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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
20 August 2020 (20.08.2020)

WIPO | PCT

(10) International Publication Number
WO 2020/168102 A1

(51) International Patent Classification:
C12N 9/22 (2006.01) *C12N 15/10* (2006.01)

(21) International Application Number:
PCT/US2020/018145

(22) International Filing Date:
13 February 2020 (13.02.2020)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/806,708 15 February 2019 (15.02.2019) US

(71) Applicant: **SIGMA-ALDRICH CO. LLC** [US/US]; 3050 Spruce Street, St. Louis, Missouri 63103 (US).

(72) Inventor: **CHEN, Fuqiang**; c/o Sigma-Aldrich Co. LLC, 3050 Spruce Street, St. Louis, Missouri 63103 (US).

(74) Agent: **SODEY, Benjamin J.** et al.; Sigma-Aldrich Co. LLC, 3050 Spruce Street, St. Louis, MO 63103 (US).

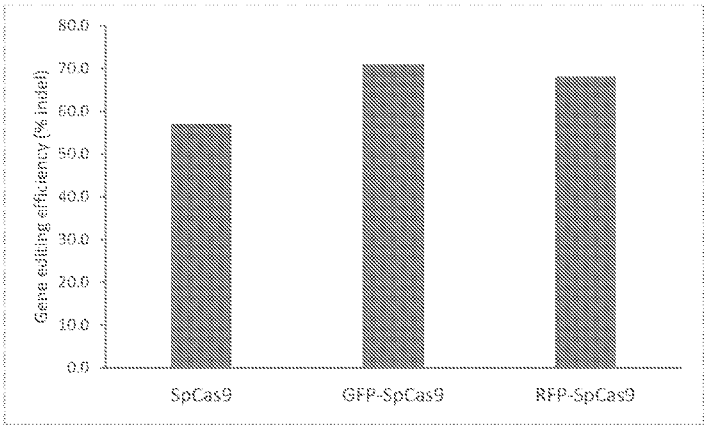
(81) **Designated States** *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR,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

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

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

(54) **Title:** CRISPR/CAS FUSION PROTEINS AND SYSTEMS

FIG. 1



(57) **Abstract:** Engineered Cas9 systems are disclosed herein.

CRISPR/CAS FUSION PROTEINS AND SYSTEMS

RELATED APPLICATIONS

[0001] The present application claims the benefit of priority of U.S. Provisional Application No. 62/806,708, filed February 15, 2019, the entirety of which is incorporated herein by reference.

FIELD

[0002] The present disclosure relates to engineered Cas9 systems, nucleic acids encoding said systems, and methods of using said systems for genome modification.

BACKGROUND

[0003] Many different types of peptide linkers have been tested to fuse GFP to Cas9, but typically result in lower activity of the underlying Cas9.

SUMMARY OF THE DISCLOSURE

[0004] Among the various aspects of the present disclosure include engineered Cas9 systems.

[0005] Other aspects and features of the disclosure are detailed bellow.

BRIEF DESCRIPTION OF THE DRAWINGS

[0006] **FIG. 1** shows that the Cas9 fusion proteins disclosed herein each retain the editing activity parallel to the level of SpCas9 protein.

[0007] **FIG. 2A** and **FIG. 2B** show that the editing efficiencies of the Cas9 fusion proteins disclosed herein were several-fold higher than that of the commercial proteins in all targets.

SEQUENCE LISTING

[0008] The instant application contains a Sequence Listing which has been submitted electronically in ASCII format and is hereby incorporated by reference in its entirety. Said ASCII copy, created on February 13, 2020, is named P19-027_WO-PCT_SL.txt and is 87,735 bytes in size.

DETAILED DESCRIPTION

[0009] Fusion of accessory proteins to CRISPR proteins creates a wide range of opportunities to localize various protein functionalities to defined locations within cells. Among other things, peptide linkers which enable fusion of heterologous proteins to CRISPR proteins in ways that preserve CRISPR functionality are disclosed.

(I) Engineered Cas9 Systems

[0010] One aspect of the present disclosure provides engineered Cas9 proteins and systems. For example, Cas9-marker fusion proteins are disclosed. In some aspects, systems include engineered Cas9 proteins and engineered guide RNAs, wherein each engineered guide RNA is designed to complex with a specific engineered Cas9 protein. These engineered Cas9 systems do not occur naturally.

(a) Engineered Cas9 Proteins

[0011] Cas9 protein is the single effector protein in Type II CRISPR systems, which are present in various bacteria. The engineered Cas9 protein disclosed herein can be from *Acaryochloris sp., Acetohalobium sp., Acidaminococcus sp., Acidithiobacillus sp., Acidothermus* sp., *Akkermansia sp., Alicyclobacillus sp., Allochromatium sp., Ammonifex sp., Anabaena sp., Arthrospira sp., Bacillus sp., Bifidobacterium sp., Burkholderiales sp., Caldicelulosiruptor sp., Campylobacter sp.*, *Candidatus sp., Clostridium sp., Corynebacterium sp., Crocosphaera sp., Cyanothece sp., Exiguobacterium sp., Fibrobacter sp., Finegoldia sp., Francisella sp.*, *Ktedonobacter sp., Lachnospiraceae sp., Lactobacillus sp., Listeria sp., Lyngbya sp., Marinobacter sp., Methanohalobium sp., Microscilla sp., Microcoleus sp., Microcystis sp., Mycoplasma sp., Natranaerobius sp., Neisseria sp., Nitratifractor sp., Nitrosococcus sp., Nocardiopsis sp., Nodularia sp., Nostoc sp., Oenococcus sp., Oscillatoria sp., Parasutterella sp., Pasteurella sp., Parvibaculum sp., Pelotomaculum sp., Petrotoga sp., Polaromonas sp., Prevotella sp., Pseudoalteromonas sp., Ralstonia sp., Rhodospirillum sp., Staphylococcus sp., Streptococcus sp., Streptomyces sp., Streptosporangium sp., Synechococcus sp., Thermosipho sp.*, *Treponema sp., Verrucomicrobia* sp., and *Wolinella sp.*

[0012] Exemplary species that the Cas9 protein or other components may be from or derived from include *Acaryochloris* spp. (*e.g., Acaryochloris marina*), *Acetohalobium* spp. (*e.g.*, *Acetohalobium arabaticum*), *Acidaminococcus* spp., *Acidithiobacillus* spp. (*e.g.*, *Acidithiobacillus caldus, Acidithiobacillus ferrooxidans*), *Acidothermus* spp., *Akkermansia* spp., *Alicyclobacillus* spp. (*e.g.*, *Alicyclobacillus acidocaldarius*), *Allochromatium* spp. (*e.g.*, *Allochromatium vinosum*), *Ammonifex* spp. (*e.g.*, *Ammonifex degensii*), *Anabaena* spp. (*e.g.*, *Anabaena variabilis*), *Arthrospira* spp. (*e.g.*, *Arthrospira maxima, Arthrospira platensis*), *Bacillus* spp. (*e.g.*, *Bacillus pseudomycoides, Bacillus selenitireducens*), *Bifidobacterium* spp., *Burkholderiales* spp. (*e.g.*, *Burkholderiales bacterium*), *Caldicelulosiruptor spp.* (*e.g.*, *Caldicelulosiruptor becscii*), *Campylobacter* spp. (*e.g.*, *Campylobacter jejuni, Campylobacter lari*), *Candidatus* spp., (*e.g.*, *Candidatus Desulforudis*), *Clostridium* spp. (*e.g.*, *Clostridium botulinum, Clostridium difficile*), *Corynebacterium* spp. (*e.g.*, *Corynebacterium diphtheria*), *Crocosphaera* spp. (*e.g.*, *Crocosphaera watsonii*), *Cyanothece* spp., *Deltaproteobacterium* spp., *Exiguobacterium* spp. (*e.g.*, *Exiguobacterium sibiricum*), (*Fibrobacter* spp. (*e.g.*, *Fibrobacter succinogene*), *Finegoldia* spp. (*e.g.*, *Finegoldia magna*), *Francisella* spp. (*e.g.*, *Francisella novicida*), *Gammaproteobacterium, Ktedonobacter* spp. (*e.g.*, *Ktedonobacter racemifer*), *Lachnospiraceae* spp., *Lactobacillus* spp. (*e.g.*, *Lactobacillus buchneri, Lactobacillus delbrueckii, Lactobacillus gasseri, Lactobacillus salivarius*), *Listeria* spp. (e.g., *Listeria innocua*), *Leptotrichia* spp., *Lyngbya* spp., *Marinobacter* spp., *Methanohalobium* spp. (*e.g.*, *Methanohalobium evestigatum*), *Microcoleus* spp. (*e.g.*, *Microcoleus chthonoplastes*), *Microscilla* spp. (*e.g.*, *Microscilla marina*), *Microcystis* spp. (*e.g.*, *Microcystis aeruginosa*), *Mycoplasma* spp., *Natranaerobius* spp. (*e.g.*, *Natranaerobius thermophilus*), *Neisseria* spp. (*e.g.*, *Neisseria cinerea, Neisseria meningitidis*), *Nitratifractor* spp., *Nitrosococcus* spp. (*e.g.*, *Nitrosococcus halophilus, Nitrosococcus watsoni*), *Nocardiopsis* spp. (*e.g.*, *Nocardiopsis dassonvillei*), *Nodularia* spp. (*e.g.*, *Nodularia spumigena*), *Nostoc* spp., *Oenococcus* spp., *Oscillatoria* spp., *Parasutterella* spp., *Parvibaculum* spp. (*e.g.*, *Parvibaculum lavamentivorans*), *Pasteurella* spp. (*e.g.*, *Pasteurella multocida*), *Pelotomaculum* spp., (*e.g.*, *Pelotomaculum thermopropionicum*), *Petrotoga* spp. (*e.g.*, *Petrotoga mobilis*), *Planctomyces spp., Polaromonas* spp. (*e.g., Polaromonas naphthalenivorans*), *Prevotella* spp., *Pseudoalteromonas* spp. (*e.g., Pseudoalteromonas haloplanktis*), *Ralstonia* spp., *Ruminococcus* spp., *Rhodospirillum* spp. (*e.g., Rhodospirillum rubrum*), *Staphylococcus* spp. (*e.g., Staphylococcus aureus*), *Streptococcus* spp. (*e.g., Streptococcus pasteurianus, Streptococcus pyogenes, Streptococcus thermophilus*), *Sutterella* spp. (*e.g., Sutterella wadsworthensis*), *Streptomyces* spp. (*e.g., Streptomyces pristinaespiralis, Streptomyces viridochromogenes, Streptomyces viridochromogenes*), *Streptosporangium* spp. (*e.g.,Streptosporangium roseum, Streptosporangium roseum*), *Synechococcus* spp., *Thermosipho* spp. (*e.g., Thermosipho africanus*), *Treponema* spp. (*e.g., Treponema denticola*), and *Verrucomicrobia* spp., *Wolinella* spp. (*e.g., Wolinella succinogenes*), and/or species delineated in bioinformatic surveys of genomic databases such as those disclosed in Makarova, Kira S., et al. "An updated evolutionary classification of CRISPR–Cas systems." Nature Reviews Microbiology 13.11 (2015): 722 and Koonin, Eugene V., Kira S. Makarova, and Feng Zhang. "Diversity, classification and evolution of CRISPR-Cas systems." Current opinion in microbiology 37 (2017): 67-78, each of which is hereby incorporated by reference herein in their entirety.

[0013] In some embodiments, the engineered Cas9 protein may be from *Streptococcus pyogenes*. In some embodiments, the engineered Cas9 protein may be from *Streptococcus thermophilus*. In some embodiments, the engineered Cas9 protein may be from *Neisseria meningitidis*. In some embodiments, the engineered Cas9 protein may be from *Staphylococcus aureus*. In some embodiments, the engineered Cas9 protein may be from *Campylobacter jejuni*.

[0014] Wild-type Cas9 proteins comprise two nuclease domains, *i.e.*, RuvC and HNH domains, each of which cleaves one strand of a double-stranded sequence. Cas9 proteins also comprise REC domains that interact with the guide RNA (*e.g.*, REC1, REC2) or the RNA/DNA heteroduplex (*e.g.*, REC3), and a domain that interacts with the protospacer-adjacent motif (PAM) (*i.e.*, PAM-interacting domain).

[0015] The Cas9 protein can be engineered to comprise one or more modifications (*i.e.*, a substitution of at least one amino acid, a deletion of at least one amino acid, an insertion of at least one amino acid) such that the Cas9 protein has altered activity, specificity, and/or stability.

[0016] For example, Cas9 protein can be engineered by one or more mutations and/or deletions to inactivate one or both of the nuclease domains. Inactivation of one nuclease domain generates a Cas9 protein that cleaves one strand of a double-stranded sequence (*i.e.*, a Cas9 nickase). The RuvC domain can be inactivated by mutations such as D10A, D8A, E762A, and/or D986A, and the HNH domain can be inactivated by mutations such as H840A, H559A, N854A, N856A, and/or N863A (with reference to the numbering system of *Streptococcus pyogenes* Cas9, SpyCas9). Inactivation of both nuclease domains generates a Cas9 protein having no cleavage activity (*i.e.*, a catalytically inactive or dead Cas9).

[0017] The Cas9 protein can also be engineered by one or more amino acid substitutions, deletions, and/or insertions to have improved targeting specificity, improved fidelity, altered PAM specificity, decreased off-target effects, and/or increased stability. Non-limiting examples of one or more mutations that improve targeting specificity, improve fidelity, and/or decrease off-target effects include N497A, R661A, Q695A, K810A, K848A, K855A, Q926A, K1003A, R1060A, and/or D1135E (with reference to the numbering system of SpyCas9).

[0018] In alternative embodiments, the Cas protein may be from a Type I CRISPR/Cas system. In some embodiments, the Cas protein may be a component of the Cascade complex of a Type-I CRISPR/Cas system. For example, the Cas protein may be a Cas3 protein. In some embodiments, the Cas protein may be from a Type III CRISPR/Cas system. In some embodiments, the Cas protein may be from a Type IV CRISPR/Cas system. In some embodiments, the Cas protein may be from a Type V CRISPR/Cas system. In some embodiments, the Cas protein may be from a Type VI CRISPR/Cas system. In some embodiments, the Cas protein may have an RNA cleavage activity. In various embodiments, the Cas protein may be classified as Cas9, Cas12a (a.k.a. Cpf1), Cas12b, Cas12c, Cas12d, Cas12e (a.k.a. CasX), Cas13a, or Cas13b.

(i) Heterologous domains

[0019] The Cas9 protein can be engineered to comprise at least one heterologous domain, *i.e.*, Cas9 is fused to one or more heterologous domains. In situations in which two or more heterologous domains are fused with Cas9, the two or more heterologous domains can be the same or they can be different. The one or more heterologous domains can be fused to the N terminal end, the C terminal end, an internal location, or combination thereof. The fusion can be direct via a chemical bond, or the linkage can be indirect via one or more linkers.

[0020] In certain preferred embodiments, the engineered Cas9 proteins described herein include one or more nuclear localization signals (NLS). Non-limiting examples of nuclear localization signals include PKKKRKV (SEQ ID NO:1), PKKKRRV (SEQ ID NO:2), KRPAATKKAGQAKKKK (SEQ ID NO:3), YGRKKRRQRRR (SEQ ID NO:4), RKKRRQRRR (SEQ ID NO:5), PAAKRVKLD (SEQ ID NO:6), RQRRNELKRSP (SEQ ID NO:7), VSRKRPRP (SEQ ID NO:8), PPKKARED (SEQ ID NO:9), PQPKKKPL (SEQ ID NO:10), SALIKKKKKMAP (SEQ ID NO:11), PKQKKRK (SEQ ID NO:12), RKLKKKIKKL (SEQ ID NO:13), REKKKFLKRR (SEQ ID NO:14), KRKGDEVDGVDEVAKKKSKK (SEQ ID NO:15), RKCLQAGMNLEARKTKK (SEQ ID NO:16), NQSSNFGPMKGGNFGGRSSGPYGGGGQYFAKPRNQGGY (SEQ ID NO:17), and RMRIZFKNKGKDTAELRRRRVEVSVELRKAKKDEQILKRRNV (SEQ ID NO:18).

[0021] In one particular embodiment, the nuclear localization signal is selected from PKKKRKV (SEQ ID NO:1) and PAAKRVKLD (SEQ ID NO:6). In another particular embodiment, the engineered Cas9 protein includes both of PKKKRKV (SEQ ID NO:1) and PAAKRVKLD (SEQ ID NO:6). In another particular embodiment, the engineered Cas9 protein includes at least two of PKKKRKV (SEQ ID NO:1) and at least one of PAAKRVKLD (SEQ ID NO:6). In another particular embodiment, the engineered Cas9 protein includes two of PKKKRKV (SEQ ID NO:1) and one of PAAKRVKLD (SEQ ID NO:6).

