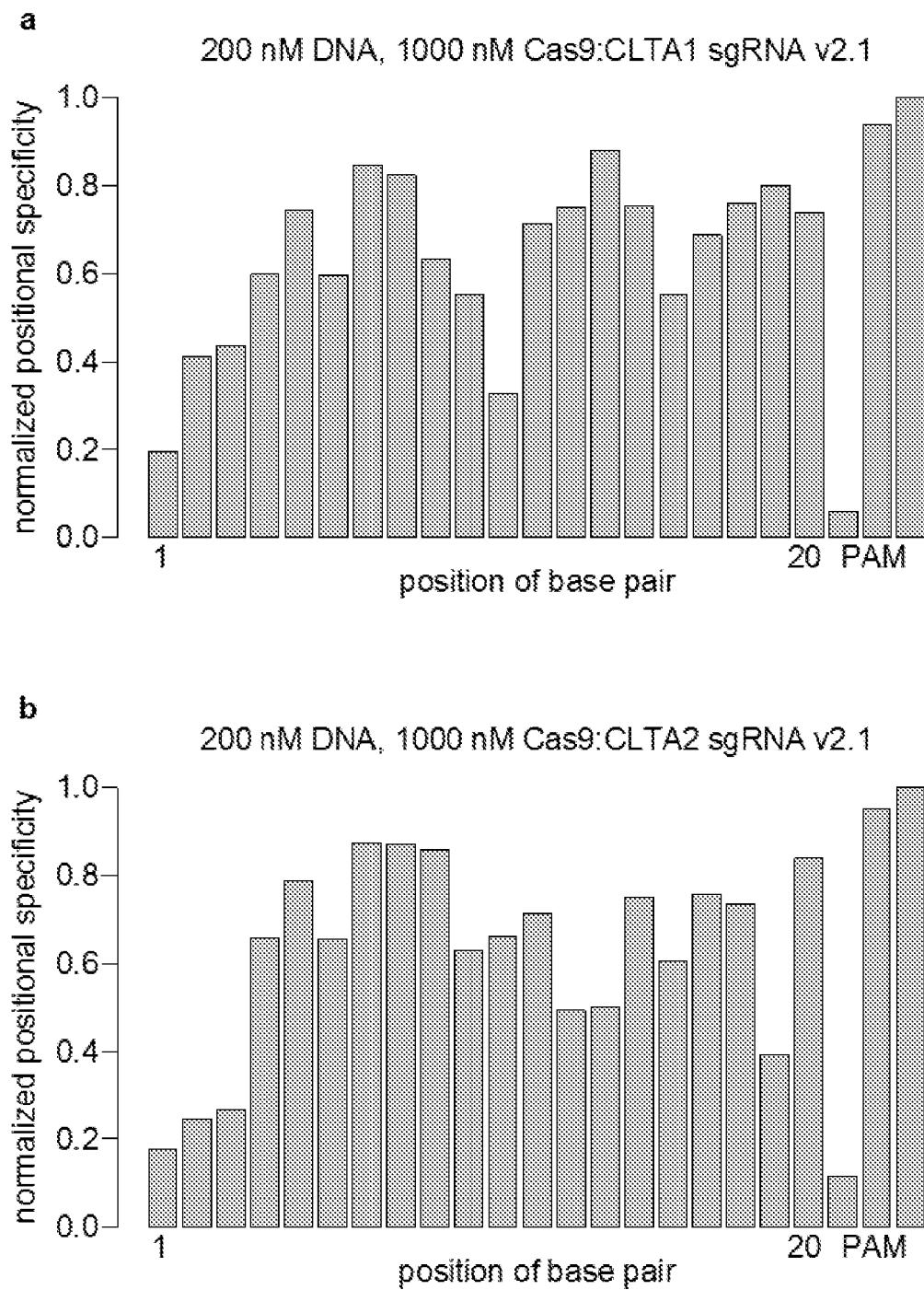
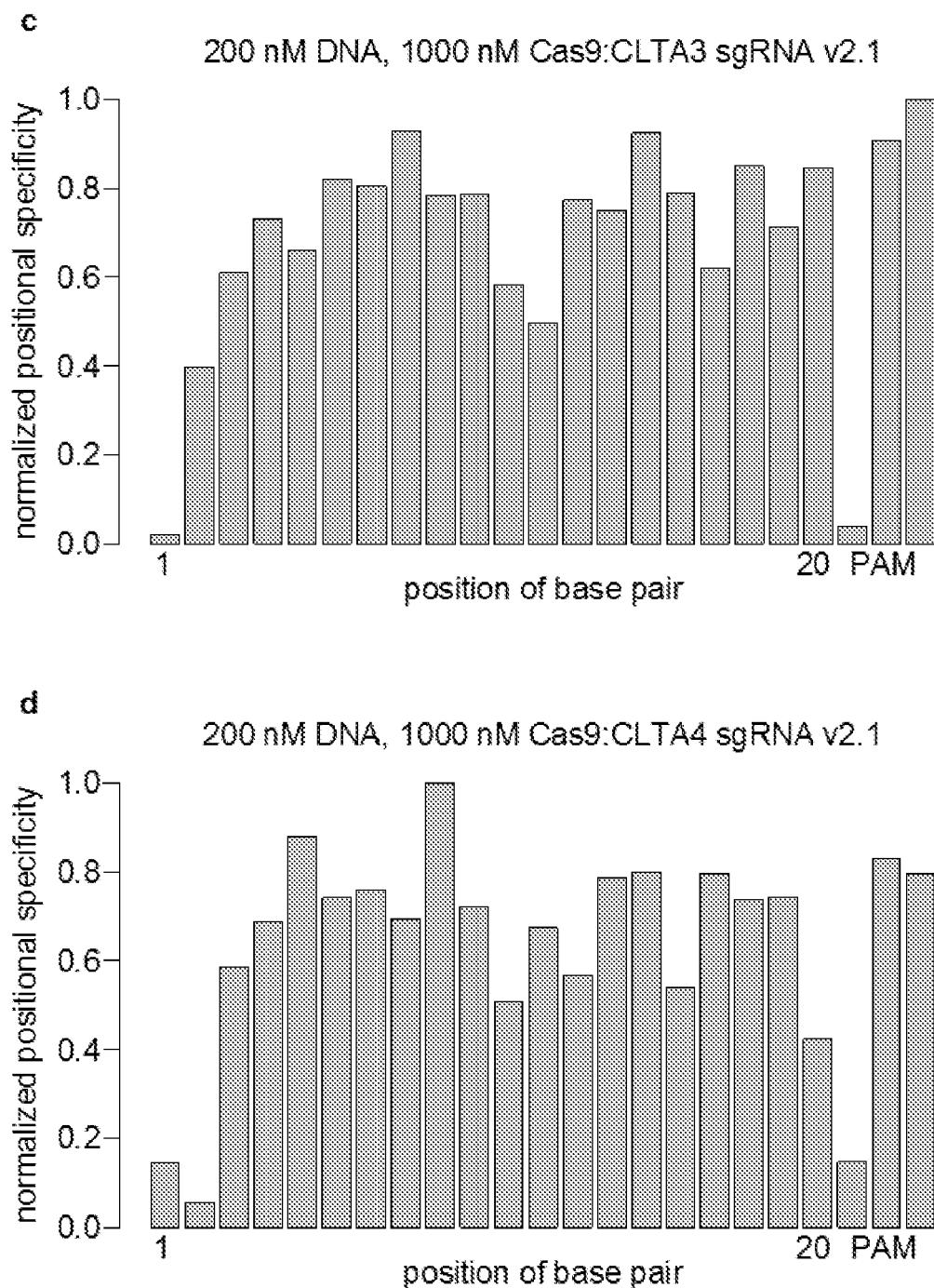
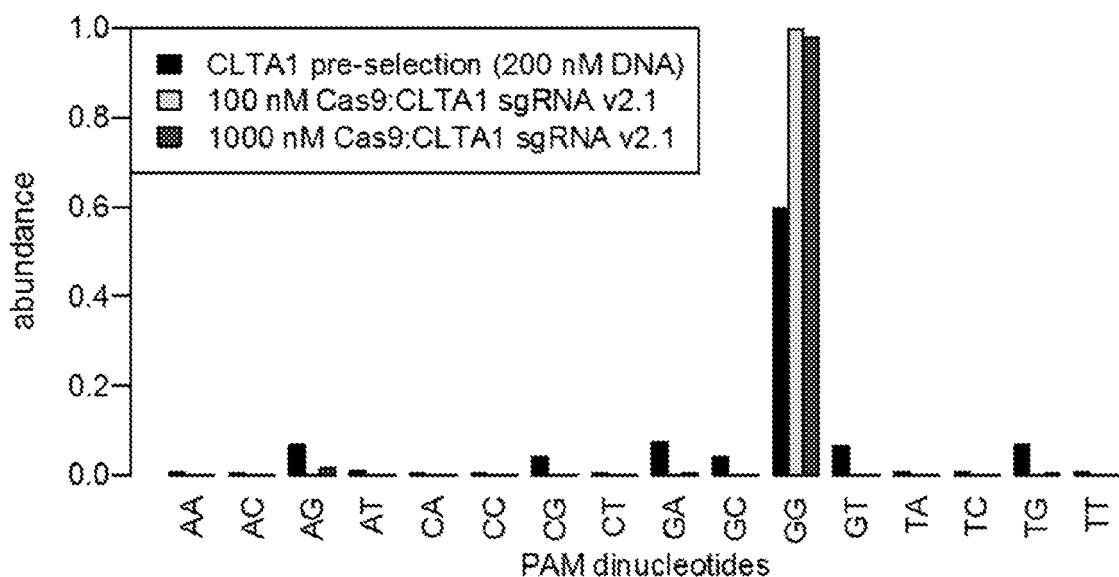
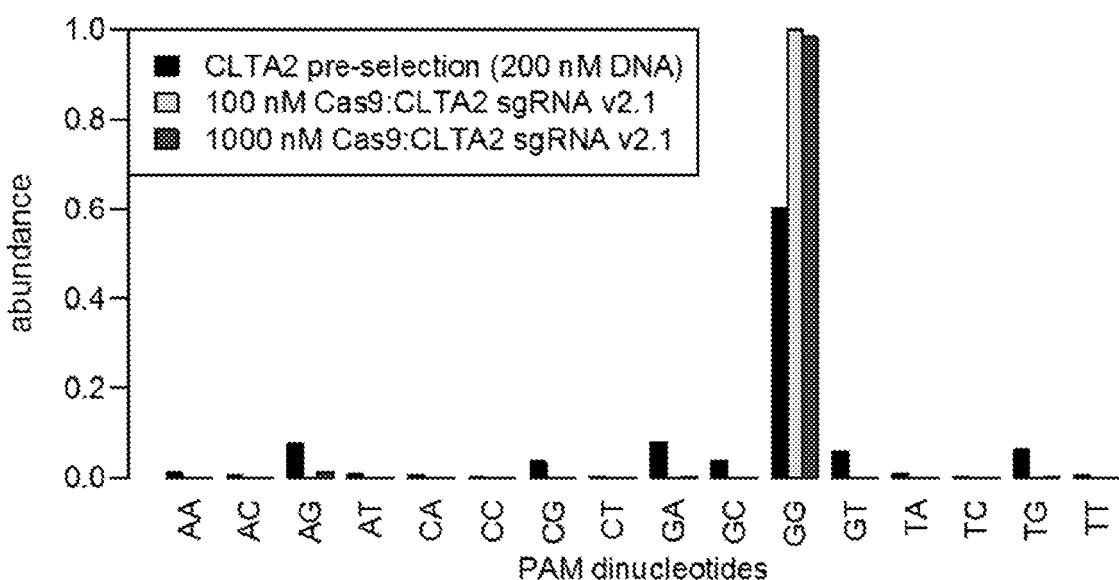
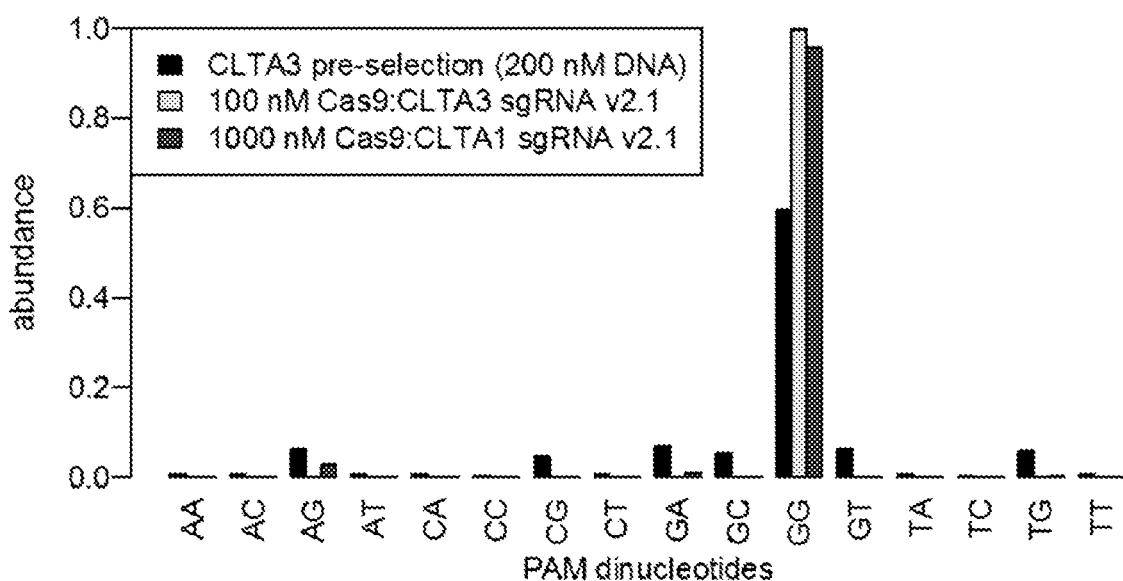
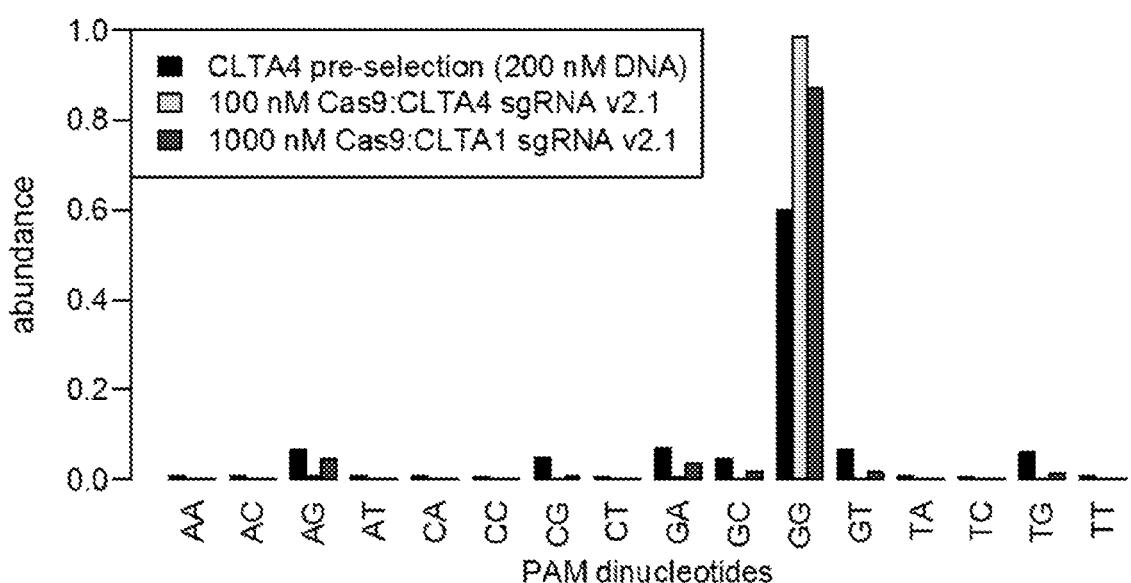


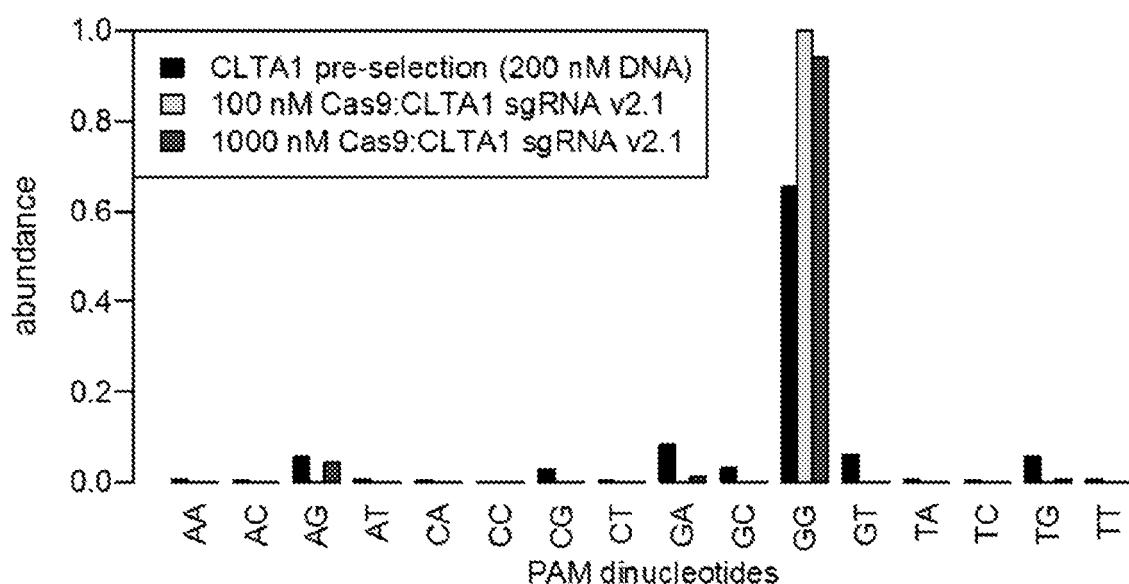
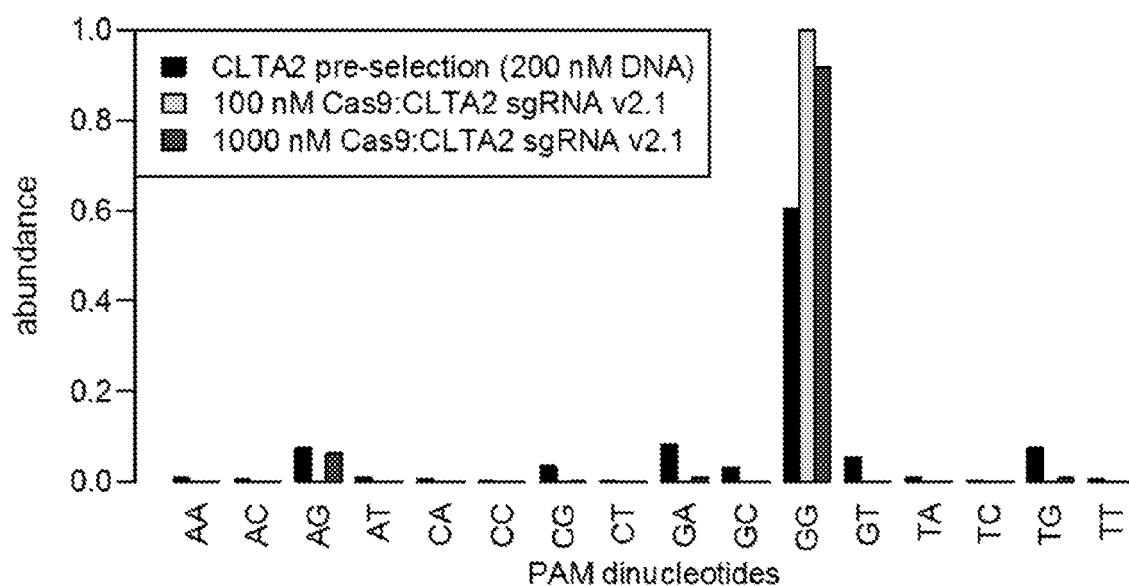
FIG. 19C-D