[0022] In these and other preferred embodiments, the engineered Cas9 proteins include one or more marker domains. Marker domains include fluorescent proteins and purification or epitope tags. Suitable fluorescent proteins include, without limit, green fluorescent proteins (*e.g.*, GFP, eGFP, GFP-2, tagGFP, turboGFP, Emerald, Azami Green, Monomeric Azami Green, CopGFP, AceGFP, ZsGreen1), yellow fluorescent proteins (*e.g.*, YFP, EYFP, Citrine, Venus, YPet, PhiYFP, ZsYellow1), blue fluorescent proteins (*e.g.*, BFP, EBFP, EBFP2, Azurite, mKalama1, GFPuv, Sapphire, T-sapphire), cyan fluorescent proteins (*e.g.,* ECFP, Cerulean, CyPet, AmCyan1, Midoriishi-Cyan), red fluorescent proteins (*e.g.*, mKate, mKate2, mPlum, DsRed monomer, mCherry, mRFP1, DsRed-Express, DsRed2, DsRed-Monomer, HcRed-Tandem, HcRed1, AsRed2, eqFP611, mRasberry, mStrawberry, Jred), orange fluorescent proteins (*e.g.*, mOrange, mKO, Kusabira-Orange, Monomeric Kusabira-Orange, mTangerine, tdTomato), or combinations thereof. The marker domain can comprise tandem repeats of one or more fluorescent proteins (*e.g.*, Suntag).

[0023] In one embodiment, the marker protein is selected from the following:

Marker Protein Sequence
MVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPWPTLVTTLTYGVQCFSRYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSKLSKDPNEKRDHMVLLEFVTAAGITLGMDELYK (SEQ ID NO:19) MVSKGEAVIKEFMRFKVHMEGSMNGHEFEIEGEGEGRPYEGTQTAKLKVTKGGPLPFSWDILSPQFMYGSRAFTKHPADIPDYYKQSFPEGFKWERVMNFEDGGAVTVTQDTSLEDGTLIYKVKLRGTNFPPDGPVMQKKTMGWEASTERLYPEDGVLKGDIKMALRLKDGGRYLADFKTTYKAKKPVQMPGAYNVDRKLDITSHNEDYTVVEQYERSEGRHSTGGMDELYK (SEQ ID NO:20)

[0024] Non-limiting examples of suitable purification or epitope tags include 6xHis (SEQ ID NO:22), FLAG® (e.g., SEQ ID NO:21), HA, GST, Myc, SAM, and the like. Non-limiting examples of heterologous fusions which facilitate detection or enrichment of CRISPR complexes include streptavidin (Kipriyanov et al., Human Antibodies, 1995, 6(3):93-101.), avidin (Airenne et al., Biomolecular Engineering, 1999, 16(1-4):87-92), monomeric forms of avidin (Laitinen et al., Journal of Biological Chemistry, 2003, 278(6):4010-

4014), peptide tags which facilitate biotinylation during recombinant production (Cull et al., Methods in Enzymology, 2000, 326:430-440).

[0025] In addition to a nuclear localization signal(s) and a marker protein(s), in various embodiments the engineered Cas9 protein may also include one or more heterologous domains such as a cell-penetrating domain, a marker domain, a chromatin disrupting domain, an epigenetic modification domain (*e.g.*, a cytidine deaminase domain, a histone acetyltransferase domain, and the like), a transcriptional regulation domain, an RNA aptamer binding domain, or a non-Cas9 nuclease domain.

[0026] In some embodiments, the one or more heterologous domains can be a cell-penetrating domain. Examples of suitable cell-penetrating domains include, without limit, GRKKRRQRRRPPQPKKKRKV (SEQ ID NO:23), PLSSIFSRIGDPPKKKRKV (SEQ ID NO:24), GALFLGWLGAAGSTMGAPKKKRKV (SEQ ID NO:25), GALFLGFLGAAGSTMGAWSQPKKKRKV (SEQ ID NO:26), KETWWETWWTEWSQPKKKRKV (SEQ ID NO:27), YARAAARQARA (SEQ ID NO:28), THRLPRRRRRR (SEQ ID NO:29), GGRRARRRRRR (SEQ ID NO:30), RRQRRTSKLMKR (SEQ ID NO:31), GWTLNSAGYLLGKINLKALAALAKKIL (SEQ ID NO:32), KALAWEAKLAKALAKALAKHLAKALAKALKCEA (SEQ ID NO:33), and RQIKIWFQNRRMKWKK (SEQ ID NO:34).

[0027] In still other embodiments, the one or more heterologous domain can be a chromatin modulating motif (CMM). Non-limiting examples of CMMs include nucleosome interacting peptides derived from high mobility group (HMG) proteins (*e.g.*, HMGB1, HMGB2, HMGB3, HMGN1, HMGN2, HMGN3a, HMGN3b, HMGN4, and HMGN5 proteins), the central globular domain of histone H1 variants (*e.g.*, histone H1.0, H1.1, H1.2, H1.3, H1.4, H1.5, H1.6, H1.7, H1.8, H1.9, and H.1.10), or DNA binding domains of chromatin remodeling complexes (*e.g.*, SWI/SNF (SWItch/Sucrose Non-Fermentable), ISWI (Imitation SWItch), CHD (Chromodomain-Helicase-DNA binding), Mi-2/NuRD (Nucleosome Remodeling and Deacetylase), INO80, SWR1, and RSC complexes. In other embodiments, CMMs also can be derived from topoisomerases, helicases, or viral proteins. The source of the CMM can and will vary. CMMs can be from humans, animals (*i.e.*, vertebrates and invertebrates), plants, algae, or yeast. Non-limiting examples of specific CMMs are listed in the table below. Persons of skill in the art can readily identify homologs in other species and/or the relevant fusion motif therein.

Protein	**Accession No.**	**Fusion Motif**
Human HMGN1	P05114	Full length
Human HMGN2	P05204	Full length
Human HMGN3a	Q15651	Full length
Human HMGN3b	Q15651-2	Full length
Human HMGN4	O00479	Full length
Human HMGN5	P82970	Nucleosome binding motif
Human HMGB1	P09429	Box A
Human histone H1.0	P07305	Globular motif
Human histone H1.2	P16403	Globular motif
Human CHD1	O14646	DNA binding motif
Yeast CHD1	P32657	DNA binding motif
Yeast ISWI	P38144	DNA binding motif
Human TOP1	P11387	DNA binding motif
Human herpesvirus 8 LANA	J9QSF0	Nucleosome binding motif
Human CMV IE1	P13202	Chromatin tethering motif
M. leprae DNA helicase	P40832	HhH binding motif

[0028] In yet other embodiments, the one or more heterologous domains can be an epigenetic modification domain. Non-limiting examples of suitable epigenetic modification domains include those with DNA deamination (*e.g.*, cytidine deaminase, adenosine deaminase, guanine deaminase), DNA methyltransferase activity (*e.g.*, cytosine methyltransferase), DNA demethylase activity, DNA amination, DNA oxidation activity, DNA helicase activity, histone acetyltransferase (HAT) activity (*e.g.*, HAT domain derived from E1A binding protein p300), histone deacetylase activity, histone methyltransferase activity, histone demethylase activity, histone kinase activity, histone phosphatase activity, histone ubiquitin ligase activity, histone deubiquitinating activity, histone adenylation activity, histone deadenylation activity, histone SUMOylating activity, histone deSUMOylating activity, histone ribosylation activity, histone deribosylation activity, histone myristoylation activity, histone demyristoylation activity, histone citrullination activity, histone alkylation activity, histone dealkylation activity, or histone oxidation activity. In specific embodiments, the epigenetic modification domain can comprise cytidine deaminase activity, adenosine deaminase activity, histone acetyltransferase activity, or DNA methyltransferase activity.

[0029] In other embodiments, the one or more heterologous domains can be a transcriptional regulation domain (*i.e.*, a transcriptional activation domain or transcriptional repressor domain). Suitable transcriptional activation domains include, without limit, herpes simplex virus VP16 domain, VP64 (*i.e.*, four tandem copies of VP16), VP160 (*i.e.*, ten tandem copies of VP16), NFκB p65 activation domain (p65) , Epstein-Barr virus R transactivator (Rta) domain, VPR (*i.e.*, VP64+p65+Rta), p300-dependent transcriptional activation domains, p53 activation domains 1 and 2, heat-shock factor 1 (HSF1) activation domains, Smad4 activation domains (SAD), cAMP response element binding protein (CREB) activation domains, E2A activation domains, nuclear factor of activated T-cells (NFAT) activation domains, or combinations thereof. Non-limiting examples of suitable transcriptional repressor domains include Kruppel-associated box (KRAB) repressor domains, Mxi repressor domains, inducible cAMP early repressor (ICER) domains, YY1 glycine rich repressor domains, Sp1-like repressors, E(spl) repressors, IκB repressors, Sin3 repressors, methyl-CpG binding protein 2 (MeCP2) repressors, or combinations thereof. Transcriptional activation or transcriptional repressor domains can be genetically fused to the Cas9 protein or bound via noncovalent protein-protein, protein-RNA, or protein-DNA interactions.

[0030] In further embodiments, the one or more heterologous domains can be an RNA aptamer binding domain (Konermann *et al.*, Nature, 2015, 517(7536):583-588; Zalatan *et al.*, Cell, 2015, 160(1-2):339-50). Examples of suitable RNA aptamer protein domains include MS2 coat protein

(MCP), PP7 bacteriophage coat protein (PCP), Mu bacteriophage Com protein, lambda bacteriophage N22 protein, stem-loop binding protein (SLBP), Fragile X mental retardation syndrome-related protein 1 (FXR1), proteins derived from bacteriophage such as AP205, BZ13, f1, f2, fd, fr, ID2, JP34/GA, JP501, JP34, JP500, KU1, M11, M12, MX1, NL95, PP7, ϕCb5, ϕCb8r, ϕCb12r, ϕCb23r, Qβ, R17, SP-β, TW18, TW19, and VK, fragments thereof, or derivatives thereof.

[0031] In yet other embodiments, the one or more heterologous domains can be a non-Cas9 nuclease domain. Suitable nuclease domains can be obtained from any endonuclease or exonuclease. Non-limiting examples of endonucleases from which a nuclease domain can be derived include, but are not limited to, restriction endonucleases and homing endonucleases. In some embodiments, the nuclease domain can be derived from a type II-S restriction endonuclease. Type II-S endonucleases cleave DNA at sites that are typically several base pairs away from the recognition/binding site and, as such, have separable binding and cleavage domains. These enzymes generally are monomers that transiently associate to form dimers to cleave each strand of DNA at staggered locations. Non-limiting examples of suitable type II-S endonucleases include BfiI, BpmI, BsaI, BsgI, BsmBI, BsmI, BspMI, FokI, MboII, and SapI. In some embodiments, the nuclease domain can be a FokI nuclease domain or a derivative thereof. The type II-S nuclease domain can be modified to facilitate dimerization of two different nuclease domains. For example, the cleavage domain of FokI can be modified by mutating certain amino acid residues. By way of non-limiting example, amino acid residues at positions 446, 447, 479, 483, 484, 486, 487, 490, 491, 496, 498, 499, 500, 531, 534, 537, and 538 of FokI nuclease domains are targets for modification. In specific embodiments, the FokI nuclease domain can comprise a first FokI half- domain comprising Q486E, I499L, and/or N496D mutations, and a second FokI half-domain comprising E490K, I538K, and/or H537R mutations.

[0032] The one or more heterologous domains can be linked directly to the Cas9 protein via one or more chemical bonds (*e.g.*, covalent bonds), or the one or more heterologous domains can be linked indirectly to the Cas9 protein via one or more linkers.

[0033] A linker is a chemical group that connects one or more other chemical groups via at least one covalent bond. Suitable linkers include amino acids, peptides, nucleotides, nucleic acids, organic linker molecules (*e.g.*, maleimide derivatives, N-ethoxybenzylimidazole, biphenyl-3,4′,5-tricarboxylic acid, p-aminobenzyloxycarbonyl, and the like), disulfide linkers, and polymer linkers (*e.g.*, PEG). The linker can include one or more spacing groups including, but not limited to alkylene, alkenylene, alkynylene, alkyl, alkenyl, alkynyl, alkoxy, aryl, heteroaryl, aralkyl, aralkenyl, aralkynyl and the like. The linker can be neutral, or carry a positive or negative charge. Additionally, the linker can be cleavable such that the linker's covalent bond that connects the linker to another chemical group can be broken or cleaved under certain conditions, including pH, temperature, salt concentration, light, a catalyst, or an enzyme. In some embodiments, the linker can be a peptide linker. The peptide linker can be a flexible amino acid linker (*e.g.*, comprising small, non-polar or polar amino acids).

[0034] In one particular embodiment, the linker is selected from the following:

Linker Protein Sequence
AEAAAKEAAAKEAAAKEAAAKALEAEAAAKEAAAKEAAAKEAAAKA (SEQ ID NO:35)
SGGSSGGSSGSETPGTSESATPESSGGSSGGS (SEQ ID NO:36)

[0035] Other non-limiting examples of flexible linkers include LEGGGS (SEQ ID NO:37), TGSG (SEQ ID NO:38), GGSGGGSG (SEQ ID NO:39), (GGGGS)$_{1-4}$ (SEQ ID NO:40), and (Gly)$_{6-8}$ (SEQ ID NO:41). Alternatively, the peptide linker can be a rigid amino acid linker. Such linkers include (EAAAK)$_{1-4}$ (SEQ ID NO:42), A(EAAAK)$_{2-5}$A (SEQ ID NO:43), PAPAP (SEQ ID NO:44), and (AP)$_{6-8}$ (SEQ ID NO:45). Additional examples of suitable linkers are well known in the art and programs to design linkers are readily available (Crasto *et al.*, Protein Eng., 2000, 13(5):309-312).

[0036] In some embodiments, the engineered Cas9 proteins can be produced recombinantly in cell-free systems, bacterial cells, or eukaryotic cells and purified using standard purification means. In other embodiments, the engineered Cas9 proteins are produced *in vivo* in eukaryotic cells of interest from nucleic acids encoding the engineered Cas9 proteins (see section (II) below).

[0037] In embodiments in which the engineered Cas9 protein comprises nuclease or nickase activity, the engineered Cas9 protein can further comprise at least cell-penetrating domain, as well as at least one chromatin disrupting domain. In embodiments in which the engineered Cas9 protein is linked to an epigenetic modification domain, the engineered Cas9 protein can further comprise at least one cell-penetrating domain, as well as at least one chromatin disrupting domain. Furthermore, in embodiments in which the engineered Cas9 protein is linked to a transcriptional regulation domain, the engineered Cas9 protein can further comprise at least one cell-penetrating domain, as well as at least one chromatin disrupting domain and/or at least one RNA aptamer binding domain.

[0038] The various fusion protein components can be combined, from N-terminus to C-terminus, in any order. For example, wherein A represents the marker protein, B represents a nuclear localization signal, and C represents the Cas9 protein, the fusion protein can be arranged, from N-terminus to C-terminus, in the following manner: A-B-C; A-C-B; B-A-C; B-C-A; C-A-B; or C-B-A, wherein a linker ("-L-") may be disposed between any two items (e.g., A-L-B-C; A-B-L-C; A-L-B-L-C; and so on).

(b) Engineered guide RNAs

[0039] The engineered guide RNAs is designed to complex with a specific engineered Cas9 protein. A guide RNA comprises (i) a CRISPR RNA (crRNA) that contains a guide sequence at the 5' end that hybridizes with a target sequence and (ii) a transacting crRNA (tracrRNA) sequence that recruits the Cas9 protein. The crRNA guide sequence of each guide RNA is different (*i.e.*, is sequence specific). The tracrRNA sequence is generally the same in guide RNAs designed to complex with a Cas9 protein from a particular bacterial species.

[0040] The crRNA guide sequence is designed to hybridize with a target sequence (*i.e.*, protospacer) in a double-stranded sequence. In general, the complementarity between the crRNA and the target sequence is at least 80%, at least 85%, at least 90%, at least 95%, or at least 99%. In specific embodiments, the complementarity is complete (*i.e.*, 100%). In various embodiments, the length of the crRNA guide sequence can range from about 15 nucleotides to about 25 nucleotides. For example, the crRNA guide sequence can be about 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, or 25 nucleotides in length. In specific embodiments, the crRNA is about 19, 20, or 21 nucleotides in length. In one embodiment, the crRNA guide sequence has a length of 20 nucleotides.