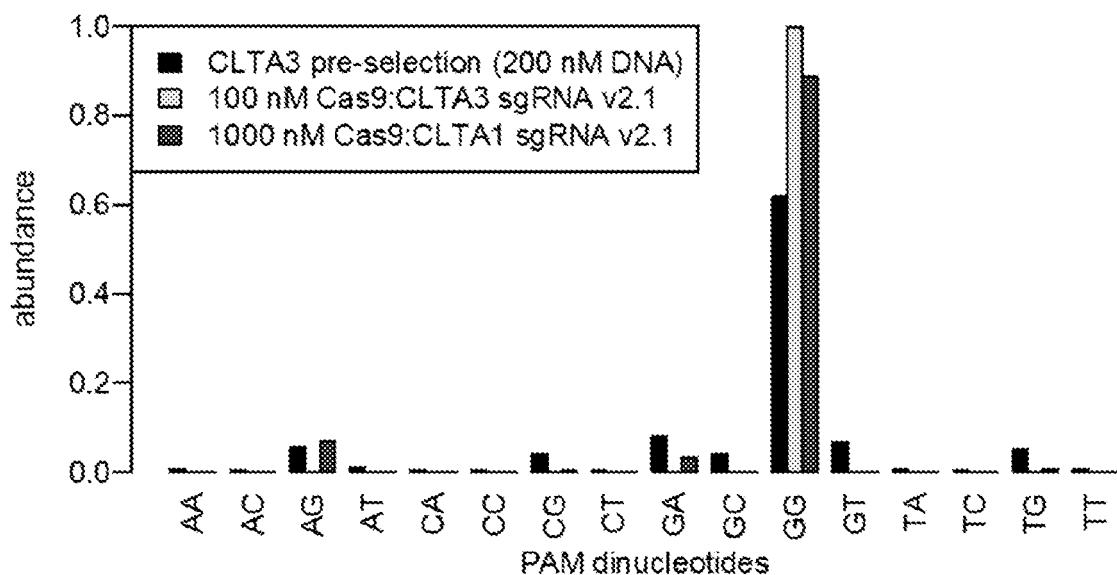
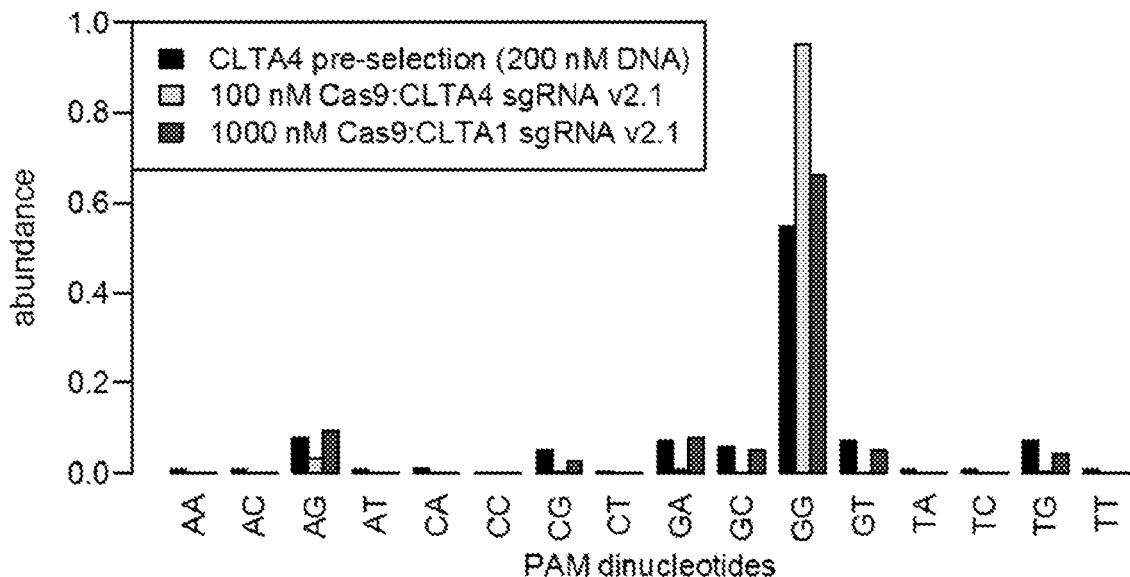
**FIG. 20A-B**

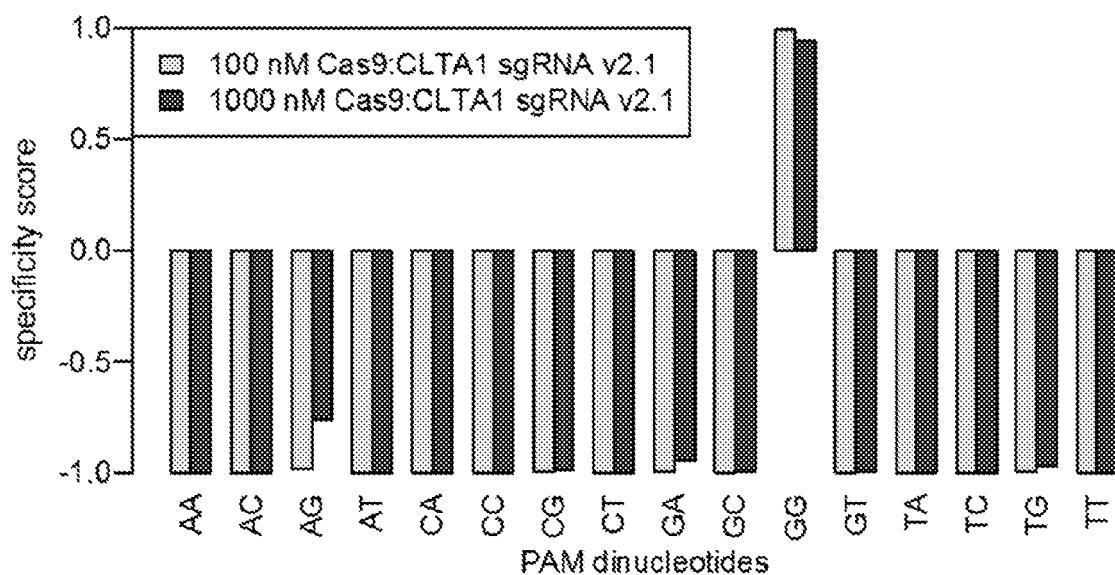
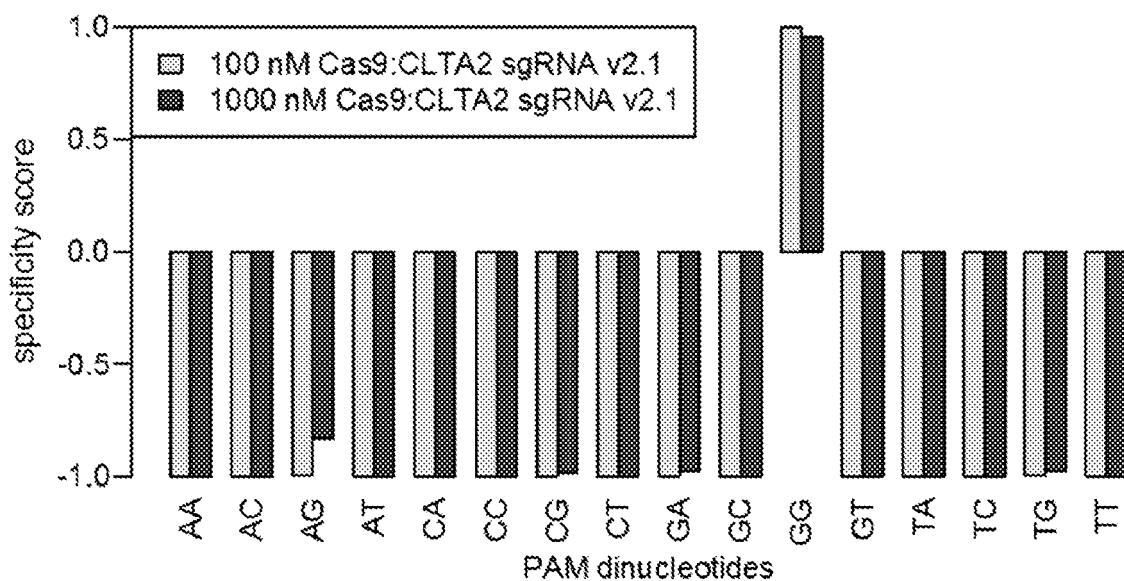
**FIG. 20C-D**

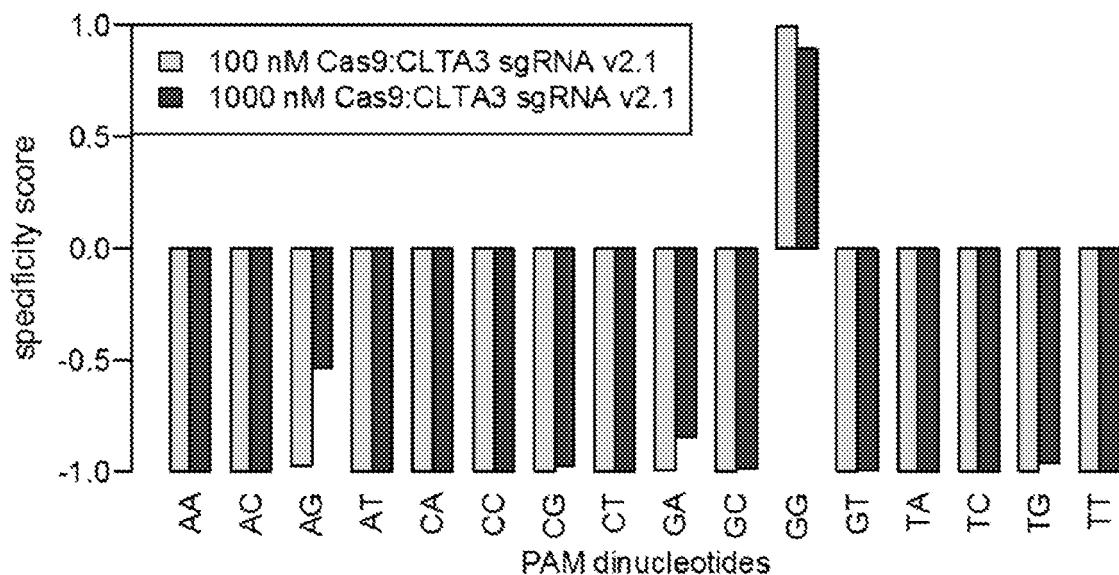
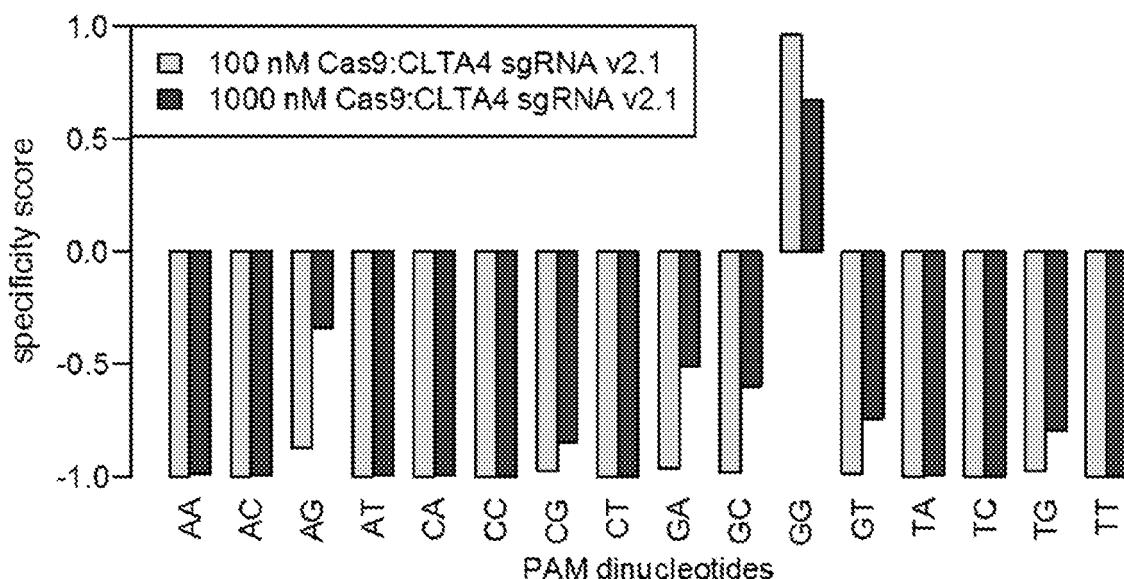
a**b****FIG. 21A-B**

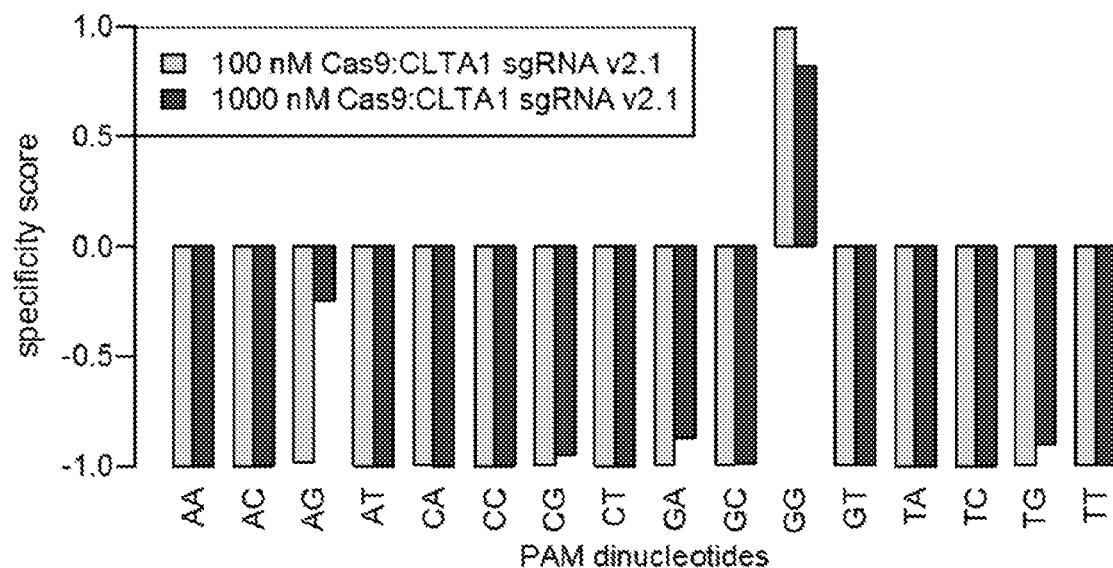
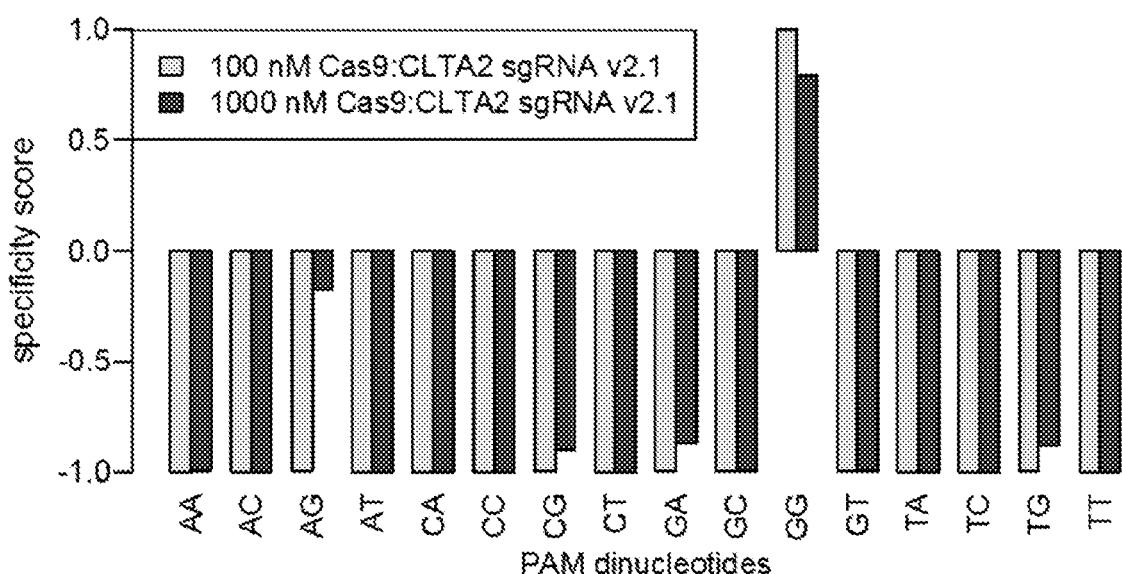
c**d****FIG. 21C-D**