[0041] The guide RNA comprises repeat sequence that forms at least one stem loop structure, which interacts with the Cas9 protein, and 3' sequence that remains single-stranded. The length of each loop and stem can vary. For example, the loop can range from about 3 to about 10 nucleotides in length, and the stem can range from about 6 to about 20 base pairs in length. The stem can comprise one or more bulges of 1 to about 10 nucleotides. The length of the single-stranded 3' region can vary. The tracrRNA sequence in the engineered guide RNA generally is based upon the coding sequence of wild type tracrRNA in the bacterial species of interest. The wild-type sequence can be modified to facilitate secondary structure formation, increased secondary structure stability, facilitate expression in eukaryotic cells, and so forth. For example, one or more nucleotide changes can be introduced into the guide RNA coding sequence (see Example 3, below). The tracrRNA sequence can range in length from about 50 nucleotides to about 300 nucleotides. In various embodiments, the tracrRNA can range in length from about 50 to about 90 nucleotides, from about 90 to about 110 nucleotides, from about 110 to about 130 nucleotides, from about 130 to about 150 nucleotides, from about 150 to about 170 nucleotides, from about 170 to about 200 nucleotides, from about 200 to about 250 nucleotides, or from about 250 to about 300 nucleotides.

[0042] In general, the engineered guide RNA is a single molecule (*i.e.*, a single guide RNA or sgRNA), wherein the crRNA sequence is linked to the tracrRNA sequence. In some embodiments, however, the engineered guide RNA can be two separate molecules. A first molecule comprising the crRNA that contains 3' sequence (comprising from about 6 to about 20 nucleotides) that is capable of base pairing with the 5' end of a second molecule, wherein the second molecule comprises the tracrRNA that contains 5’ sequence (comprising from about 6 to about 20 nucleotides) that is capable of base pairing with the 3’ end of the first molecule.

[0043] In some embodiments, the tracrRNA sequence of the engineered guide RNA can be modified to comprise one or more aptamer sequences (Konermann *et al.*, Nature, 2015, 517(7536):583-588; Zalatan *et al.*, Cell, 2015, 160(1-2):339-50). Suitable aptamer sequences include those that bind adaptor proteins chosen from MCP, PCP, Com, SLBP, FXR1, AP205, BZ13, f1, f2, fd, fr, ID2, JP34/GA, JP501, JP34, JP500, KU1, M11, M12, MX1, NL95, PP7, ϕCb5, ϕCb8r, ϕCb12r, ϕCb23r, Qβ, R17, SP-β, TW18, TW19, VK, fragments thereof, or derivatives thereof. Those of skill in the art appreciate that the length of the aptamer sequence can vary.

[0044] In other embodiments, the guide RNA can further comprise at least one detectable label. The detectable label can be a fluorophore (*e.g.*, FAM, TMR, Cy3, Cy5, Texas Red, Oregon Green, Alexa Fluors, Halo tags, or suitable fluorescent dye), a detection tag (*e.g.*, biotin, digoxigenin, and the like), quantum dots, or gold particles.

[0045] The guide RNA can comprise standard ribonucleotides and/or modified ribonucleotides. In some embodiment, the guide RNA can comprise standard or modified deoxyribonucleotides. In embodiments in which the guide RNA is enzymatically synthesized (*i.e.*, *in vivo* or *in vitro*), the guide RNA generally comprises standard ribonucleotides. In embodiments in which the guide RNA is chemically synthesized, the guide RNA can comprise standard or modified ribonucleotides and/or deoxyribonucleotides. Modified ribonucleotides and/or deoxyribonucleotides include base modifications (*e.g.*, pseudouridine, 2-thiouridine, N6-methyladenosine, and the like) and/or sugar modifications (*e.g.*, 2’-O-methy, 2’-fluoro, 2’-amino, locked nucleic acid (LNA), and so forth). The backbone of the guide RNA can also be modified to comprise phosphorothioate linkages, boranophosphate linkages, or peptide nucleic acids.

(c) PAM Sequence

[0046] In some embodiments, the target sequence may be adjacent to a protospacer adjacent motif (PAM), a short sequence recognized by a CRISPR/Cas9 complex. In some embodiments, the PAM may be adjacent to or within 1, 2, 3, or 4, nucleotides of the 3' end of the target sequence. The length and the sequence of the PAM may depend on the Cas9 protein used. For example, the PAM may be selected from a consensus or a particular PAM sequence for a specific Cas9 protein or Cas9 ortholog, including those disclosed in FIG. 1 of Ran et al., Nature, 520: 186-191 (2015), which is incorporated herein by reference. In some embodiments, the PAM may comprise 2, 3, 4, 5, 6, 7, 8, 9, or 10 nucleotides in length. Non-limiting exemplary PAM sequences include NGG, NGGNG, NG, NAAAAN, NNAAAAW, NNNNACA, GNNNCNNA, and NNNNGATT (wherein N is defined as any nucleotide, and W is defined as either A or T). In some embodiments, the PAM sequence may be NGG. In some embodiments, the PAM sequence may be NGGNG. In some embodiments, the PAM sequence may be NNAAAAW.

[0047] It will be understood that different CRISPR proteins recognize different PAM sequences. For example, PAM sequences for Cas9 proteins include 5'-NGG, 5'-NGGNG, 5'-NNAGAAW, 5'-NNNNGATT, 5-NNNNRYAC, 5'-NNNNCAAA, 5'-NGAAA, 5'-NNAAT, 5'-NNNRTA, 5'-NNGG, 5'-NNNRTA, 5'-MMACCA, 5'-NNNNGRY, 5'-NRGNK, 5'-GGGRG, 5'-NNAMMMC, and 5'-NNG, and PAM sequences for Cas12a proteins include 5'-TTN and 5'-TTTV, wherein N is defined as any nucleotide, R is defined as either G or A, W is defined as either A or T, Y is defined an either C or T, and V is defined as A, C, or G. In general, Cas9 PAMs are located 3' of the target sequence, and Cas12a PAMs are located 5' of the target sequence. Various PAM sequences and the CRISPR proteins that recognize them are known in the art, *e.g.*, U.S. Patent Application Publication 2019/0249200; Leenay, Ryan T., et al. "Identifying and visualizing functional PAM diversity across CRISPR-Cas systems." Molecular cell 62.1 (2016): 137-147; and Kleinstiver, Benjamin P., et al. "Engineered CRISPR-Cas9 nucleases with altered PAM specificities." Nature 523.7561 (2015): 481, each of which are incorporated by reference herein in their entirety.

[0048] Additionally or alternatively, the PAM for each of the engineered Cas9 systems disclosed herein is presented below.

PAM Sequences	
Engineered Cas9system	PAM (5’-3’)*

Bacillus smithii Cas9 (BsmCas9)	NNNNCAAA
Lactobacillus rhamnosus Cas9 (LrhCas9)	NGAAA
Parasutterella excrementihominis Cas9 (PexCas9)	NGG
Mycoplasma canis Cas9 (McaCas9)	NNGG
Mycoplasma gallisepticum Cas9 (MgaCas9)	NNAAT
Akkermansia glycaniphila Cas9 (AglCas9)	NNNRTA
Akkermansia muciniphila Cas9 (AmuCas9)	MMACCA
Oenococcus kitaharae Cas9 (OkiCas9)	NNG
Bifidobacterium bombi Cas9 (BboCas9)	NNNNGRY
Acidothermus cellulolyticus Cas9 (AceCas9)	NGG
Alicyclobacillus hesperidum Cas9 (AheCas9)	NGG
Wolinella succinogenes Cas9 (WsuCas9)	NGG
Nitratifractor salsuginis Cas9 (NsaCas9)	NRGNK
Ralstonia syzygii Cas9 (RsyCas9)	GGGRG
Corynebacterium diphtheria Cas9 (CdiCas9)	NNAMMMC

* K is G or T; M is A or C; R is A or G; Y is C or T; and N is A, C, G, or T.

See, e.g., U.S. Patent Application Publication No. 2019/0249200 (hereby incorporated by reference herein in its entirety.

(II) Nucleic Acids

[0049] A further aspect of the present disclosure provides nucleic acids encoding the engineered Cas9 systems described above in section (I). The systems can be encoded by single nucleic acids or multiple nucleic acids. The nucleic acids can be DNA or RNA, linear or circular, single-stranded or double-stranded. The RNA or DNA can be codon optimized for efficient translation into protein in the eukaryotic cell of interest. Codon optimization programs are available as freeware or from commercial sources.

[0050] In some embodiments, nucleic acid encodes a protein having at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 95%, or at least about 99% sequence identity to the amino acid sequence of SEQ ID NO:48, 49, or 50. In certain embodiments, the nucleic acid encoding the engineered Cas9 protein can have at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 95%, or at least about 99% sequence identity to the DNA sequence of SEQ ID NO:48, 49, or 50. In certain embodiments, the DNA encoding the engineered Cas9 protein has the DNA sequence of SEQ ID NO:48, 49, or 50. In additional embodiments, the nucleic acid encodes a protein having at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 95%, or at least about 99% sequence identity to the amino acid sequence of SEQ ID NO:48, 49, or 50.

[0051] In some embodiments, the nucleic acid encoding the engineered Cas9 protein can be RNA. The RNA can be enzymatically synthesized *in vitro*. For this, DNA encoding the engineered Cas9 protein can be operably linked to a promoter sequence that is recognized by a phage RNA polymerase for *in vitro* RNA synthesis. For example, the promoter sequence can be a T7, T3, or SP6 promoter sequence or a variation of a T7, T3, or SP6 promoter sequence. The DNA encoding the engineered protein can be part of a vector, as detailed below. In such embodiments, the *in vitro*-transcribed RNA can be purified, capped, and/or polyadenylated. In other embodiments, the RNA encoding the engineered Cas9 protein can be part of a self-replicating RNA (Yoshioka *et al.*, Cell Stem Cell, 2013, 13:246-254). The self-replicating RNA can be derived from a noninfectious, self-replicating Venezuelan equine encephalitis (VEE) virus RNA replicon, which is a positive-sense, single-stranded RNA that is capable of self-replicating for a limited number of cell divisions, and which can be modified to code proteins of interest (Yoshioka *et al.*, Cell Stem Cell, 2013, 13:246-254).

[0052] In other embodiments, the nucleic acid encoding the engineered Cas9 protein can be DNA. The DNA coding sequence can be operably linked to at least one promoter control sequence for expression in the cell of interest. In certain embodiments, the DNA coding sequence can be operably linked to a promoter sequence for expression of the engineered Cas9 protein in bacterial (*e.g.*, *E. coli*) cells or eukaryotic (*e.g.*, yeast, insect, or mammalian) cells. Suitable bacterial promoters include, without limit, T7 promoters, *lac* operon promoters, *trp* promoters, *tac* promoters (which are hybrids of *trp* and *lac* promoters), variations of any of the foregoing, and combinations of any of the foregoing. Non-limiting examples of suitable eukaryotic promoters include constitutive, regulated, or cell- or tissue-specific promoters. Suitable eukaryotic constitutive promoter control sequences include, but are not limited to, cytomegalovirus immediate early promoter (CMV), simian virus (SV40) promoter, adenovirus major late promoter, Rous sarcoma virus (RSV) promoter, mouse mammary tumor virus (MMTV) promoter, phosphoglycerate kinase (PGK) promoter, elongation factor (ED1)-alpha promoter, ubiquitin promoters, actin promoters, tubulin promoters, immunoglobulin promoters, fragments thereof, or combinations of any of the foregoing. Examples of suitable eukaryotic regulated promoter control sequences include without limit those regulated by heat shock, metals, steroids, antibiotics, or alcohol. Non-limiting examples of tissue-specific promoters include B29 promoter, CD14 promoter, CD43 promoter, CD45 promoter, CD68 promoter, desmin promoter, elastase-1 promoter, endoglin promoter, fibronectin promoter, Flt-1 promoter, GFAP promoter, GPIIb promoter, ICAM-2 promoter, INF-β promoter, Mb promoter, Nphsl promoter, OG-2 promoter, SP-B promoter, SYN1 promoter, and WASP promoter. The promoter sequence can be wild type or it can be modified for more efficient or efficacious expression. In some embodiments, the DNA coding sequence also can be linked to a polyadenylation signal (*e.g.*, SV40 polyA signal, bovine growth hormone (BGH) polyA signal, *etc.*) and/or at least one transcriptional termination sequence. In some situations, the engineered Cas9 protein can be purified from the bacterial or eukaryotic cells.

[0053] In still other embodiments, the engineered guide RNA can be encoded by DNA. In some instances, the DNA encoding the engineered guide RNA can be operably linked to a promoter sequence that is recognized by a phage RNA polymerase for *in vitro* RNA synthesis. For example, the promoter sequence can be a T7, T3, or SP6 promoter sequence or a variation of a T7, T3, or SP6 promoter sequence. In other instances, the DNA encoding the engineered guide RNA can be operably linked to a promoter sequence that is recognized by RNA polymerase III (Pol III) for expression in eukaryotic cells of interest. Examples of suitable Pol III promoters include, but are not limited to, mammalian U6, U3, H1, and 7SL RNA promoters.

[0054] In various embodiments, the nucleic acid encoding the engineered Cas9 protein can be present in a vector. In some embodiments, the vector can further comprise nucleic acid encoding the engineered guide RNA. Suitable vectors include plasmid vectors, viral vectors, and self-replicating RNA (Yoshioka *et al.*, Cell Stem Cell, 2013, 13:246-254). In some embodiments, the nucleic acid encoding the complex or fusion protein can be present in a plasmid vector. Non-limiting examples of suitable plasmid vectors include pUC, pBR322, pET, pBluescript, and variants thereof. In other embodiments, the nucleic acid encoding the complex or fusion protein can be part of a viral vector (*e.g.*, lentiviral vectors, adeno-associated viral vectors, adenoviral vectors, and so forth). The plasmid or viral vector can comprise additional expression control sequences (*e.g.*, enhancer sequences, Kozak sequences, polyadenylation sequences, transcriptional termination sequences, *etc.*), selectable marker sequences (*e.g.*, antibiotic resistance genes), origins of replication, and the like. Additional information about vectors and use thereof can be found in “Current Protocols in Molecular Biology” Ausubel *et al.*, John Wiley & Sons, New York, 2003 or “Molecular Cloning: A Laboratory Manual” Sambrook & Russell, Cold Spring Harbor Press, Cold Spring Harbor, NY, 3rd edition, 2001.

(III) Eukaryotic Cells

[0055] Another aspect of the present disclosure comprises eukaryotic cells comprising at least one engineered Cas9 system as detailed above in section (I) and/or at least one nucleic acid encoding and engineered Cas9 protein and/or engineered guide RNA as detailed above in section (II).

[0056] The eukaryotic cell can be a human cell, a non-human mammalian cell, a non-mammalian vertebrate cell, an invertebrate cell, a plant cell, or a single cell eukaryotic organism. Examples of suitable eukaryotic cells are detailed below in section (IV)(c). The eukaryotic cell can be *in vitro*, *ex vivo*, or *in vivo*.

[0057] By way of example, in some embodiments, the eukaryotic cell, or a population of eukaryotic cells, is a T-cell, a CD8$^+$ T-cell, a CD8$^+$ naive T cell, a central memory T cell, an effector memory T-cell, a CD4$^+$ T-cell, a stem cell memory T-cell, a helper T-cell, a regulatory T-cell, a cytotoxic T-cell, a natural killer T-cell, a hematopoietic stem cell, a long term hematopoietic stem cell, a short term hematopoietic stem cell, a multipotent progenitor cell, a lineage restricted progenitor cell, a lymphoid progenitor cell, a pancreatic progenitor cell, an endocrine progenitor cell, an exocrine progenitor cell, a myeloid progenitor cell, a common myeloid progenitor cell, an erythroid progenitor cell, a megakaryocyte erythroid progenitor cell, a monocytic precursor cell, an endocrine precursor cell, an exocrine cell, a fibroblast, a hepatoblast, a myoblast, a macrophage, an islet beta-cell, a cardiomyocyte, a blood cell, a ductal cell, an acinar cell, an alpha cell, a beta cell, a delta cell, a PP cell, a cholangiocyte, a retinal cell, a photoreceptor cell, a rod cell, a cone cell, a retinal pigmented epithelium cell, a trabecular meshwork cell, a cochlear hair cell, an outer hair cell, an inner hair cell, a pulmonary epithelial cell, a bronchial epithelial cell, an alveolar epithelial cell, a pulmonary epithelial progenitor cell, a striated muscle cell, a cardiac muscle cell, a muscle satellite cell, a myocyte, a neuron, a neuronal stem cell, a mesenchymal stem cell, an induced pluripotent stem (iPS) cell, an embryonic stem cell, a monocyte, a megakaryocyte, a neutrophil, an eosinophil, a basophil, a mast cell, a reticulocyte, a B cell, e.g. a progenitor B cell, a Pre B cell, a Pro B cell, a memory B cell, a plasma B cell, a gastrointestinal epithelial cell, a biliary epithelial cell, a pancreatic ductal epithelial cell, an intestinal stem cell, a hepatocyte, a liver stellate cell, a Kupffer cell, an osteoblast, an osteoclast, an adipocyte (e.g., a brown adipocyte, or a white adipocyte), a preadipocyte, a pancreatic precursor cell, a pancreatic islet cell, a pancreatic beta cell, a pancreatic alpha cell, a pancreatic delta cell, a pancreatic exocrine cell, a Schwann cell, or an oligodendrocyte, or a population of such cells.

(IV) Methods for Modifying Chromosomal Sequences

[0058] A further aspect of the present disclosure encompasses methods for modifying a chromosomal sequence in eukaryotic cells. In general, the methods comprise introducing into the eukaryotic cell of interest at least one engineered Cas9 system as detailed above in section (I) and/or at least one nucleic acid encoding said engineered Cas9 system as detailed above in section (II).

[0059] In embodiments in which the engineered Cas9 protein comprises nuclease or nickase activity, the chromosomal sequence modification can comprise a substitution of at least one nucleotide, a deletion of at least one nucleotide, an insertion of at least one nucleotide. In some iterations, the method comprises introducing into the eukaryotic cell one engineered Cas9 system comprising nuclease activity or two engineered Cas9 systems comprising nickase activity and no donor polynucleotide, such that the engineered Cas9 system or systems introduce a double-stranded break in the target site in the chromosomal sequence and repair of the double-stranded break by cellular DNA repair processes introduces at least one nucleotide change (*i.e.*, indel), thereby inactivating the chromosomal sequence (*i.e.*, gene knock-out). In other iterations, the method comprises introducing into the eukaryotic cell one engineered Cas9 system comprising nuclease activity or two engineered Cas9 systems comprising nickase activity, as well as the donor polynucleotide, such that the engineered Cas9 system or systems introduce a double-stranded break in the target site in the chromosomal sequence and repair of the double-stranded break by cellular DNA repair processes leads to insertion or exchange of sequence in the donor polynucleotide into the target site in the chromosomal sequence (i.e., gene correction or gene knock-in).

[0060] In embodiments, in which the engineered Cas9 protein comprises epigenetic modification activity or transcriptional regulation activity, the chromosomal sequence modification can comprise a conversion of at least one nucleotide in or near the target site, a modification of at least one nucleotide in or near the target site, a modification of at least one histone protein in or near the target site, and/or a change in transcription in or near the target site in the chromosomal sequence.

(a) Introduction into the Cell

[0061] As mentioned above, the method comprises introducing into the eukaryotic cell at least one engineered Cas9 system and/or nucleic acid encoding said system (and optional donor polynucleotide). The at least one system and/or nucleic acid/donor polynucleotide can be introduced into the cell of interest by a variety of means.

[0062] In some embodiments, the cell can be transfected with the appropriate molecules (*i.e.*, protein, DNA, and/or RNA). Suitable transfection methods include nucleofection (or electroporation), calcium phosphate-mediated transfection, cationic polymer transfection (*e.g.*, DEAE-dextran or polyethylenimine), viral transduction, virosome transfection, virion transfection, liposome transfection, cationic liposome transfection, immunoliposome transfection, nonliposomal lipid transfection, dendrimer transfection, heat shock transfection, magnetofection, lipofection, gene gun delivery, impalefection, sonoporation, optical transfection, and proprietary agent-enhanced uptake of nucleic acids. Transfection methods are well known in the art (*see*, *e.g.*, "Current Protocols in Molecular Biology" Ausubel *et al.*, John Wiley & Sons, New York, 2003 or "Molecular Cloning: A Laboratory Manual" Sambrook & Russell, Cold Spring Harbor Press, Cold Spring Harbor, NY, 3rd edition, 2001). In other embodiments, the molecules can be introduced into the cell by microinjection. For example, the molecules can be injected into the cytoplasm or nuclei of the cells of interest. The amount of each molecule introduced into the cell can vary, but those skilled in the art are familiar with means for determining the appropriate amount.

[0063] The various molecules can be introduced into the cell simultaneously or sequentially. For example, the engineered Cas9 system (or its encoding nucleic acid) and the donor polynucleotide can be introduced at the same time. Alternatively, one can be introduced first and then the other can be introduced later into the cell.

[0064] In general, the cell is maintained under conditions appropriate for cell growth and/or maintenance. Suitable cell culture conditions are well known in the art and are described, for example, in Santiago *et al.*, Proc. Natl. Acad. Sci. USA, 2008, 105:5809-5814; Moehle *et al.* Proc. Natl. Acad. Sci. USA, 2007, 104:3055-3060; Urnov *et al.*, Nature, 2005, 435:646-651; and Lombardo *et al.*, Nat. Biotechnol., 2007, 25:1298-1306. Those of skill in the art appreciate that methods for culturing cells are known in the art and can and will vary depending on the cell type. Routine optimization may be used, in all cases, to determine the best techniques for a particular cell type.

(b) Optional Donor Polynucleotide

[0065] In embodiments in which the engineered Cas9 protein comprises nuclease or nickase activity, the method can further comprise introducing at least one donor polynucleotide into the cell. The donor polynucleotide can be single-stranded or double-stranded, linear or circular, and/or RNA or DNA. In some embodiments, the donor polynucleotide can be a vector, *e.g.*, a plasmid vector.

[0066] The donor polynucleotide comprises at least one donor sequence. In some aspects, the donor sequence of the donor polynucleotide can be a modified version of an endogenous or native chromosomal sequence. For example, the donor sequence can be essentially identical to a portion of the chromosomal sequence at or near the sequence targeted by the engineered Cas9 system, but which comprises at least one nucleotide change. Thus, upon integration or exchange with the native sequence, the sequence at the targeted chromosomal location comprises at least one nucleotide change. For example, the change can be an insertion of one or more nucleotides, a deletion of one or more nucleotides, a substitution of one or more nucleotides, or combinations thereof. As a consequence of the "gene correction" integration of the modified sequence, the cell can produce a modified gene product from the targeted chromosomal sequence.

[0067] In other aspects, the donor sequence of the donor polynucleotide can be an exogenous sequence. As used herein, an "exogenous" sequence refers to a sequence that is not native to the cell, or a sequence whose native location is in a different location in the genome of the cell. For example, the exogenous sequence can comprise protein coding sequence, which can be operably linked to an exogenous promoter control sequence such that, upon integration into the genome, the cell is able to express the protein coded by the integrated sequence. Alternatively, the exogenous sequence can be integrated into the chromosomal sequence such that its expression is regulated by an endogenous promoter control sequence. In other iterations, the exogenous sequence can be a transcriptional control sequence, another expression control sequence, an RNA coding sequence, and so forth. As noted above, integration of an exogenous sequence into a chromosomal sequence is termed a "knock in."

[0068] As can be appreciated by those skilled in the art, the length of the donor sequence can and will vary. For example, the donor sequence can vary in length from several nucleotides to hundreds of nucleotides to hundreds of thousands of nucleotides.

[0069] Typically, the donor sequence in the donor polynucleotide is flanked by an upstream sequence and a downstream sequence, which have substantial sequence identity to sequences located upstream and downstream, respectively, of the sequence targeted by the engineered Cas9 system. Because of these sequence similarities, the upstream and downstream sequences of the donor polynucleotide permit homologous recombination between the donor polynucleotide and the targeted chromosomal sequence such that the donor sequence can be integrated into (or exchanged with) the chromosomal sequence.

[0070] The upstream sequence, as used herein, refers to a nucleic acid sequence that shares substantial sequence identity with a chromosomal sequence upstream of the sequence targeted by the engineered Cas9 system. Similarly, the downstream sequence refers to a nucleic acid sequence that shares substantial sequence identity with a chromosomal sequence downstream of the sequence targeted by the engineered Cas9 system. As used herein, the phrase “substantial sequence identity” refers to sequences having at least about 75% sequence identity. Thus, the upstream and downstream sequences in the donor polynucleotide can have about 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% sequence identity with sequence upstream or downstream to the target sequence. In an exemplary embodiment, the upstream and downstream sequences in the donor polynucleotide can have about 95% or 100% sequence identity with chromosomal sequences upstream or downstream to the sequence targeted by the engineered Cas9 system.

[0071] In some embodiments, the upstream sequence shares substantial sequence identity with a chromosomal sequence located immediately upstream of the sequence targeted by the engineered Cas9 system. In other embodiments, the upstream sequence shares substantial sequence identity with a chromosomal sequence that is located within about one hundred (100) nucleotides upstream from the target sequence. Thus, for example, the upstream sequence can share substantial sequence identity with a chromosomal sequence that is located about 1 to about 20, about 21 to about 40, about 41 to about 60, about 61 to about 80, or about 81 to about 100 nucleotides upstream from the target sequence. In some embodiments, the downstream sequence shares substantial sequence identity with a chromosomal sequence located immediately downstream of the sequence targeted by the engineered Cas9 system. In other embodiments, the downstream sequence shares substantial sequence identity with a chromosomal sequence that is located within about one hundred (100) nucleotides downstream from the target sequence. Thus, for example, the downstream sequence can share substantial sequence identity with a chromosomal sequence that is located about 1 to about 20, about 21 to about 40, about 41 to about 60, about 61 to about 80, or about 81 to about 100 nucleotides downstream from the target sequence.

[0072] Each upstream or downstream sequence can range in length from about 20 nucleotides to about 5000 nucleotides. In some embodiments, upstream and downstream sequences can comprise about 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1100, 1200, 1300, 1400, 1500, 1600, 1700, 1800, 1900, 2000, 2100, 2200, 2300, 2400, 2500, 2600, 2800, 3000, 3200, 3400, 3600, 3800, 4000, 4200, 4400, 4600, 4800, or 5000 nucleotides. In specific embodiments, upstream and downstream sequences can range in length from about 50 to about 1500 nucleotides.

(c) Cell Types

[0073] A variety of eukaryotic cells are suitable for use in the methods disclosed herein. For example, the cell can be a human cell, a non-human mammalian cell, a non-mammalian vertebrate cell, an invertebrate cell, an insect cell, a plant cell, a yeast cell, or a single cell eukaryotic organism. In some embodiments, the cell can be a one cell embryo. For example, a non-human mammalian embryo including rat, hamster, rodent, rabbit, feline, canine, ovine, porcine, bovine, equine, and primate embryos. In still other embodiments, the cell can be a stem cell such as embryonic stem cells, ES-like stem cells, fetal stem cells, adult stem cells, and the like. In one embodiment, the stem cell is not a human embryonic stem cell. Furthermore, the stem cells may include those made by the techniques disclosed in WO2003/046141, which is incorporated herein in its entirety, or Chung *et al.* (Cell Stem Cell, 2008, 2:113-117). The cell can be *in vitro* (*i.e.*, in culture), *ex vivo* (*i.e.*, within tissue isolated from an organism), or *in vivo* (*i.e.*, within an organism). In exemplary embodiments, the cell is a mammalian cell or mammalian cell line. In particular embodiments, the cell is a human cell or human cell line.

[0074] By way of example, in some embodiments, the eukaryotic cell, or a population of eukaryotic cells, is a T-cell, a $CD8^+$ T-cell, a $CD8^+$ naive T cell, a central memory T cell, an effector memory T-cell, a $CD4^+$ T-cell, a stem cell memory T-cell, a helper T-cell, a regulatory T-cell, a cytotoxic T-cell, a natural killer T-cell, a hematopoietic stem cell, a long term hematopoietic stem cell, a short term hematopoietic stem cell, a multipotent progenitor cell, a lineage restricted progenitor cell, a lymphoid progenitor cell, a pancreatic progenitor cell, an endocrine progenitor cell, an exocrine progenitor cell, a myeloid progenitor cell, a common myeloid progenitor cell, an erythroid progenitor cell, a megakaryocyte erythroid progenitor cell, a monocytic precursor cell, an endocrine precursor cell, an exocrine cell, a fibroblast, a hepatoblast, a myoblast, a macrophage, an islet beta-cell, a cardiomyocyte, a blood cell, a ductal cell, an acinar cell, an alpha cell, a beta cell, a delta cell, a PP cell, a cholangiocyte, a retinal cell, a photoreceptor cell, a rod cell, a cone cell, a retinal pigmented epithelium cell, a trabecular meshwork cell, a cochlear hair cell, an outer hair cell, an inner hair cell, a pulmonary epithelial cell, a bronchial epithelial cell, an alveolar epithelial cell, a pulmonary epithelial progenitor cell, a striated muscle cell, a cardiac muscle cell, a muscle satellite cell, a myocyte, a neuron, a neuronal stem cell, a mesenchymal stem cell, an induced pluripotent stem (iPS) cell, an embryonic stem cell, a monocyte, a megakaryocyte, a neutrophil, an eosinophil, a basophil, a mast cell, a reticulocyte, a B cell, e.g. a progenitor B cell, a Pre B cell, a Pro B cell, a memory B cell, a plasma B cell, a gastrointestinal epithelial cell, a biliary epithelial cell, a pancreatic ductal epithelial cell, an intestinal stem cell, a hepatocyte, a liver stellate cell, a Kupffer cell, an osteoblast, an osteoclast, an adipocyte (e.g., a brown adipocyte, or a white adipocyte), a preadipocyte, a pancreatic precursor cell, a pancreatic islet cell, a pancreatic beta cell, a pancreatic alpha cell, a pancreatic delta cell, a pancreatic exocrine cell, a Schwann cell, or an oligodendrocyte, or a population of such cells.

[0075] Non-limiting examples of suitable mammalian cells or cell lines include human embryonic kidney cells (HEK293, HEK293T); human cervical carcinoma cells (HELA); human lung cells (W138); human liver cells (Hep G2); human U2-OS osteosarcoma cells, human A549 cells, human A-431 cells, and human K562 cells; Chinese hamster ovary (CHO) cells, baby hamster kidney (BHK) cells; mouse myeloma NS0 cells, mouse embryonic fibroblast 3T3 cells (NIH3T3), mouse B lymphoma A20 cells; mouse melanoma B16 cells; mouse myoblast C2C12 cells; mouse myeloma SP2/0 cells; mouse embryonic mesenchymal C3H-10T1/2 cells; mouse carcinoma CT26 cells, mouse prostate DuCuP cells; mouse breast EMT6 cells; mouse hepatoma Hepa1c1c7 cells; mouse myeloma J5582 cells; mouse epithelial MTD-1A cells; mouse myocardial MyEnd cells; mouse renal RenCa cells; mouse pancreatic RIN-5F cells; mouse melanoma X64 cells; mouse lymphoma YAC-1 cells; rat glioblastoma 9L cells; rat B lymphoma RBL cells; rat neuroblastoma B35 cells; rat hepatoma cells (HTC); buffalo rat liver BRL 3A cells; canine kidney cells (MDCK); canine mammary (CMT) cells; rat osteosarcoma D17 cells; rat monocyte/macrophage DH82 cells; monkey kidney SV-40 transformed fibroblast (COS7) cells; monkey kidney CVI-76 cells; African green monkey kidney (VERO-76) cells. An extensive list of mammalian cell lines may be found in the American Type Culture Collection catalog (ATCC, Manassas, VA).

(V) Applications

[0076] The compositions and methods disclosed herein can be used in a variety of therapeutic, diagnostic, industrial, and research applications. In some embodiments, the present disclosure can be used to modify any chromosomal sequence of interest in a cell, animal, or plant in order to model and/or study the function of genes, study genetic or epigenetic conditions of interest, or study biochemical pathways involved in various diseases or disorders. For example, transgenic organisms can be created that model diseases or disorders, wherein the expression of one or more nucleic acid sequences associated with a disease or disorder is altered. The disease model can be used to study the effects of mutations on the organism, study the development and/or progression of the disease, study the effect of a pharmaceutically active compound on the disease, and/or assess the efficacy of a potential gene therapy strategy.