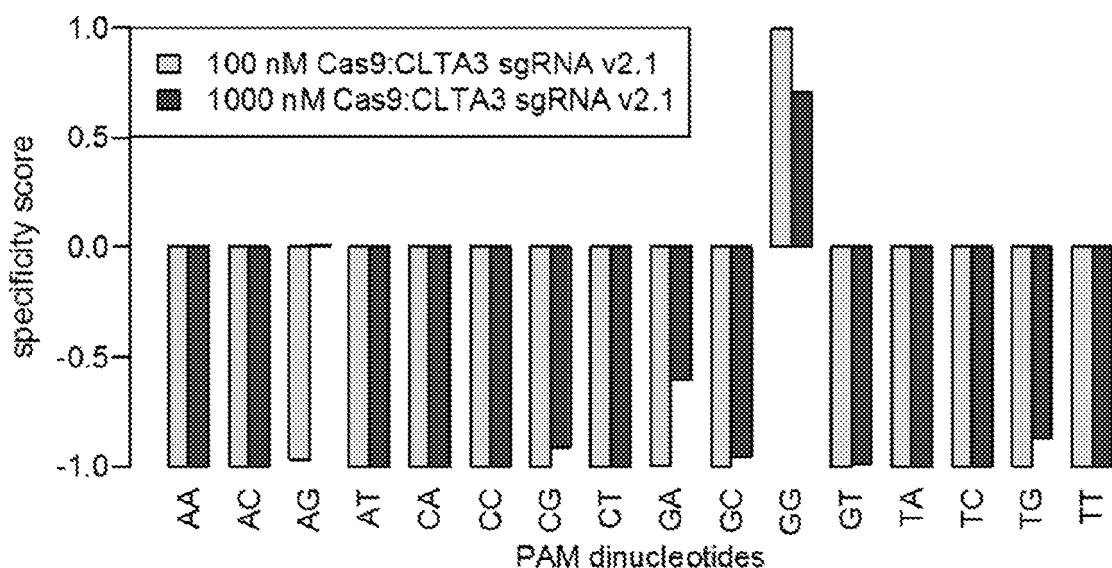
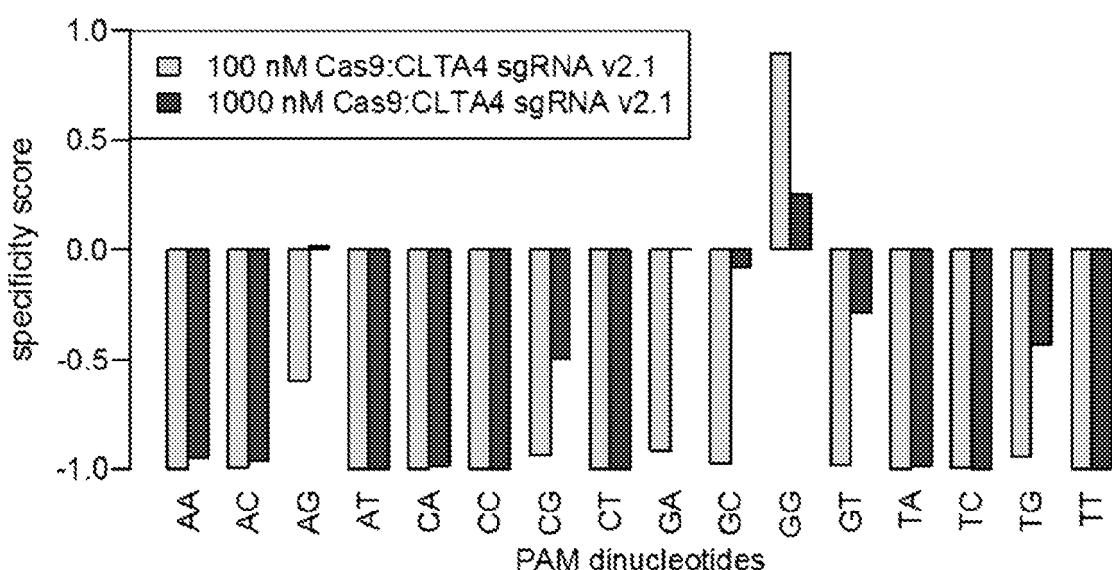
a**b****FIG. 22A-B**