[0077] In other embodiments, the compositions and methods can be used to perform efficient and cost effective functional genomic screens, which can be used to study the function of genes involved in a particular biological process and how any alteration in gene expression can affect the biological process, or to perform saturating or deep scanning mutagenesis of genomic loci in conjunction with a cellular phenotype. Saturating or deep scanning mutagenesis can be used to determine critical minimal features and discrete vulnerabilities of functional elements required for gene expression, drug resistance, and reversal of disease, for example.

[0078] In further embodiments, the compositions and methods disclosed herein can be used for diagnostic tests to establish the presence of a disease or disorder and/or for use in determining treatment options. Examples of suitable diagnostic tests include detection of specific mutations in cancer cells (*e.g.*, specific mutation in EGFR, HER2, and the like), detection of specific mutations associated with particular diseases (*e.g.*, trinucleotide repeats, mutations in β-globin associated with sickle cell disease, specific SNPs, *etc.*), detection of hepatitis, detection of viruses (*e.g.*, Zika), and so forth.

[0079] In additional embodiments, the compositions and methods disclosed herein can be used to correct genetic mutations associated with a particular disease or disorder such as, *e.g.*, correct globin gene mutations associated with sickle cell disease or thalassemia, correct mutations in the adenosine deaminase gene associated with severe combined immune deficiency (SCID), reduce the expression of HTT, the disease-causing gene of Huntington's disease, or correct mutations in the rhodopsin gene for the treatment of retinitis pigmentosa. Such modifications may be made in cells *ex vivo*.

[0080] In still other embodiments, the compositions and methods disclosed herein can be used to generate crop plants with improved traits or increased resistance to environmental stresses. The present disclosure can also be used to generate farm animal with improved traits or production animals. For example, pigs have many features that make them attractive as biomedical models, especially in regenerative medicine or xenotransplantation.

[0081] In still other embodiments, the compositions and methods disclosed herein can be used to determine chromosome identity and location within a living cell or chemically fixed cell (such as formalin fixation used in formalin-fixed paraffin embedded clinical samples). For example, a CRIPSR complex linked via a peptide sequence disclosed herein to a fluorescent protein maybe targeted in single or multiple copies to a genetic locus, and such complexes detected by microscopy to determine chromosomal locus copy number and/or location. Example genetic loci for tracking might include centromeric regions, telomeric regions, or other repetitive regions of the genome to which multiple copies of a single identical CRISPR complex may bind.

DEFINITIONS

[0082] Unless defined otherwise, all technical and scientific terms used herein have the meaning commonly understood by a person skilled in the art to which this invention belongs. The following references provide one of skill with a general definition of many of the terms used in this invention: Singleton *et al.*, Dictionary of Microbiology and Molecular Biology (2nd Ed. 1994); The Cambridge Dictionary of Science and Technology (Walker ed., 1988); The Glossary of Genetics, 5th Ed., R. Rieger *et al.* (eds.), Springer Verlag (1991); and Hale & Marham, The Harper Collins Dictionary of Biology (1991). As used herein, the following terms have the meanings ascribed to them unless specified otherwise.

[0083] When introducing elements of the present disclosure or the preferred embodiments(s) thereof, the articles “a”, “an”, “the” and “said" are intended to mean that there are one or more of the elements. The terms “comprising”, “including” and “having” are intended to be inclusive and mean that there may be additional elements other than the listed elements.

[0084] The term “about” when used in relation to a numerical value, x, for example means x ± 5%.

[0085] As used herein, the terms “complementary” or “complementarity” refer to the association of double-stranded nucleic acids by base pairing through specific hydrogen bonds. The base paring may be standard Watson-Crick base pairing (*e.g.*, 5'-A G T C-3' pairs with the complementary sequence 3'-T C A G-5'). The base pairing also may be Hoogsteen or reversed Hoogsteen hydrogen bonding. Complementarity is typically measured with respect to a duplex region and thus, excludes overhangs, for example. Complementarity between two strands of the duplex region may be partial and expressed as a percentage (*e.g.*, 70%), if only some (*e.g.*, 70%) of the bases are complementary. The bases that are not complementary are “mismatched.” Complementarity may also be complete (*i.e.*, 100%), if all the bases in the duplex region are complementary.

[0086] As used herein, the term “CRISPR/Cas system” or “Cas9 system” refers to a complex comprising a Cas9 protein (*i.e.*, nuclease, nickase, or catalytically dead protein) and a guide RNA.

[0087] The term “endogenous sequence,” as used herein, refers to a chromosomal sequence that is native to the cell.

[0088] As used herein, the term “exogenous” refers to a sequence that is not native to the cell, or a chromosomal sequence whose native location in the genome of the cell is in a different chromosomal location.

[0089] A “gene,” as used herein, refers to a DNA region (including exons and introns) encoding a gene product, as well as all DNA regions which regulate the production of the gene product, whether or not such regulatory sequences are adjacent to coding and/or transcribed sequences. Accordingly, a gene includes, but is not necessarily limited to, promoter sequences, terminators, translational regulatory sequences such as ribosome binding sites and internal ribosome entry sites, enhancers, silencers, insulators, boundary elements, replication origins, matrix attachment sites, and locus control regions.

[0090] The term “heterologous” refers to an entity that is not endogenous or native to the cell of interest. For example, a heterologous protein refers to a protein that is derived from or was originally derived from an exogenous source, such as an exogenously introduced nucleic acid sequence. In some instances, the heterologous protein is not normally produced by the cell of interest.

[0091] The term “nickase” refers to an enzyme that cleaves one strand of a double-stranded nucleic acid sequence (*i.e.*, nicks a double-stranded sequence). For example, a nuclease with double strand cleavage activity can be modified by mutation and/or deletion to function as a nickase and cleave only one strand of a double-stranded sequence.

[0092] The term “nuclease,” as used herein, refers to an enzyme that cleaves both strands of a double-stranded nucleic acid sequence.

[0093] The terms “nucleic acid” and “polynucleotide” refer to a deoxyribonucleotide or ribonucleotide polymer, in linear or circular conformation, and in either single- or double-stranded form. For the purposes of the present disclosure, these terms are not to be construed as limiting with respect to the length of a polymer. The terms can encompass known analogs of natural nucleotides, as well as nucleotides that are modified in the base, sugar and/or phosphate moieties (*e.g.*, phosphorothioate backbones). In general, an analog of a particular nucleotide has the same base-pairing specificity; *i.e.*, an analog of A will base-pair with T.

[0094] The term “nucleotide” refers to deoxyribonucleotides or ribonucleotides. The nucleotides may be standard nucleotides (*i.e.*, adenosine, guanosine, cytidine, thymidine, and uridine), nucleotide isomers, or nucleotide analogs. A nucleotide analog refers to a nucleotide having a modified purine or pyrimidine base or a modified ribose moiety. A nucleotide analog may be a naturally occurring nucleotide (*e.g.*, inosine, pseudouridine, etc.) or a non-naturally occurring nucleotide. Non-limiting examples of modifications on the sugar or base moieties of a nucleotide include the addition (or removal) of acetyl groups, amino groups, carboxyl groups, carboxymethyl groups, hydroxyl groups, methyl groups, phosphoryl groups, and thiol groups, as well as the substitution of the carbon and nitrogen atoms of the bases with other atoms (*e.g.*, 7-deaza purines). Nucleotide analogs also include dideoxy nucleotides, 2’-O-methyl nucleotides, locked nucleic acids (LNA), peptide nucleic acids (PNA), and morpholinos.

[0095] The terms “polypeptide” and “protein” are used interchangeably to refer to a polymer of amino acid residues.

[0096] The terms “target sequence,” “target chromosomal sequence,” and “target site” are used interchangeably to refer to the specific sequence in chromosomal DNA to which the engineered Cas9 system is targeted, and the site at which the engineered Cas9 system modifies the DNA or protein(s) associated with the DNA.

[0097] Techniques for determining nucleic acid and amino acid sequence identity are known in the art. Typically, such techniques include determining the nucleotide sequence of the mRNA for a gene and/or determining the amino acid sequence encoded thereby, and comparing these sequences to a second nucleotide or amino acid sequence. Genomic sequences can also be determined and compared in this fashion. In general, identity refers to an exact nucleotide-to-nucleotide or amino acid-to-amino acid correspondence of two polynucleotides or polypeptide sequences, respectively. Two or more sequences (polynucleotide or amino acid) can be compared by determining their percent identity. The percent identity of two sequences, whether nucleic acid or amino acid sequences, is the number of exact matches between two aligned sequences divided by the length of the shorter sequences and multiplied by 100. An approximate alignment for nucleic acid sequences is provided by the local homology algorithm of Smith and Waterman, Advances in Applied Mathematics 2:482-489 (1981). This algorithm can be applied to amino acid sequences by using the scoring matrix developed by Dayhoff, Atlas of Protein Sequences and Structure, M. O. Dayhoff ed., 5 suppl. 3:353-358, National Biomedical Research Foundation, Washington, D.C., USA, and normalized by Gribskov, Nucl. Acids Res. 14(6):6745-6763 (1986). An exemplary implementation of this algorithm to determine percent identity of a sequence is provided by the Genetics Computer Group (Madison, Wis.) in the “BestFit” utility application. Other suitable programs for calculating the percent identity or similarity between sequences are generally known in the art, for example, another alignment program is BLAST, used with default parameters. For example, BLASTN and BLASTP can be used using the following default parameters: genetic code=standard; filter=none; strand=both; cutoff=60; expect=10; Matrix=BLOSUM62; Descriptions=50 sequences; sort by=HIGH SCORE; Databases=non-redundant, GenBank+EMBL+DDBJ+PDB+GenBank CDS translations+Swiss protein+Spupdate+PIR. Details of these programs can be found on the GenBank website.

[0098] As various changes could be made in the above-described cells and methods without departing from the scope of the invention, it is intended that all matter contained in the above description and in the examples given below, shall be interpreted as illustrative and not in a limiting sense.

EXAMPLES

[0099] The following examples illustrate certain aspects of the disclosure.

Example 1: Human cell gene editing using GFP-SpCas9 and RFP-SpCas9 fusion proteins

[0100] Human K562 cells (0.35 x 10^6) were transfected with 60 pmol of SpCas9, GFP-SpCas9, or RFP-SpCas9 recombinant protein and 180 pmol of an in vitro transcribed single guide RNA (sgRNA) targeting the human EMX1 locus with the guide sequence 5'-GCUCCCAUCACAUCAACCGG-3'. Transfection was carried out using Nucleofection Solution V and an Amaxa instrument. Cells were maintained at 37°C and 5% CO_2 for three days before harvested for gene editing analysis. Genomic DNA was prepared using QuickExtract DNA extraction solution. Targeted EMX1 region was PCR amplified using primers consisting of target-specific sequences and next generation sequencing (NGS) adaptors. The forward primer is 5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNNNNNNAGTCTTCCCATCAGGCTCTCA-3' (SEQ ID NO:46) and the reverse primer is GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNNNNNAGAGTCCAGCTTGGGCC-3' (SEQ ID NO:47), where the target-specific sequences are underlined, and N represents A, T, G, or C. PCR amplicons were analyzed by NGS using the Illumina MiSeq to determine the editing efficiency of each Cas9 protein. The results displayed in **FIG. 1** show that the GFP-SpCas9 and RFP-SpCas9 fusing proteins each retain the editing activity parallel to the level by SpCas9 protein.

[0101] Table 1 presents the human codon optimized DNA and protein sequences of engineered Cas9/NLS proteins, wherein the NLS sequences are presented in bold text and the linker between the marker protein and Cas9 is presented in underlined text.

Table 1. Engineered Cas9 Systems
Amino acid sequence of GFP-SpCas9 fusion protein MVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPWPTLVTTLTYGVQCFSRYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSKLSKDPNEKRDHMVLLEFVTAAGITLGMDELYKVDAEAAAKEAAAKEAAAKEAAAKALEAEAAAKEAAAKEAAAKEAAAKA**PAAKRVKLD**GGGGSTGMDKKYSIGLDIGTNSVGWAVITDEYKVPSKKFKVLGNTDRHSIKKNLIGALLFDSGETAEATRLKRTARRRYTRRKNRICYLQEIFSNEMAKVDDSFFHRLEESFLVEEDKKHERHPIFGNIVDEVAYHEKYPTIYHLRKKLVDSTDKADLRLIYLALAHMIKFRGHFLIEGDLNPDNSDVDKLFIQLVQTYNQLFEENPINASGVDAKAILSARLSKSRRLENLIAQLPGEKKNGLFGNLIALSLGLTPNFKSNFDLAEDAKLQLSKDTYDDDLDNLLAQIGDQYADLFLAAKNLSDAILLSDILRVNTEITKAPLSASMIKRYDEHHQDLTLLKALVRQQLPEKYKEIFFDQSKNGYAGYIDGGASQEEFYKFIKPILEKMDGTEELLVKLNREDLLRKQRTFDNGSIPHQIHLGELHAILRRQEDFYPFLKDNREKIEKILTFRIPYYVGPLARGNSRFAWMTRKSEETITPWNFEEVVDKGASAQSFIERMTNFDKNLPNEKVLPKHSLLYEYFTVYNELTKVKYVTEGMRKPAFLSGEQKKAIVDLLFKTNRKVTVKQLKEDYFKKIECFDSVEISGVEDRFNASLGTYHDLLKIIKDKDFLDNEENEDILEDIVLTLTLFEDREMIEERLKTYAHLFDDKVMKQLKRRRYTGWGRLSRKLINGIRDKQSGKTILDFLKSDGFANRNFMQLIHDDSLTFKEDIQKAQVSGQGDSLHEHIANLAGSPAIKKGILQTVKVVDELVKVMGRHKPENIVIEMARENQTTQKGQKNSRERMKRIEEGIKELGSQILKEHPVENTQLQNEKLYLYYLQNGRDMYVDQELDINRLSDYDVDHIVPQSFLKDDSIDNKVLTRSDKNRGKSDNVPSEEVVKKMKNYWRQLLNAKLITQRKFDNLTKAERGGLSELDKAGFIKRQLVETRQITKHVAQILDSRMNTKYDENDKLIREVKVITLKSKLVSDFRKDFQFYKVREINNYHHAHDAYLNAVVGTALIKKYPKLESEFVYGDYKVYDVRKMIAKSEQEIGKATAKYFFYSNIMNFFKTEITLANGEIRKRPLIETNGETGEIVWDKGRDFATVRKVLSMPQVNIVKKTEVQTGGFSKESILPKRNSDKLIARKKDWDPKKYGGFDSPTVAYSVLVVAKVEKGKSKKLKSVKELLGITIMERSSFEKNPIDFLEAKGYKEVKKDLIIKLPKYSLFELENGRKRMLASAGELQKGNELALPSKYVNFLYLASHYEKLKGSPEDNEQKQLFVEQHKHYLDEIIEQISEFSKRVILADANLDKVLSAYNKHRDKPIREQAENIIHLFTLTNLGAPAAFKYFDTTIDRKRYTSTKEVLDATLIHQSITGLYETRIDLSQLGGDEF**PKKKRKV**GGGGS**PKKKRKV** (SEQ ID NO: 48) Underlined: Linker between GFP and SpCas9 Bold: Nuclear localization signals
Amino acid sequence of RFP-SpCas9 fusion protein MVSKGEAVIKEFMRFKVHMEGSMNGHEFEIEGEGEGRPYEGTQTAKLKVTKGGPLPFSWDILSPQFMYGSRAFTKHPADIPDYYKQSFPEGFKWERVMNFEDGGAVTVTQDTSLEDGTLIYKVKLRGTNFPPDGPVMQKKTMGWEASTERLYPEDGVLKGDIKMALRLKDGGRYLADFKTTYKAKKPVQMPGAYNVDRKLDITSHNEDYTVVEQYERSEGRHSTGGMDELYKVDSGGSSGGSSGSETPGTSESATPESSGGSSGGS**PAAKRVKLD**GGGGSTGMDKKYSIGLDI

GTNSVGWAVITDEYKVPSKKFKVLGNTDRHSIKKNLIGALLFDSGETAEAT
RLKRTARRRYTRRKNRICYLQEIFSNEMAKVDDSFFHRLEESFLVEEDKK
HERHPIFGNIVDEVAYHEKYPTIYHLRKKLVDSTDKADLRLIYLALAHMIKFR
GHFLIEGDLNPDNSDVDKLFIQLVQTYNQLFEENPINASGVDAKAILSARLS
KSRRLENLIAQLPGEKKNGLFGNLIALSLGLTPNFKSNFDLAEDAKLQLSK
DTYDDDLDNLLAQIGDQYADLFLAAKNLSDAILLSDILRVNTEITKAPLSAS
MIKRYDEHHQDLTLLKALVRQQLPEKYKEIFFDQSKNGYAGYIDGGASQE
EFYKFIKPILEKMDGTEELLVKLNREDLLRKQRTFDNGSIPHQIHLGELHAIL
RRQEDFYPFLKDNREKIEKILTFRIPYYVGPLARGNSRFAWMTRKSEETIT
PWNFEEVVDKGASAQSFIERMTNFDKNLPNEKVLPKHSLLYEYFTVYNEL
TKVKYVTEGMRKPAFLSGEQKKAIVDLLFKTNRKVTVKQLKEDYFKKIECF
DSVEISGVEDRFNASLGTYHDLLKIIKDKDFLDNEENEDILEDIVLTLTLFED
REMIEERLKTYAHLFDDKVMKQLKRRRYTGWGRLSRKLINGIRDKQSGKT
ILDFLKSDGFANRNFMQLIHDDSLTFKEDIQKAQVSGQGDSLHEHIANLAG
SPAIKKGILQTVKVVDELVKVMGRHKPENIVIEMARENQTTQKGQKNSRE
RMKRIEEGIKELGSQILKEHPVENTQLQNEKLYLYYLQNGRDMYVDQELDI
NRLSDYDVDHIVPQSFLKDDSIDNKVLTRSDKNRGKSDNVPSEEVVKKMK
NYWRQLLNAKLITQRKFDNLTKAERGGLSELDKAGFIKRQLVETRQITKHV
AQILDSRMNTKYDENDKLIREVKVITLKSKLVSDFRKDFQFYKVREINNYHH
AHDAYLNAVVGTALIKKYPKLESEFVYGDYKVYDVRKMIAKSEQEIGKATA
KYFFYSNIMNFFKTEITLANGEIRKRPLIETNGETGEIVWDKGRDFATVRKV
LSMPQVNIVKKTEVQTGGFSKESILPKRNSDKLIARKKDWDPKKYGGFDS
PTVAYSVLVVAKVEKGKSKKLKSVKELLGITIMERSSFEKNPIDFLEAKGYK
EVKKDLIIKLPKYSLFELENGRKRMLASAGELQKGNELALPSKYVNFLYLAS
HYEKLKGSPEDNEQKQLFVEQHKHYLDEIIEQISEFSKRVILADANLDKVLS
AYNKHRDKPIREQAENIIHLFTLTNLGAPAAFKYFDTTIDRKRYTSTKEVLD
ATLIHQSITGLYETRIDLSQLGGDEF**PKKKRKV**GGGGS**PKKKRKV** (SEQ
ID NO: 49)

Underlined: Linker between RFP and SpCas9
Bold: Nuclear localization signals

Amino acid sequence of GFP-eSpCas9 fusion protein

MVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTT
GKLPVPWPTLVTTLTYGVQCFSRYPDHMKQHDFFKSAMPEGYVQERTIF
FKDDGNYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNV
YIMADKQKNGIKVNFKIRHNIEDGSVQLADHYQQNTPIGDGPVLLPDNHYL
STQSKLSKDPNEKRDHMVLLEFVTAAGITLGMDELYKVDSGGSSGGSSG
SETPGTSESATPESSGGSSGGS**PAAKRVKLD**GGGGSTGMDKKYSIGLDI
GTNSVGWAVITDEYKVPSKKFKVLGNTDRHSIKKNLIGALLFDSGETAEAT
RLKRTARRRYTRRKNRICYLQEIFSNEMAKVDDSFFHRLEESFLVEEDKK
HERHPIFGNIVDEVAYHEKYPTIYHLRKKLVDSTDKADLRLIYLALAHMIKFR
GHFLIEGDLNPDNSDVDKLFIQLVQTYNQLFEENPINASGVDAKAILSARLS
KSRRLENLIAQLPGEKKNGLFGNLIALSLGLTPNFKSNFDLAEDAKLQLSK
DTYDDDLDNLLAQIGDQYADLFLAAKNLSDAILLSDILRVNTEITKAPLSAS
MIKRYDEHHQDLTLLKALVRQQLPEKYKEIFFDQSKNGYAGYIDGGASQE
EFYKFIKPILEKMDGTEELLVKLNREDLLRKQRTFDNGSIPHQIHLGELHAIL
RRQEDFYPFLKDNREKIEKILTFRIPYYVGPLARGNSRFAWMTRKSEETIT

PWNFEEVVDKGASAQSFIERMTNFDKNLPNEKVLPKHSLLYEYFTVYNEL
TKVKYVTEGMRKPAFLSGEQKKAIVDLLFKTNRKVTVKQLKEDYFKKIECF
DSVEISGVEDRFNASLGTYHDLLKIIKDKDFLDNEENEDILEDIVLTLTLFED
REMIEERLKTYAHLFDDKVMKQLKRRRYTGWGRLSRKLINGIRDKQSGKT
ILDFLKSDGFANRNFMQLIHDDSLTFKEDIQKAQVSGQGDSLHEHIANLAG
SPAIKKGILQTVKVVDELVKVMGRHKPENIVIEMARENQTTQKGQKNSRE
RMKRIEEGIKELGSQILKEHPVENTQLQNEKLYLYYLQNGRDMYVDQELDI
NRLSDYDVDHIVPQSFLADDSIDNKVLTRSDKNRGKSDNVPSEEVVKKMK
NYWRQLLNAKLITQRKFDNLTKAERGGLSELDKAGFIKRQLVETRQITKHV
AQILDSRMNTKYDENDKLIREVKVITLKSKLVSDFRKDFQFYKVREINNYHH
AHDAYLNAVVGTALIKKYPALESEFVYGDYKVYDVRKMIAKSEQEIGKATA
KYFFYSNIMNFFKTEITLANGEIRKAPLIETNGETGEIVWDKGRDFATVRKV
LSMPQVNIVKKTEVQTGGFSKESILPKRNSDKLIARKKDWDPKKYGGFDS
PTVAYSVLVVAKVEKGKSKKLKSVKELLGITIMERSSFEKNPIDFLEAKGYK
EVKKDLIIKLPKYSLFELENGRKRMLASAGELQKGNELALPSKYVNFLYLAS
HYEKLKGSPEDNEQKQLFVEQHKHYLDEIIEQISEFSKRVILADANLDKVLS
AYNKHRDKPIREQAENIIHLFTLTNLGAPAAFKYFDTTIDRKRYTSTKEVLD
ATLIHQSITGLYETRIDLSQLGGDEF**PKKKRKV**GGGGS**PKKKRKV** (SEQ ID NO:50)

Underlined: Linker between GFP and eSpCas9
Bold: Nuclear localization signals

[0102] Human codon optimized DNA sequences used to produce the three proteins are as follows:

Human codon optimized GFP-SpCas9 DNA sequence

ATGGTTAGCAAAGGTGAAGAACTGTTTACAGGTGTTGTTCCGATTCTG
GTTGAACTGGATGGTGATGTTAATGGCCACAAATTTTCAGTTAGCGGT
GAAGGCGAAGGTGATGCAACCTATGGTAAACTGACCCTGAAATTTATC
TGTACCACCGGCAAACTGCCGGTTCCGTGGCCGACACTGGTTACCAC
ACTGACCTATGGTGTTCAGTGTTTTAGCCGTTATCCGGATCACATGAAA
CAGCACGATTTTTTCAAAAGCGCAATGCCGGAAGGTTATGTTCAAGAA
CGTACCATCTTCTTCAAAGATGACGGCAACTATAAAACCCGTGCCGAA
GTTAAATTTGAAGGTGATACCCTGGTGAATCGCATTGAACTGAAAGGC
ATCGATTTTAAAGAGGATGGTAATATCCTGGGCCACAAACTGGAATATA
ATTATAATAGCCACAACGTGTACATCATGGCCGACAAACAGAAAAATG
GCATCAAAGTGAACTTCAAGATCCGCCATAATATTGAAGATGGTTCAGT
TCAGCTGGCCGATCATTATCAGCAGAATACCCCGATTGGTGATGGTCC
GGTTCTGCTGCCGGATAATCATTATCTGAGCACCCAGAGCAAACTGAG
CAAAGATCCGAATGAAAAACGTGATCACATGGTGCTGCTGGAATTTGT
TACCGCAGCAGGTATTACCTTAGGTATGGATGAACTGTATAAAGTCGA
CGCAGAAGCAGCAGCAAAAGAAGCCGCTGCCAAAGAAGCGGCAGCGA
AAGAGGCAGCCGCAAAAGCACTGGAAGCCGAGGCTGCGGCTAAAGAG
GCTGCTGCAAAAGAAGCAGCCGCTAAAGAAGCTGCGGCTAAGGCACC
GGCAGCAAAACGTGTTAAACTGGACGGTGGTGGTGGTAGCACCGGTA
TGGACAAGAAATACAGCATCGGTTTGGATATTGGCACGAATAGCGTGG

```
GTTGGGCCGTTATTACCGACGAGTACAAAGTGCCGTCCAAGAAATTCA
AAGTGCTGGGCAATACCGATCGCCATAGCATCAAGAAAAATCTGATTG
GCGCACTGCTGTTCGACAGCGGTGAGACTGCCGAAGCTACGCGTCTG
AAGCGTACGGCGCGTCGTCGCTACACCCGCCGTAAGAACCGTATTTG
CTATCTGCAAGAAATCTTCAGCAACGAAATGGCCAAAGTTGATGATAG
CTTTTTTCACCGCCTGGAAGAGAGCTTTCTGGTGGAAGAGGATAAGAA
ACACGAGCGCCATCCGATTTTTGGTAACATTGTCGATGAAGTGGCATA
CCATGAGAAGTACCCGACCATCTACCACCTTCGTAAGAAACTGGTGGA
CAGCACCGATAAAGCTGATCTGCGTCTGATTTACCTGGCGCTGGCCCA
CATGATTAAGTTTCGCGGTCATTTTCTGATCGAGGGCGATCTGAATCC
GGACAATTCTGATGTTGACAAGCTGTTTATTCAACTTGTACAGACCTAC
AACCAGTTGTTCGAAGAGAACCCGATCAATGCGAGCGGTGTTGATGCC
AAAGCAATTCTGAGCGCACGCCTGAGCAAATCTCGCCGTTTGGAGAAC
CTGATTGCACAGCTGCCGGGTGAGAAGAAAAACGGTCTGTTCGGCAAT
CTGATTGCACTGTCCCTGGGCTTGACCCCGAATTTTAAGAGCAACTTC
GACCTGGCCGAAGATGCGAAGCTCCAATTGAGCAAAGACACCTACGA
CGATGACCTGGACAATCTGCTGGCCCAGATTGGCGACCAGTACGCAG
ATCTGTTCTTGGCTGCGAAAAACCTGAGCGATGCAATTCTGCTGTCGG
ACATCCTGCGCGTGAATACGGAAATCACGAAAGCGCCTCTGAGCGCG
TCTATGATCAAGCGCTATGACGAGCACCACCAAGATCTGACCCTGCTG
AAAGCTCTGGTGAGACAACAATTGCCAGAGAAGTATAAAGAAATTTTCT
TTGACCAGAGCAAAAACGGCTATGCGGGTTACATTGACGGTGGCGCC
AGCCAAGAAGAGTTCTACAAATTCATTAAGCCTATCCTGGAGAAAATGG
ATGGCACCGAAGAACTGCTGGTAAAGCTGAATCGTGAAGATCTGCTGC
GCAAACAGCGCACTTTTGATAACGGTAGCATTCCGCACCAGATCCATC
TGGGTGAGTTGCACGCGATTTTGCGTCGCCAGGAAGATTTTTATCCGT
TCTTGAAAGACAACCGTGAGAAAATCGAGAAAATTCTGACGTTCCGTAT
CCCGTATTATGTCGGCCCGCTGGCGCGTGGTAATAGCCGCTTCGCGT
GGATGACCCGCAAATCAGAGGAAACGATTACCCCGTGGAATTTTGAGG
AAGTTGTTGATAAGGGTGCAAGCGCGCAGTCGTTCATTGAGCGTATGA
CCAACTTTGACAAGAATTTGCCGAATGAAAAAGTCTTGCCGAAGCACT
CTCTGCTGTACGAGTATTTTACCGTTTACAACGAATTGACCAAGGTTAA
ATACGTCACCGAAGGCATGCGCAAACCGGCCTTCCTGAGCGGCGAGC
AGAAAAAAGCAATCGTTGACCTCTTGTTTAAGACCAACCGCAAGGTTAC
GGTCAAACAACTGAAAGAGGACTATTTCAAGAAAATTGAATGTTTTGAC
TCCGTAGAGATCTCCGGTGTTGAGGACCGTTTCAACGCGAGCCTGGG
CACCTACCATGATCTGCTGAAAATTATTAAAGACAAAGATTTTCTGGAC
AACGAAGAGAACGAAGATATTCTGGAAGATATCGTTCTGACCCTGACG
CTGTTCGAAGATCGTGAGATGATTGAGGAACGTCTGAAAACCTACGCA
CACTTGTTCGATGACAAAGTTATGAAACAGCTGAAGCGTCGTCGTTAC
ACAGGTTGGGGCCGTCTGAGCCGTAAGCTTATCAATGGTATCCGTGAC
AAACAGAGCGGTAAGACGATTCTGGACTTTCTGAAGTCAGATGGCTTC
GCCAATCGCAACTTTATGCAACTGATTCATGACGACTCTCTGACGTTCA
AGGAAGATATCCAAAAGGCACAGGTGAGCGGTCAGGGTGATAGCCTG
CATGAGCATATCGCGAACCTGGCGGGTAGCCCGGCTATCAAAAAGGG
TATCTTACAGACTGTGAAAGTTGTGGATGAATTGGTTAAGGTTATGGGT
CGTCACAAACCGGAAAATATTGTGATCGAGATGGCACGTGAAAATCAG
ACGACGCAAAAGGGTCAAAAAAATTCTCGTGAGCGCATGAAACGTATT
GAAGAGGGTATCAAAGAATTGGGCAGCCAAATTCTGAAAGAACACCCG
GTCGAGAACACCCAGCTGCAAAACGAAAAACTGTATTTATACTATCTGC
```