c**d****FIG. 22C-D**

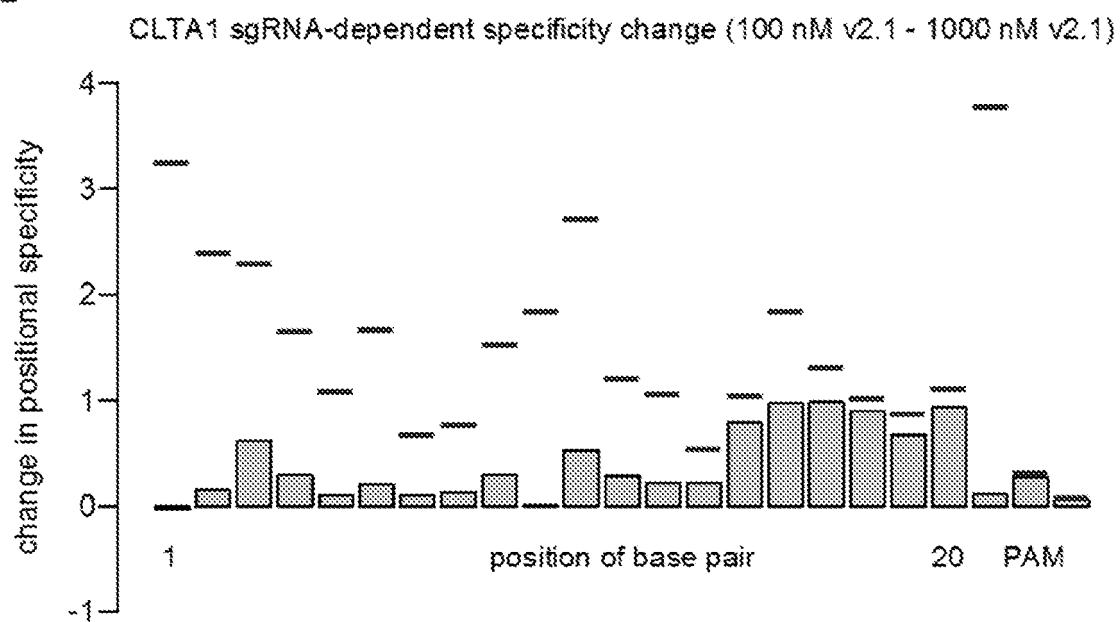
a**b****FIG. 23A-B**

c**d****FIG. 23C-D**

a**b****FIG. 24A-B**

c**d****FIG. 24C-D**

a



b

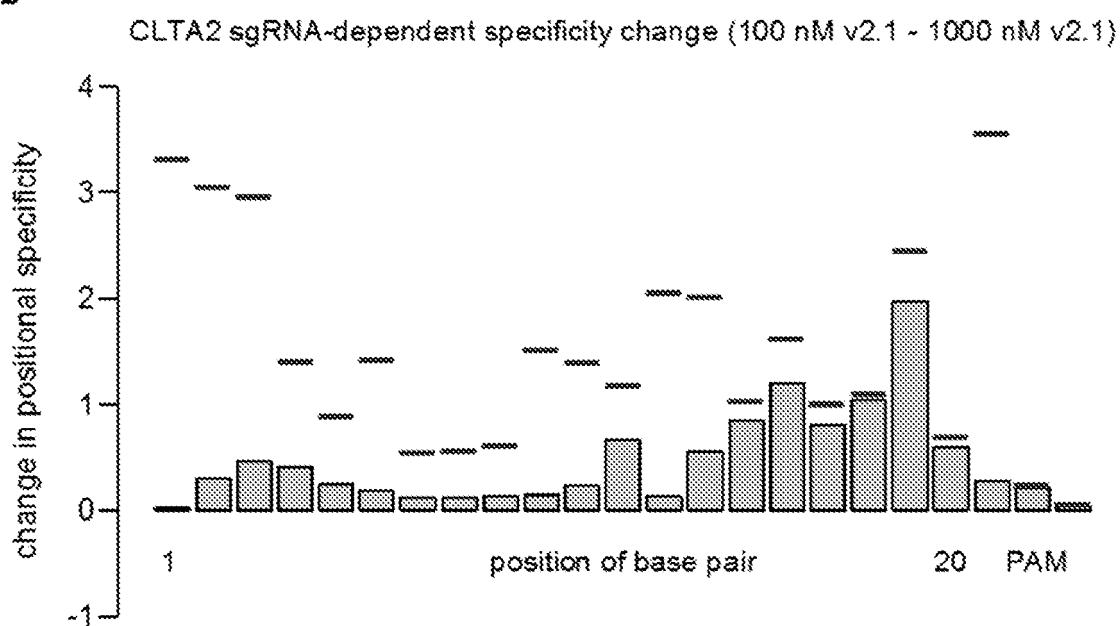


FIG. 25A-B

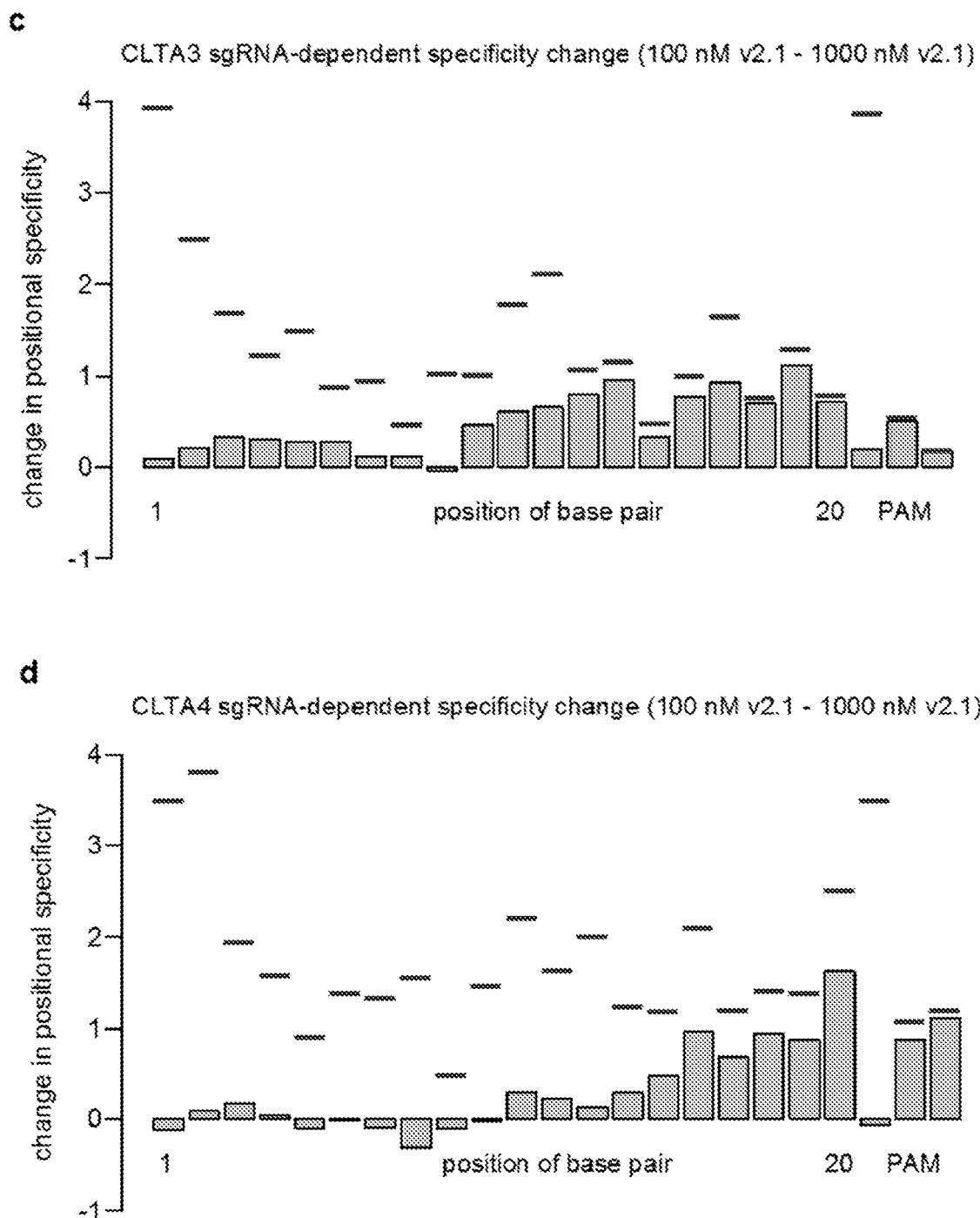


FIG. 25C-D

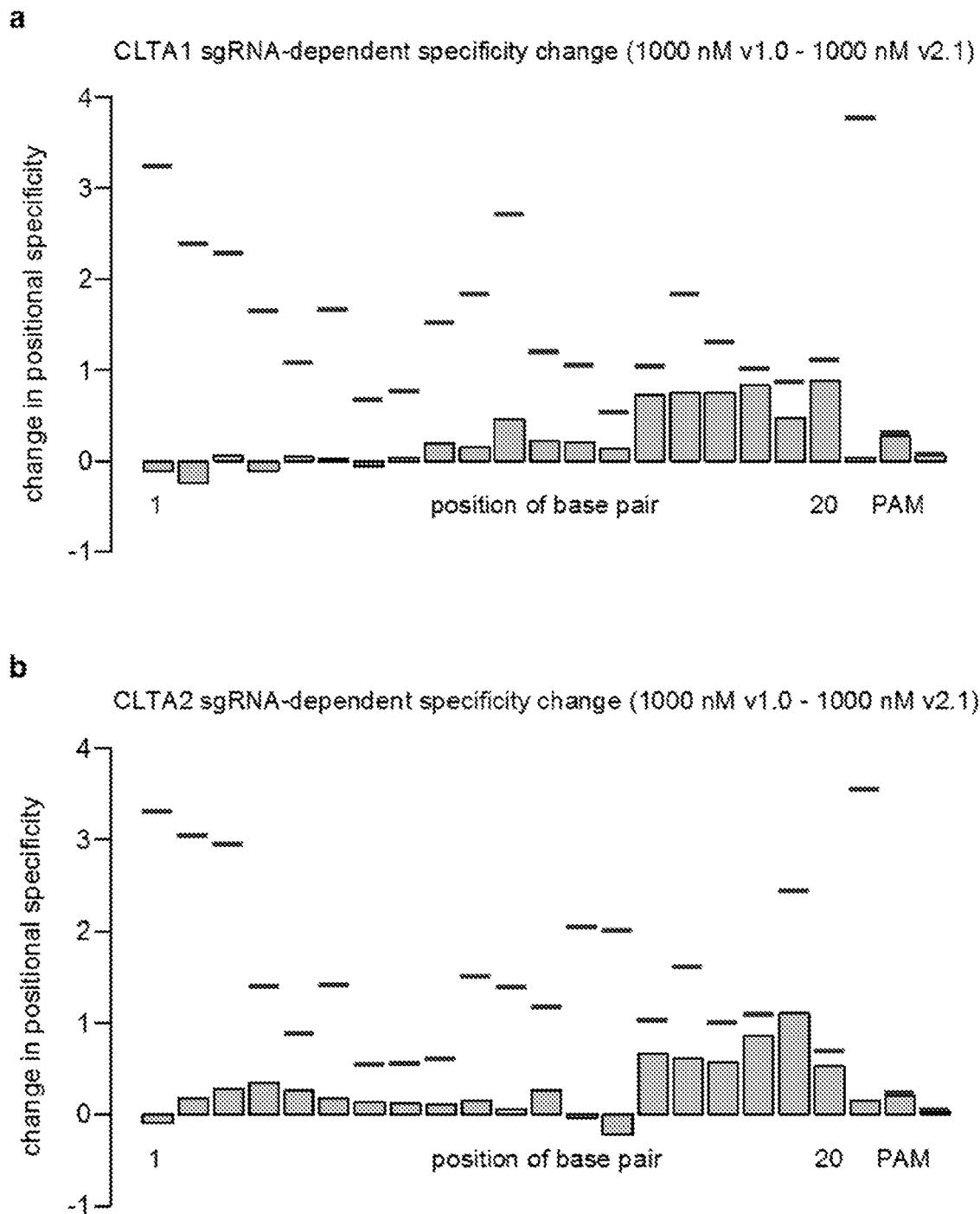
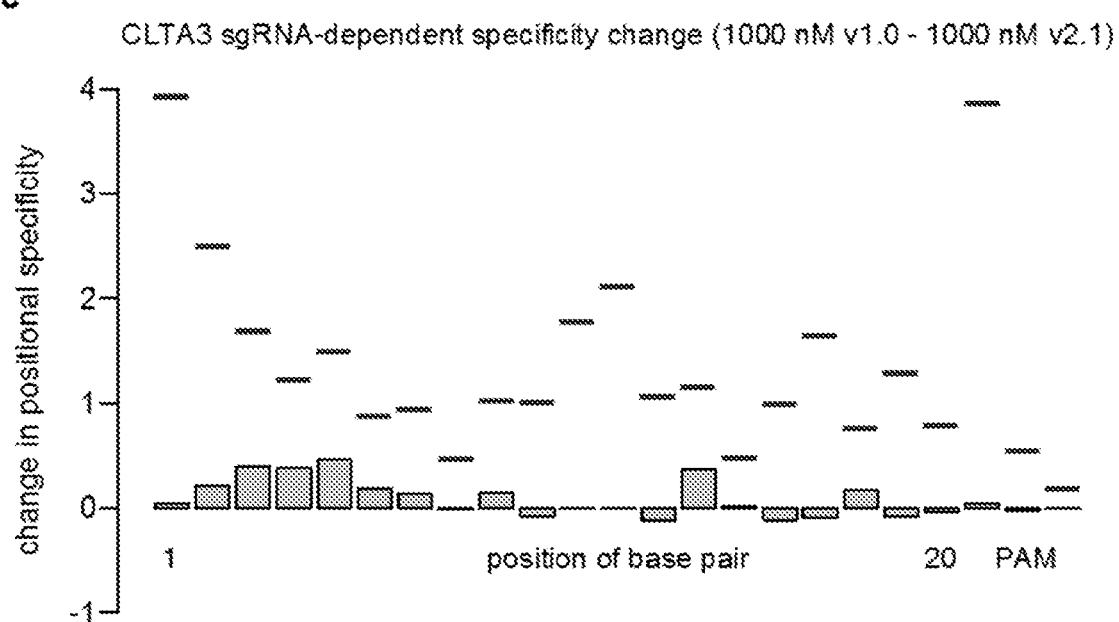
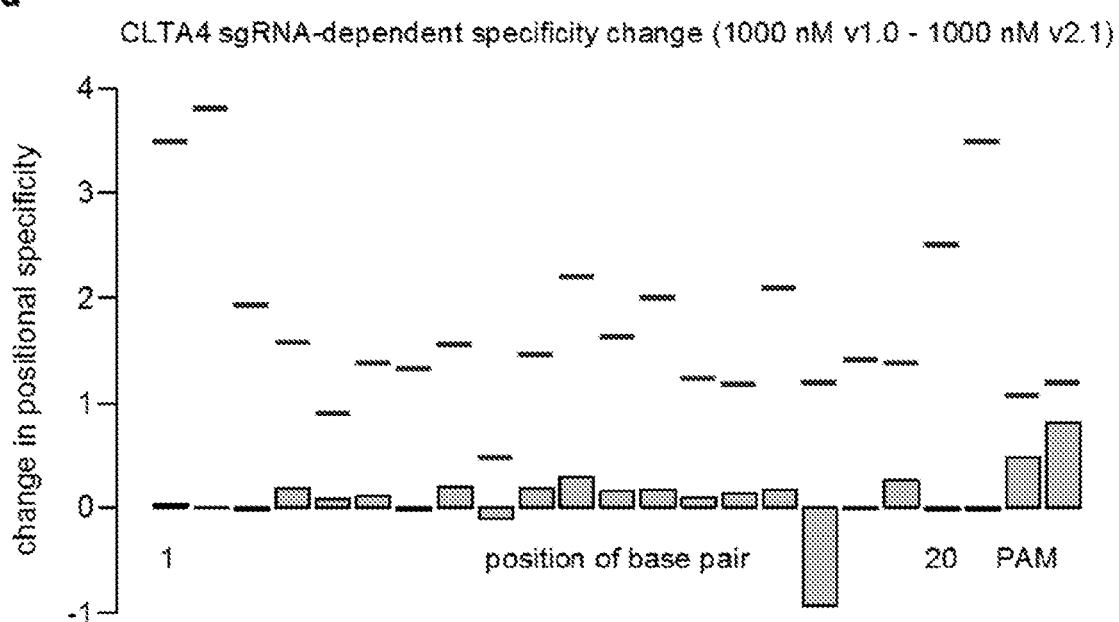
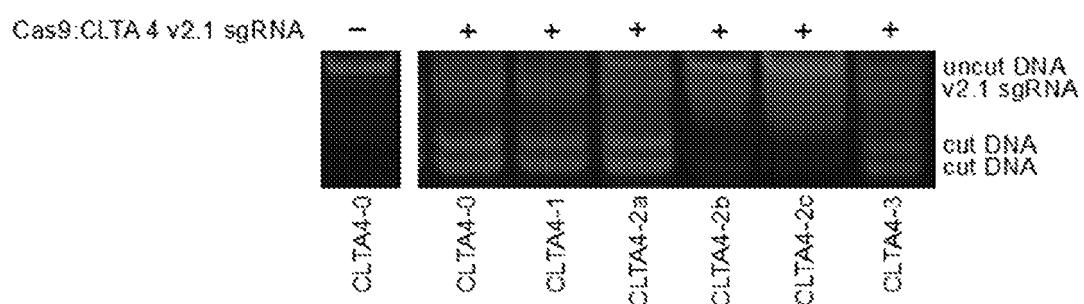