AGAACGGTCGTGACATGTACGTGGATCAAGAACTGGACATCAATCGTTTGAGCGATTACGATGTTGATCATATTGTGCCTCAGAGCTTTCTGAAAGACGATTCGATCGACAACAAAGTGCTGACCCGTAGCGACAAGAATCGTGGTAAGAGCGATAACGTGCCGAGCGAAGAAGTCGTTAAGAAAATGAAAAACTACTGGCGTCAGCTGCTGAACGCCAAGCTGATTACCCAGCGTAAGTTCGATAACCTGACGAAAGCCGAGCGTGGAGGCCTGAGCGAGCTGGACAAGGCCGGCTTTATCAAGCGTCAACTGGTGGAAACCCGTCAGATCACTAAACATGTGGCACAGATCCTGGACTCCCGCATGAATACGAAATATGACGAGAATGACAAGTTGATCCGTGAAGTCAAAGTTATTACGCTGAAAAGCAAACTGGTGTCCGATTTCCGTAAAGACTTCCAGTTCTATAAAGTCCGTGAAATCAACAACTATCATCACGCCCACGATGCGTACTTGAACGCTGTTGTGGGCACCGCACTGATCAAGAAATACCCTAAGCTCGAAAGCGAGTTTGTCTATGGTGACTATAAAGTTTACGACGTGCGTAAGATGATCGCCAAGAGCGAGCAAGAAATTGGTAAGGCTACCGCAAAGTACTTTTTCTACAGCAACATCATGAACTTCTTCAAAACCGAGATTACCCTGGCGAACGGTGAGATCCGTAAACGGCCGCTGATTGAGACTAATGGCGAAACGGGCGAGATTGTGTGGGACAAGGGTCGCGATTTCGCTACGGTTCGTAAGGTCCTGAGCATGCCGCAAGTTAACATTGTCAAGAAAACTGAAGTGCAGACGGGTGGCTTTAGCAAAGAATCCATCCTGCCGAAGCGTAATAGCGATAAACTTATCGCGCGTAAAAAAGACTGGGACCCAAAGAAATATGGCGGCTTTGATAGCCCGACCGTCGCGTATAGCGTGTTAGTGGTCGCGAAAGTTGAAAAGGGCAAGAGCAAGAAACTGAAGTCCGTCAAAGAACTTCTGGGTATCACCATCATGGAACGTAGCTCCTTTGAGAAGAACCCGATTGACTTCTTAGAGGCGAAGGGTTATAAAGAAGTCAAAAAAGACCTGATTATCAAGCTGCCGAAGTACAGCCTGTTTGAGTTGGAGAATGGTCGTAAGCGCATGCTGGCGAGCGCGGGTGAGCTGCAAAAGGGCAACGAACTGGCGCTGCCGTCGAAATACGTCAATTTTCTGTACCTGGCCAGCCACTACGAAAAGCTGAAGGGTTCTCCGGAAGATAACGAACAAAAGCAACTGTTCGTTGAGCAACATAAACACTACTTGGACGAAATCATCGAGCAAATTAGCGAATTTAGCAAACGTGTCATCCTGGCGGACGCGAATCTGGACAAGGTCCTGTCTGCATACAATAAGCATCGCGACAAACCAATTCGTGAGCAAGCGGAGAATATCATCCACCTGTTTACGCTGACCAACCTAGGTGCGCCGGCGGCATTCAAGTATTTCGATACGACCATCGACCGCAAGCGCTATACCAGCACCAAAGAGGTCCTGGACGCGACCCTGATCCACCAGAGCATTACCGGCTTATACGAAACCCGTATTGATTTGAGCCAACTGGGTGGCGATGAATTCCCGAAAAAAAAGCGCAAAGTTGGTGGCGGTGGTAGCCCGAAAAAGAAACGTAAAGTG (SEQ ID NO:62)
Human codon optimized RFP-SpCas9 DNA sequence ATGGTTAGCAAAGGTGAAGCCGTGATTAAAGAATTTATGCGCTTTAAGGTTCACATGGAAGGTAGCATGAATGGCCATGAATTTGAAATTGAAGGTGAAGGCGAAGGTCGTCCGTATGAAGGCACCCAGACCGCAAAACTGAAAGTTACCAAAGGTGGTCCGCTGCCGTTTAGCTGGGATATTCTGAGTCCGCAGTTTATGTATGGTAGCCGTGCATTTACCAAACATCCGGCAGATATTCCGGATTATTACAAACAGAGCTTTCCGGAAGGTTTTAAATGGGAACGTGTGATGAATTTTGAAGATGGTGGTGCAGTTACCGTTACACAGGATACCAGCCTGGAAGATGGCACCCTGATCTATAAAGTTAAACTGCGTGGCACCAATTTTCCGCCTGATGGTCCGGTTATGCAGAAAAAAACAATGGGTTGGGAAGCAAGCACCGAACGTCTGTATCCTGAAGATGGCGTTCTGAAAGGTGATATCAAAATGGCACTGCGTCTGAAAGATGGCGGTCGTTATCTGGCA

```
GATTTCAAAACCACCTATAAAGCCAAAAAACCTGTTCAGATGCCTGGTG
CCTATAATGTTGATCGTAAACTGGATATTACCAGCCACAACGAAGATTA
TACCGTTGTGGAACAGTATGAACGTAGCGAAGGCCGTCATAGCACAG
GTGGTATGGATGAACTGTATAAAGTCGACAGCGGTGGTAGCAGCGGT
GGTTCAAGCGGTAGCGAAACACCGGGTACAAGCGAAAGCGCAACACC
GGAAAGCAGTGGTGGTAGTTCAGGTGGTAGTCCGGCAGCAAAACGTG
TGAAACTGGATGGCGGTGGCGGTAGCACCGGTATGGACAAGAAATAC
AGCATCGGTTTGGATATTGGCACGAATAGCGTGGGTTGGGCCGTTATT
ACCGACGAGTACAAAGTGCCGTCCAAGAAATTCAAAGTGCTGGGCAAT
ACCGATCGCCATAGCATCAAGAAAAATCTGATTGGCGCACTGCTGTTC
GACAGCGGTGAGACTGCCGAAGCTACGCGTCTGAAGCGTACGGCGC
GTCGTCGCTACACCCGCCGTAAGAACCGTATTTGCTATCTGCAAGAAA
TCTTCAGCAACGAAATGGCCAAAGTTGATGATAGCTTTTTTCACCGCCT
GGAAGAGAGCTTTCTGGTGGAAGAGGATAAGAAACACGAGCGCCATC
CGATTTTTGGTAACATTGTCGATGAAGTGGCATACCATGAGAAGTACC
CGACCATCTACCACCTTCGTAAGAAACTGGTGGACAGCACCGATAAAG
CTGATCTGCGTCTGATTTACCTGGCGCTGGCCCACATGATTAAGTTTC
GCGGTCATTTTCTGATCGAGGGCGATCTGAATCCGGACAATTCTGATG
TTGACAAGCTGTTTATTCAACTTGTACAGACCTACAACCAGTTGTTCGA
AGAGAACCCGATCAATGCGAGCGGTGTTGATGCCAAAGCAATTCTGAG
CGCACGCCTGAGCAAATCTCGCCGTTTGGAGAACCTGATTGCACAGCT
GCCGGGTGAGAAGAAAAACGGTCTGTTCGGCAATCTGATTGCACTGTC
CCTGGGCTTGACCCCGAATTTTAAGAGCAACTTCGACCTGGCCGAAGA
TGCGAAGCTCCAATTGAGCAAAGACACCTACGACGATGACCTGGACAA
TCTGCTGGCCCAGATTGGCGACCAGTACGCAGATCTGTTCTTGGCTGC
GAAAAACCTGAGCGATGCAATTCTGCTGTCGGACATCCTGCGCGTGAA
TACGGAAATCACGAAAGCGCCTCTGAGCGCGTCTATGATCAAGCGCTA
TGACGAGCACCACCAAGATCTGACCCTGCTGAAAGCTCTGGTGAGACA
ACAATTGCCAGAGAAGTATAAAGAAATTTTCTTTGACCAGAGCAAAAAC
GGCTATGCGGGTTACATTGACGGTGGCGCCAGCCAAGAAGAGTTCTA
CAAATTCATTAAGCCTATCCTGGAGAAAATGGATGGCACCGAAGAACT
GCTGGTAAAGCTGAATCGTGAAGATCTGCTGCGCAAACAGCGCACTTT
TGATAACGGTAGCATTCCGCACCAGATCCATCTGGGTGAGTTGCACGC
GATTTTGCGTCGCCAGGAAGATTTTTATCCGTTCTTGAAAGACAACCGT
GAGAAAATCGAGAAAATTCTGACGTTCCGTATCCCGTATTATGTCGGC
CCGCTGGCGCGTGGTAATAGCCGCTTCGCGTGGATGACCCGCAAATC
AGAGGAAACGATTACCCCGTGGAATTTTGAGGAAGTTGTTGATAAGGG
TGCAAGCGCGCAGTCGTTCATTGAGCGTATGACCAACTTTGACAAGAA
TTTGCCGAATGAAAAAGTCTTGCCGAAGCACTCTCTGCTGTACGAGTA
TTTTACCGTTTACAACGAATTGACCAAGGTTAAATACGTCACCGAAGGC
ATGCGCAAACCGGCCTTCCTGAGCGGCGAGCAGAAAAAAGCAATCGT
TGACCTCTTGTTTAAGACCAACCGCAAGGTTACGGTCAAACAACTGAA
AGAGGACTATTTCAAGAAAATTGAATGTTTTGACTCCGTAGAGATCTCC
GGTGTTGAGGACCGTTTCAACGCGAGCCTGGGCACCTACCATGATCT
GCTGAAAATTATTAAAGACAAAGATTTTCTGGACAACGAAGAGAACGAA
GATATTCTGGAAGATATCGTTCTGACCCTGACGCTGTTCGAAGATCGT
GAGATGATTGAGGAACGTCTGAAAACCTACGCACACTTGTTCGATGAC
AAAGTTATGAAACAGCTGAAGCGTCGTCGTTACACAGGTTGGGGCCGT
CTGAGCCGTAAGCTTATCAATGGTATCCGTGACAAACAGAGCGGTAAG
ACGATTCTGGACTTTCTGAAGTCAGATGGCTTCGCCAATCGCAACTTTA
```

TGCAACTGATTCATGACGACTCTCTGACGTTCAAGGAAGATATCCAAAA
GGCACAGGTGAGCGGTCAGGGTGATAGCCTGCATGAGCATATCGCGA
ACCTGGCGGGTAGCCCGGCTATCAAAAAGGGTATCTTACAGACTGTGA
AAGTTGTGGATGAATTGGTTAAGGTTATGGGTCGTCACAAACCGGAAA
ATATTGTGATCGAGATGGCACGTGAAAATCAGACGACGCAAAAGGGTC
AAAAAAATTCTCGTGAGCGCATGAAACGTATTGAAGAGGGTATCAAAG
AATTGGGCAGCCAAATTCTGAAAGAACACCCGGTCGAGAACACCCAGC
TGCAAAACGAAAAACTGTATTTATACTATCTGCAGAACGGTCGTGACAT
GTACGTGGATCAAGAACTGGACATCAATCGTTTGAGCGATTACGATGT
TGATCATATTGTGCCTCAGAGCTTTCTGAAAGACGATTCGATCGACAAC
AAAGTGCTGACCCGTAGCGACAAGAATCGTGGTAAGAGCGATAACGT
GCCGAGCGAAGAAGTCGTTAAGAAAATGAAAAACTACTGGCGTCAGCT
GCTGAACGCCAAGCTGATTACCCAGCGTAAGTTCGATAACCTGACGAA
AGCCGAGCGTGGAGGCCTGAGCGAGCTGGACAAGGCCGGCTTTATCA
AGCGTCAACTGGTGGAAACCCGTCAGATCACTAAACATGTGGCACAGA
TCCTGGACTCCCGCATGAATACGAAATATGACGAGAATGACAAGTTGA
TCCGTGAAGTCAAAGTTATTACGCTGAAAAGCAAACTGGTGTCCGATTT
CCGTAAAGACTTCCAGTTCTATAAAGTCCGTGAAATCAACAACTATCAT
CACGCCCACGATGCGTACTTGAACGCTGTTGTGGGCACCGCACTGAT
CAAGAAATACCCTAAGCTCGAAAGCGAGTTTGTCTATGGTGACTATAAA
GTTTACGACGTGCGTAAGATGATCGCCAAGAGCGAGCAAGAAATTGGT
AAGGCTACCGCAAAGTACTTTTTCTACAGCAACATCATGAACTTCTTCA
AAACCGAGATTACCCTGGCGAACGGTGAGATCCGTAAACGGCCGCTG
ATTGAGACTAATGGCGAAACGGGCGAGATTGTGTGGGACAAGGGTCG
CGATTTCGCTACGGTTCGTAAGGTCCTGAGCATGCCGCAAGTTAACAT
TGTCAAGAAAACTGAAGTGCAGACGGGTGGCTTTAGCAAAGAATCCAT
CCTGCCGAAGCGTAATAGCGATAAACTTATCGCGCGTAAAAAAGACTG
GGACCCAAAGAAATATGGCGGCTTTGATAGCCCGACCGTCGCGTATA
GCGTGTTAGTGGTCGCGAAAGTTGAAAAGGGCAAGAGCAAGAAACTG
AAGTCCGTCAAAGAACTTCTGGGTATCACCATCATGGAACGTAGCTCC
TTTGAGAAGAACCCGATTGACTTCTTAGAGGCGAAGGGTTATAAAGAA
GTCAAAAAAGACCTGATTATCAAGCTGCCGAAGTACAGCCTGTTTGAG
TTGGAGAATGGTCGTAAGCGCATGCTGGCGAGCGCGGGTGAGCTGCA
AAAGGGCAACGAACTGGCGCTGCCGTCGAAATACGTCAATTTTCTGTA
CCTGGCCAGCCACTACGAAAAGCTGAAGGGTTCTCCGGAAGATAACG
AACAAAAGCAACTGTTCGTTGAGCAACATAAACACTACTTGGACGAAAT
CATCGAGCAAATTAGCGAATTTAGCAAACGTGTCATCCTGGCGGACGC
GAATCTGGACAAGGTCCTGTCTGCATACAATAAGCATCGCGACAAACC
AATTCGTGAGCAAGCGGAGAATATCATCCACCTGTTTACGCTGACCAA
CCTAGGTGCGCCGGCGGCATTCAAGTATTTCGATACGACCATCGACC
GCAAGCGCTATACCAGCACCAAAGAGGTCCTGGACGCGACCCTGATC
CACCAGAGCATTACCGGCTTATACGAAACCCGTATTGATTTGAGCCAA
CTGGGTGGCGATGAATTCCCGAAAAAAAAGCGCAAAGTTGGTGGCGG
TGGTAGCCCGAAAAAGAAACGTAAAGTG (SEQ ID NO:63)

Human codon optimized GFP-eSpCas9 DNA sequence

ATGGTTAGCAAAGGTGAAGAACTGTTTACAGGTGTTGTTCCGATTCTG
GTTGAACTGGATGGTGATGTTAATGGCCACAAATTTTCAGTTAGCGGT
GAAGGCGAAGGTGATGCAACCTATGGTAAACTGACCCTGAAATTTATC
TGTACCACCGGCAAACTGCCGGTTCCGTGGCCGACACTGGTTACCAC