FIG. 26A-B

c**d****FIG. 26C-D**



	sequence	<i>In vitro</i> selection enrichment value	% cut
CLTA4-0	GCAGATGTAGTGTTCACAGGG	7.9	85%
CLTA4-1	GaAGATGTAGTGTTCACAGGG	27.5	84%
CLTA4-2a	GaAGATGTAGTGTTCACtGGG	43.9	79%
CLTA4-2b	GCAGATGgAGgGTTCACAGGG	1.0	36%
CLTA4-2c	GCAGATGTAGTGTtaCCAgAGGG	0.064	none detected
CLTA4-3	GgGATGTAGTGTTCACtGGG	95.9	72%

FIG. 27

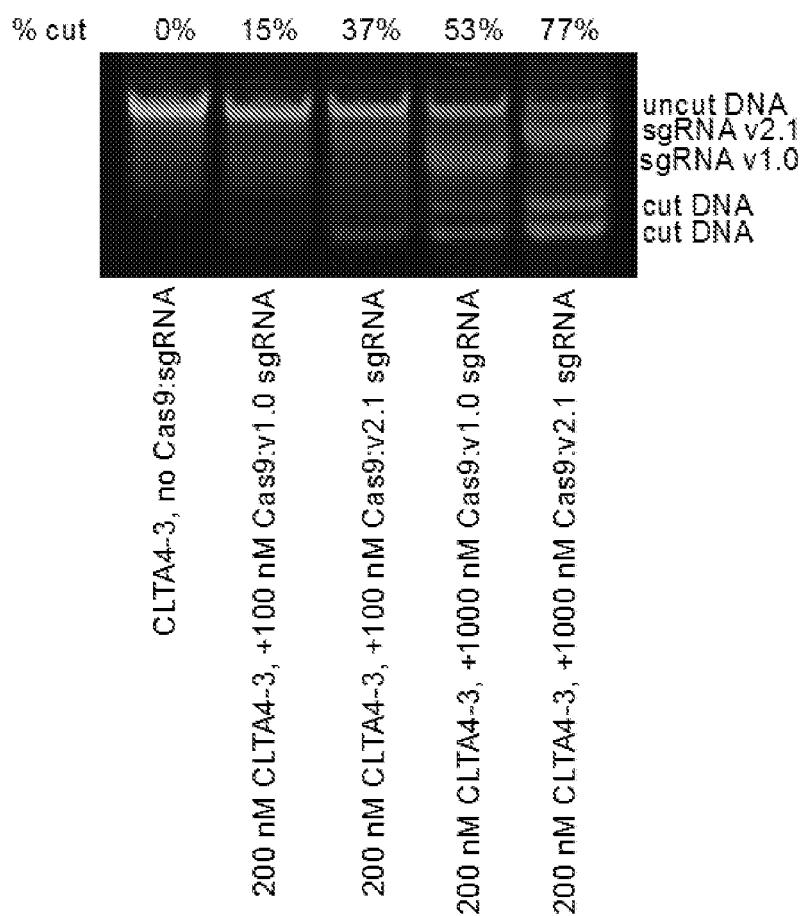


FIG. 28

REFERENCES CITED IN THE DESCRIPTION

This list of references cited by the applicant is for the reader's convenience only. It does not form part of the European patent document. Even though great care has been taken in compiling the references, errors or omissions cannot be excluded and the EPO disclaims all liability in this regard.

Patent documents cited in the description

- WO 2013066438 A2 [0002]
- WO 2013066438 A [0005] [0048]
- US 2010055131 W [0088]
- US 61864289 B [0137]
- US 14320370 B [0137]
- US 14320413 B [0137]

Non-patent literature cited in the description

- PATTANAYAK, V. ; RAMIREZ, C.L. ; JOUNG, J.K. ; LIU, D.R. Revealing off-target cleavage specificities of zinc-finger nucleases by in vitro selection. *Nature methods*, 2011, vol. 8, 765-770 [0005] [0048] [0136]
- JINEK M. ; CHYLINSKI K. ; FONFARA I. ; HAUER M. ; DOUDNA J.A. ; CHARPENTIER E. *Science*, 2012, vol. 337, 816-821 [0019]
- FERRETTI J.J. ; MCSHAN W.M. ; AJDIC D.J. ; SAVIC D.J. ; SAVIC G. ; LYON K. ; PRIMEAUX C. ; SEZATE S. ; SUVOROV A.N. ; KENTON S. Complete genome sequence of an M1 strain of *Streptococcus pyogenes*. *Proc. Natl. Acad. Sci. U.S.A.*, 2001, vol. 98, 4658-4663 [0019] [0034]
- DELTCHEVA E. ; CHYLINSKI K. ; SHARMA C.M. ; GONZALES K. ; CHAO Y. ; PIRZADA Z.A. ; ECKER M.R. ; VOGEL J. ; CHARPENTIER E. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature*, 2011, vol. 471, 602-607 [0019] [0034]
- JINEK M. ; CHYLINSKI K. ; FONFARA I. ; HAUER M. ; DOUDNA J.A. ; CHARPENTIER E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, 2012, vol. 337, 816-821 [0019] [0034]
- CHYLINSKI ; RHUN ; CHARPENTIER. The tracrRNA and Cas9 families of type II CRISPR-Cas immunity systems. *RNA Biology*, 2013, vol. 10 (5), 726-737 [0019]
- K. C. NICOLAOU ; A. L. SMITH ; E. W. YUE. Chemistry and biology of natural and designed enediynes. *PNAS*, 1993, vol. 90 (13), 5881-5888 [0024]
- S. WALKER ; R. LANDOVITZ ; W.D. DING ; G.A. ELLESTAD ; D. KAHNE. Cleavage behavior of calicheamicin gamma 1 and calicheamicin T. *Proc Natl Acad Sci U.S.A.*, 1992, vol. 89 (10), 4608-12 [0024]
- ADRIAN L. SMITH ; K. C. BICOLAOU. The Enediyne Antibiotics. *J. Med. Chem.*, 1996, vol. 39 (11), 2103-2117 [0024]
- DONALD BORDERS. Enediyne antibiotics as anti-tumor agents. *Informa Healthcare*, 23 November 1994 [0024]
- EDGELL DR. Selfish DNA: homing endonucleases find a home. *Curr Biol*, February 2009, vol. 19 (3), R115-R117 [0025]
- JASIN M. Genetic manipulation of genomes with rare-cutting endonucleases. *Trends Genet*, June 1996, vol. 12 (6), 224-8 [0025]
- BURTA ; KOUFOPANOU V. Homing endonuclease genes: the rise and fall and rise again of a selfish element. *Curr Opin Genet Dev*, December 2004, vol. 14 (6), 609-15 [0025]
- CONG, L. et al. Multiplex genome engineering using CRISPR/Cas systems. *Science*, 2013, vol. 339, 819-823 [0035] [0136]
- MALI, P. et al. RNA-guided human genome engineering via Cas9. *Science*, 2013, vol. 339, 823-826 [0035] [0136]
- HWANG, W.Y. et al. Efficient genome editing in zebrafish using a CRISPR-Cas system. *Nature biotechnology*, 2013, vol. 31, 227-229 [0035] [0136]
- JINEK, M. et al. RNA-programmed genome editing in human cells. *eLife*, 2013, vol. 2, e00471 [0035] [0136]
- DICARLO, J.E. et al. Genome engineering in *Saccharomyces cerevisiae* using CRISPR-Cas systems. *Nucleic acids research*, 2013 [0035] [0136]
- JIANG, W. et al. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nature biotechnology*, 2013, vol. 31, 233-239 [0035]
- ESVELT ; WANG. *Molecular Systems Biology*, 2013, vol. 9, 641 [0039]
- MILLER, JEFFREY. A TALE nuclease architecture for efficient genome editing. *Nature Biotechnology*, February 2011, vol. 29 (2), 143-8 [0040]
- ZHANG, FENG. Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription. *Nature Biotechnology*, February 2011, vol. 29 (2), 149-53 [0040] [0041]
- Transcriptional Activators of Human Genes with Programmable DNA-Specificity. GEIBLER, R. ; SCHOLZE, H. ; HAHN, S. ; STREUBEL, J. ; BONNAS, U. ; BEHRENS, S. E. ; BOCH, J. *PLoS ONE*, 2011, vol. 6, e19509 [0040] [0041]

- **BOCH, JENS.** TALEs of genome targeting. *Nature Biotechnology*, February 2011, vol. 29 (2), 135-6 [0040]
- **BOCH, JENS.** Breaking the Code of DNA Binding Specificity of TAL-Type III Effectors. *Science*, December 2009, vol. 326 (5959), 1509-12 [0040]
- **MOSCOU, MATTHEW J. ; ADAM J. BOGDANOVA.** A Simple Cipher Governs DNA Recognition by TAL Effectors. *Science*, December 2009, vol. 326 (5959), 1501 [0040]
- **CERMAK, T. ; DOYLE, E. L. ; CHRISTIAN, M. ; WANG, L. ; ZHANG, Y. ; SCHMIDT, C. ; BALLER, J. A. ; SOMIA, N. V. et al.** Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic Acids Research*, 2011 [0041]
- **MORBITZER, R. ; ELSAESER, J. ; HAUSNER, J. ; LAHAYE, T.** Assembly of custom TALE-type DNA binding domains by modular cloning. *Nucleic Acids Research*, 2011 [0041]
- **LI, T. ; HUANG, S. ; ZHAO, X. ; WRIGHT, D. A. ; CARPENTER, S. ; SPALDING, M. H. ; WEEKS, D. P. ; YANG, B.** Modularly assembled designer TAL effector nucleases for targeted gene knockout and gene replacement in eukaryotes. *Nucleic Acids Research*, 2011 [0041]
- Assembly of Designer TAL Effectors by Golden Gate Cloning. **WEBER, E. ; GRUETZNER, R. ; WERNER, S. ; ENGLER, C. ; MARILLONNET, S.** PLoS ONE. 2011, vol. 6, e19722 [0041]
- **KLUG A ; RHODES D.** Zinc fingers: a novel protein fold for nucleic acid recognition. *Cold Spring Harb. Symp. Quant. Biol.*, 1987, vol. 52, 473-82 [0043]
- **KRISHNA SS ; MAJUMDAR I ; GRISHIN NV.** Structural classification of zinc fingers: survey and summary. *Nucleic Acids Res.*, January 2003, vol. 31 (2), 532-50 [0043]
- **PABO CO ; PEISACH E ; GRANT RA.** Design and selection of novel cys₂His₂ Zinc finger proteins. *Annual Review of Biochemistry*, 2001, vol. 70, 313-340 [0043]
- **JAMIESON AC ; MILLER JC ; PABO CO.** Drug discovery with engineered zinc-finger proteins. *Nature Reviews Drug Discovery*, 2003, vol. 2 (5), 361-368 [0043]
- **LIU Q ; SEGAL DJ ; GHIARA JB ; BARBAS CF.** Design of polydactyl zinc-finger proteins for unique addressing within complex genomes. *Proc. Natl. Acad. Sci. U.S.A.*, May 1997, vol. 94 (11) [0043]
- **PAVLETICH NP ; PABO CO.** Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 Å. *Science*, May 1991, vol. 252 (5007), 809-17 [0044]
- **PATTANAYAK, V. et al.** *Nature Methods*, 2011, vol. 8, 765-770 [0049]
- **A. R. GENNARO.** Remington's The Science and Practice of Pharmacy. Lippincott, Williams & Wilkins, 2006 [0088]
- **HOCKEMEYER, D. et al.** Genetic engineering of human pluripotent cells using TALE nucleases. *Nature biotechnology*, 2011, vol. 29, 731-734 [0136]
- **ZOU, J. et al.** Gene targeting of a disease-related gene in human induced pluripotent stem and embryonic stem cells. *Cell stem cell*, 2009, vol. 5, 97-110 [0136]
- **HOCKEMEYER, D. et al.** Efficient targeting of expressed and silent genes in human ESCs and iPSCs using zinc-finger nucleases. *Nature biotechnology*, 2009, vol. 27, 851-857 [0136]
- **DOYON, Y. et al.** Heritable targeted gene disruption in zebrafish using designed zinc-finger nucleases. *Nature biotechnology*, 2008, vol. 26, 702-708 [0136]
- **MENG, X. ; NOYES, M.B. ; ZHU, L.J. ; LAWSON, N.D. ; WOLFE, S.A.** Targeted gene inactivation in zebrafish using engineered zinc-finger nucleases. *Nature biotechnology*, 2008, vol. 26, 695-701 [0136]
- **SANDER, J.D. et al.** Targeted gene disruption in somatic zebrafish cells using engineered TALENs. *Nature biotechnology*, 2011, vol. 29, 697-698 [0136]
- **TESSON, L. et al.** Knockout rats generated by embryo microinjection of TALENs. *Nature biotechnology*, 2011, vol. 29, 695-696 [0136]
- **CUI, X. et al.** Targeted integration in rat and mouse embryos with zinc-finger nucleases. *Nature biotechnology*, 2011, vol. 29, 64-67 [0136]
- **PEREZ, E.E. et al.** Establishment of HIV-1 resistance in CD4+ T cells by genome editing using zinc-finger nucleases. *Nature biotechnology*, 2008, vol. 26, 808-816 [0136]
- **JINEK, M. et al.** A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, 2012, vol. 337, 816-821 [0136]
- **JIANG, W. ; BIKARD, D. ; COX, D. ; ZHANG, F. ; MARRAFFINI, L.A.** RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nature biotechnology*, 2013, vol. 31, 233-239 [0136]
- **SAPRANAUSKAS, R. et al.** The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli. *Nucleic acids research*, 2011, vol. 39, 9275-9282 [0136]
- **SEMENOVA, E. et al.** Interference by clustered regularly interspaced short palindromic repeat (CRISPR)RNA is governed by a seed sequence. *Proceedings of the National Academy of Sciences of the United States of America*, 2011, vol. 108, 10098-10103 [0136]
- **QI, L.S. et al.** Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression. *Cell*, 2013, vol. 152, 1173-1183 [0136]
- **DOYON, J.B. ; PATTANAYAK, V. ; MEYER, C.B. ; LIU, D.R.** Directed evolution and substrate specificity profile of homing endonuclease I-SceI. *Journal of the American Chemical Society*, 2006, vol. 128, 2477-2484 [0136]

- **SCHNEIDER, T.D. ; STEPHENS, R.M.** Sequence logos: a new way to display consensus sequences.
Nucleic acids research, 1990, vol. 18, 6097-6100
[0136]