ACTGACCTATGGTGTTCAGTGTTTTAGCCGTTATCCGGATCACATGAAA
CAGCACGATTTTTTCAAAAGCGCAATGCCGGAAGGTTATGTTCAAGAA
CGTACCATCTTCTTCAAAGATGACGGCAACTATAAAACCCGTGCCGAA
GTTAAATTTGAAGGTGATACCCTGGTGAATCGCATTGAACTGAAAGGC
ATCGATTTTAAAGAGGATGGTAATATCCTGGGCCACAAACTGGAATATA
ATTATAATAGCCACAACGTGTACATCATGGCCGACAAACAGAAAAATG
GCATCAAAGTGAACTTCAAGATCCGCCATAATATTGAAGATGGTTCAGT
TCAGCTGGCCGATCATTATCAGCAGAATACCCCGATTGGTGATGGTCC
GGTTCTGCTGCCGGATAATCATTATCTGAGCACCCAGAGCAAACTGAG
CAAAGATCCGAATGAAAAACGTGATCACATGGTGCTGCTGGAATTTGT
TACCGCAGCAGGTATTACCTTAGGTATGGATGAACTGTATAAAGTCGA
CAGCGGTGGTAGCAGCGGTGGTTCAAGCGGTAGCGAAACACCGGGTA
CAAGCGAAAGCGCAACACCGGAAAGCAGTGGTGGTAGCTCAGGTGGT
AGTCCGGCAGCAAAACGTGTTAAACTGGACGGTGGTGGTGGTAGCAC
CGGTATGGACAAGAAATACAGCATCGGTTTGGATATTGGCACGAATAG
CGTGGGTTGGGCCGTTATTACCGACGAGTACAAAGTGCCGTCCAAGA
AATTCAAAGTGCTGGGCAATACCGATCGCCATAGCATCAAGAAAAATC
TGATTGGCGCACTGCTGTTCGACAGCGGTGAGACTGCCGAAGCTACG
CGTCTGAAGCGTACGGCGCGTCGTCGCTACACCCGCCGTAAGAACCG
TATTTGCTATCTGCAAGAAATCTTCAGCAACGAAATGGCCAAAGTTGAT
GATAGCTTTTTTCACCGCCTGGAAGAGAGCTTTCTGGTGGAAGAGGAT
AAGAAACACGAGCGCCATCCGATTTTTGGTAACATTGTCGATGAAGTG
GCATACCATGAGAAGTACCCGACCATCTACCACCTTCGTAAGAAACTG
GTGGACAGCACCGATAAAGCTGATCTGCGTCTGATTTACCTGGCGCTG
GCCCACATGATTAAGTTTCGCGGTCATTTTCTGATCGAGGGCGATCTG
AATCCGGACAATTCTGATGTTGACAAGCTGTTTATTCAACTTGTACAGA
CCTACAACCAGTTGTTCGAAGAGAACCCGATCAATGCGAGCGGTGTTG
ATGCCAAAGCAATTCTGAGCGCACGCCTGAGCAAATCTCGCCGTTTGG
AGAACCTGATTGCACAGCTGCCGGGTGAGAAGAAAAACGGTCTGTTC
GGCAATCTGATTGCACTGTCCCTGGGCTTGACCCCGAATTTTAAGAGC
AACTTCGACCTGGCCGAAGATGCGAAGCTCCAATTGAGCAAAGACACC
TACGACGATGACCTGGACAATCTGCTGGCCCAGATTGGCGACCAGTA
CGCAGATCTGTTCTTGGCTGCGAAAAACCTGAGCGATGCAATTCTGCT
GTCGGACATCCTGCGCGTGAATACGGAAATCACGAAAGCGCCTCTGA
GCGCGTCTATGATCAAGCGCTATGACGAGCACCACCAAGATCTGACCC
TGCTGAAAGCTCTGGTGAGACAACAATTGCCAGAGAAGTATAAAGAAA
TTTTCTTTGACCAGAGCAAAAACGGCTATGCGGGTTACATTGACGGTG
GCGCCAGCCAAGAAGAGTTCTACAAATTCATTAAGCCTATCCTGGAGA
AAATGGATGGCACCGAAGAACTGCTGGTAAAGCTGAATCGTGAAGATC
TGCTGCGCAAACAGCGCACTTTTGATAACGGTAGCATTCCGCACCAGA
TCCATCTGGGTGAGTTGCACGCGATTTTGCGTCGCCAGGAAGATTTTT
ATCCGTTCTTGAAAGACAACCGTGAGAAAATCGAGAAAATTCTGACGTT
CCGTATCCCGTATTATGTCGGCCCGCTGGCGCGTGGTAATAGCCGCTT
CGCGTGGATGACCCGCAAATCAGAGGAAACGATTACCCCGTGGAATTT
TGAGGAAGTTGTTGATAAGGGTGCAAGCGCGCAGTCGTTCATTGAGC
GTATGACCAACTTTGACAAGAATTTGCCGAATGAAAAAGTCTTGCCGAA
GCACTCTCTGCTGTACGAGTATTTTACCGTTTACAACGAATTGACCAAG
GTTAAATACGTCACCGAAGGCATGCGCAAACCGGCCTTCCTGAGCGG
CGAGCAGAAAAAAGCAATCGTTGACCTCTTGTTTAAGACCAACCGCAA
GGTTACGGTCAAACAACTGAAAGAGGACTATTTCAAGAAAATTGAATGT

```
TTTGACTCCGTAGAGATCTCCGGTGTTGAGGACCGTTTCAACGCGAGC
CTGGGCACCTACCATGATCTGCTGAAAATTATTAAAGACAAAGATTTTC
TGGACAACGAAGAGAACGAAGATATTCTGGAAGATATCGTTCTGACCC
TGACGCTGTTCGAAGATCGTGAGATGATTGAGGAACGTCTGAAAACCT
ACGCACACTTGTTCGATGACAAAGTTATGAAACAGCTGAAGCGTCGTC
GTTACACAGGTTGGGGCCGTCTGAGCCGTAAGCTTATCAATGGTATCC
GTGACAAACAGAGCGGTAAGACGATTCTGGACTTTCTGAAGTCAGATG
GCTTCGCCAATCGCAACTTTATGCAACTGATTCATGACGACTCTCTGAC
GTTCAAGGAAGATATCCAAAAGGCACAGGTGAGCGGTCAGGGTGATA
GCCTGCATGAGCATATCGCGAACCTGGCGGGTAGCCCGGCTATCAAA
AAGGGTATCTTACAGACTGTGAAAGTTGTGGATGAATTGGTTAAGGTTA
TGGGTCGTCACAAACCGGAAAATATTGTGATCGAGATGGCACGTGAAA
ATCAGACGACGCAAAAGGGTCAAAAAAATTCTCGTGAGCGCATGAAAC
GTATTGAAGAGGGTATCAAAGAATTGGGCAGCCAAATTCTGAAAGAAC
ACCCGGTCGAGAACACCCAGCTGCAAAACGAAAAACTGTATTTATACT
ATCTGCAGAACGGTCGTGACATGTACGTGGATCAAGAACTGGACATCA
ATCGTTTGAGCGATTACGATGTTGATCATATTGTGCCTCAGAGCTTTCT
GGCGGACGATTCGATCGACAACAAAGTGCTGACCCGTAGCGACAAGA
ATCGTGGTAAGAGCGATAACGTGCCGAGCGAAGAAGTCGTTAAGAAAA
TGAAAAACTACTGGCGTCAGCTGCTGAACGCCAAGCTGATTACCCAGC
GTAAGTTCGATAACCTGACGAAAGCCGAGCGTGGAGGCCTGAGCGAG
CTGGACAAGGCCGGCTTTATCAAGCGTCAACTGGTGGAAACCCGTCA
GATCACTAAACATGTGGCACAGATCCTGGACTCCCGCATGAATACGAA
ATATGACGAGAATGACAAGTTGATCCGTGAAGTCAAAGTTATTACGCTG
AAAAGCAAACTGGTGTCCGATTTCCGTAAAGACTTCCAGTTCTATAAAG
TCCGTGAAATCAACAACTATCATCACGCCCACGATGCGTACTTGAACG
CTGTTGTGGGCACCGCACTGATCAAGAAATACCCTGCACTCGAAAGCG
AGTTTGTCTATGGTGACTATAAAGTTTACGACGTGCGTAAGATGATCGC
CAAGAGCGAGCAAGAAATTGGTAAGGCTACCGCAAAGTACTTTTTCTA
CAGCAACATCATGAACTTCTTCAAAACCGAGATTACCCTGGCGAACGG
TGAGATCCGTAAAGCGCCGCTGATTGAGACTAATGGCGAAACGGGCG
AGATTGTGTGGGACAAGGGTCGCGATTTCGCTACGGTTCGTAAGGTCC
TGAGCATGCCGCAAGTTAACATTGTCAAGAAAACTGAAGTGCAGACGG
GTGGCTTTAGCAAAGAATCCATCCTGCCGAAGCGTAATAGCGATAAAC
TTATCGCGCGTAAAAAAGACTGGGACCCAAAGAAATATGGCGGCTTTG
ATAGCCCGACCGTCGCGTATAGCGTGTTAGTGGTCGCGAAAGTTGAAA
AGGGCAAGAGCAAGAAACTGAAGTCCGTCAAAGAACTTCTGGGTATCA
CCATCATGGAACGTAGCTCCTTTGAGAAGAACCCGATTGACTTCTTAG
AGGCGAAGGGTTATAAAGAAGTCAAAAAAGACCTGATTATCAAGCTGC
CGAAGTACAGCCTGTTTGAGTTGGAGAATGGTCGTAAGCGCATGCTGG
CGAGCGCGGGTGAGCTGCAAAAGGGCAACGAACTGGCGCTGCCGTC
GAAATACGTCAATTTTCTGTACCTGGCCAGCCACTACGAAAAGCTGAA
GGGTTCTCCGGAAGATAACGAACAAAAGCAACTGTTCGTTGAGCAACA
TAAACACTACTTGGACGAAATCATCGAGCAAATTAGCGAATTTAGCAAA
CGTGTCATCCTGGCGGACGCGAATCTGGACAAGGTCCTGTCTGCATA
CAATAAGCATCGCGACAAACCAATTCGTGAGCAAGCGGAGAATATCAT
CCACCTGTTTACGCTGACCAACCTAGGTGCGCCGGCGGCATTCAAGTA
TTTCGATACGACCATCGACCGCAAGCGCTATACCAGCACCAAAGAGGT
CCTGGACGCGACCCTGATCCACCAGAGCATTACCGGCTTATACGAAAC
CCGTATTGATTTGAGCCAACTGGGTGGCGATGAATTCCCGAAAAAAAA
```

GCGCAAAGTTGGTGGCGGTGGTAGCCCGAAAAAGAAACGTAAAGTG (SEQ ID NO:64)

Example 2: Editing efficiency comparison with commercial products

[0103] Two commercial GFP-SpCas9 fusion protein products, GenCrispr NLS-Cas9-EGFP Nuclease and ArciTect Cas9-eGFP Nuclease, were purchased from GenScript (Piscataway, NJ) and Stemcell Technologies (Vancouver, Canada), respectively. Human U2OS cells (0.2 x 10^6) and HEK293 cells (0.3 x 10^6) were transfected with 50 pmol of GenCrispr NLS-Cas9-EGFP Nuclease, or ArciTect Cas9-eGFP Nuclease, or the GFP-SpCas9 protein of the current invention, in combination with 150 pmol each of four chemically synthesized sgRNAs targeting the Human EMX1, HEKSite4, VEGFA3, HPRT loci. The guide sequences are: 5'-GAGUCCGAGCAGAAGAAGAA-3' (EMX1) (SEQ ID NO:51), 5'-GGCACUGCGGCUGGAGGUGG-3' (HEKSite4) (SEQ ID NO:52), 5'GGUGAGUGAGUGUGUGCGUG-3' (VEGFA3), and 5'-GGUCACUUUUAACACACCCA-3' (HPRT) (SEQ ID NO:53). Transfection was carried out using Nucleofection Solution V and an Amaxa instrument. Cells were maintained at 37°C and 5% CO_2 for three days before harvested for gene editing analysis. Genomic DNA was prepared using QuickExtract DNA extraction solution. Each targeted genomic region was PCR amplified using a pair of primers consisting of target-specific sequences and next generation sequencing (NGS) adaptors. The primers are listed in the following table:

NGS primer sequences	
Target	Primer sequence (5'-3')
EMX1	Forward: TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNNNNNNCCCCAGTGGCTGCTCT (SEQ ID NO:54) Reverse: GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNNNNNCCAGGCCTCCCCAAAGC (SEQ ID NO:55)
HEKSite4	Forward: TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNNNNNNGGAACCCAGGTAGCCAGAGA (SEQ ID NO:56)

	Reverse: GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNNNNNGGGGTGGGGTCAGACGT (SEQ ID NO:57)
VEGF A3	Forward: TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNNNNNNGCCCATTCCCTCTTTAGCCA (SEQ ID NO:58) Reverse: GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNNNNNGGAGCAGGAAAGTGAGGTTAC (SEQ ID NO:59)
HPRT	Forward: TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNNNNNNAATGGACACATGGGTAGTCAGG (SEQ ID NO:60) Reverse: GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNNNNNGGCTTATATCCAACACTTCGTGGG (SEQ ID NO:61)

[0104] PCR amplicons were analyzed by NGS using the Illumina MiSeq to determine the editing efficiency of each Cas9 protein. The results in **FIG. 2A** and **FIG. 2B** show that the editing efficiencies by the GFP-SpCas9 protein of the current invention were several-fold higher than that of the commercial proteins in all targets.

WHAT IS CLAIMED IS:

1. A fusion protein comprising a Cas9 protein linked to at least one marker protein.

2. The fusion protein of claim 1, wherein the at least one marker protein is linked to the Cas9 protein directly via a chemical bond, directly or indirectly via a linker, or a combination thereof.

3. The fusion protein of claim 1, wherein the at least one marker protein is linked to the Cas9 protein directly or indirectly via a linker.

4. The fusion protein of any one of claims 1 to 3, wherein the at least one marker protein is linked to the Cas9 protein at its N-terminus, C-terminus, an internal location, or a combination thereof.

5. The fusion protein of any one of claims 1 to 4, further comprising at least one nuclear localization signal.

6. The fusion protein of claim 5, wherein the nuclear localization signal, the marker protein, the linker, and the Cas9 protein are arranged in the following order (N-terminus to C-terminus):

   marker protein – linker – nuclear localization signal – Cas9 protein;

   marker protein – nuclear localization signal – linker – Cas9 protein;

   nuclear localization signal – linker – marker protein – Cas9 protein;

   nuclear localization signal – marker protein – linker – Cas9 protein;

   marker protein – linker – nuclear localization signal – linker – Cas9 protein; or nuclear localization signal – linker – marker protein – linker – Cas9 protein.

7. The fusion protein of claim 6, wherein the nuclear localization signal, the marker protein, the linker, and the Cas9 protein are arranged in the following order (N-terminus to C-terminus):

   marker protein – linker – nuclear localization signal – Cas9 protein.

8. The fusion protein of any one of claims 1 to 7, further comprising at least one heterologous domain.

9. The fusion protein of claim 8, wherein the at least one heterologous domain is a cell-penetrating domain, a chromatin modulating motif, an epigenetic modification domain, a transcriptional regulation domain, an RNA aptamer binding domain, or combination thereof.

10. The fusion protein of any one of claims 1 to 9, wherein the fusion protein is a nuclease and cleaves both strands of a double-stranded sequence, is a nickase and cleaves one strand of a double-stranded sequence, or has no nuclease or nickase activity.

11. The fusion protein of any one of claims 1 to 10, wherein the cas9 protein is from *Streptococcus pyogenes, Streptococcus thermophilus*, *Neisseria meningitidis*, *Staphylococcus aureus*, or *Campylobacter jejuni*.

12. The fusion protein of any one of claims 1 to 11, wherein the marker protein has an amino acid sequence having at least 90% sequence identity with SEQ ID NO:19 or 20.

13. The fusion protein of any one of claims 1 to 11, wherein the marker protein has an amino acid sequence set forth in SEQ ID NO:19 or 20.

14. The fusion protein of any one of claims 3 to 13, wherein the linker has an amino acid sequence having at least 90% sequence identity with SEQ ID NO:35 or 36.

15. The fusion protein of any one of claims 3 to 13, wherein the linker has an amino acid sequence set forth in SEQ ID NO:35 or 36.

16. The fusion protein of any one of claims 3 to 15, wherein the fusion protein has an amino acid sequence having at least 90% sequence identity with SEQ ID NO:48, 49, or 50.

17. The fusion protein of any one of claims 3 to 15, wherein the fusion protein has an amino acid sequence as set forth in SEQ ID NO:48, 49, or 50.

18. A system comprising the fusion protein of any one of claims 1 to 17 and an engineered guide RNA.

19. The system of claim 18, wherein the engineered guide RNA is a single molecule.

20. The system of any one of claims 18 to 19, wherein the engineered guide RNA sequence is optimized to facilitate base-paring within the engineered guide RNA, minimize base-paring within the engineered guide RNA, increase stability of the engineered guide RNA, facilitate transcription of the engineered guide RNA in a eukaryotic cell, or a combination thereof.

21. A plurality of nucleic acids encoding the fusion protein of any one of claims 1 to 17.

22. A plurality of nucleic acids encoding the system of any one of claims 18 to 20, the plurality of nucleic acid comprising at least one nucleic acid encoding the fusion protein, and at least one nucleic acid encoding the engineered guide RNA.

23. The plurality of nucleic acids of claim 14, wherein the at least one nucleic acid encoding the fusion protein is RNA.

24. The plurality of nucleic acids of claim 14, wherein the at least one nucleic acid encoding the fusion protein is DNA.

25. The plurality of nucleic acids of any one of claims 21 to 24, wherein the at least one nucleic acid encoding the fusion protein is codon optimized for expression in a eukaryotic cell.

26. The plurality of nucleic acids of claim 25, wherein the eukaryotic cell is a human cell, a non-human mammalian cell, a non-mammalian vertebrate cell, an invertebrate cell, a plant cell, or a single cell eukaryotic organism.

27. The plurality of nucleic acids of any one of claims 22 to 26, wherein the at least one nucleic acid encoding the engineered guide RNA is DNA.

28. The plurality of nucleic acids of any one of claims 22 to 27, wherein the at least one nucleic acid encoding the fusion protein is operably linked to a phage promoter sequence for *in vitro* RNA synthesis or protein expression in a bacterial cell, and the at least one nucleic acid encoding the engineered guide RNA is operably linked to a phage promoter sequence for *in vitro* RNA synthesis.

29. The plurality of nucleic acids of any one of claims 22 to 27, wherein the at least one nucleic acid encoding the fusion protein is operably linked to a eukaryotic promoter sequence for expression in a eukaryotic cell, and the at least one nucleic acid encoding the engineered guide RNA is operably linked to a eukaryotic promoter sequence for expression in a eukaryotic cell.

30. At least one vector comprising the plurality of nucleic acids of any one of claims 21 to 29.

31. The at least one vector of claim 30, which is a plasmid vector, a viral vector, or a self-replicating viral RNA replicon.

32. A eukaryotic cell comprising at least one system comprising a fusion protein as defined in claims 1 to 17, a system as defined in claims 18 to 20, a plurality of nucleic acids as defined in claims 21 to 29, or at least one vector as defined in claims 30 to 31.

33. The eukaryotic cell of claim 32, which is a human cell, a non-human mammalian cell, a plant cell, a non-mammalian vertebrate cell, an invertebrate cell, or a single cell eukaryotic organism.

34. The eukaryotic cell of any one of claims 32 to 33, which is *in vivo*, *ex vivo*, or *in vitro*.

35. A method for determining chromosome identity and location within a living eukaryotic cell or chemically fixed eukaryotic cell, the method comprising introducing the fusion protein, system, plurality of nucleic acids, or vector of any one of claims 1 to 31 into the living or chemically fixed eukaryotic cell and detecting a signal from the marker protein.

36. The method of claim 35, wherein the eukaryotic cell is a human cell, a non-human mammalian cell, a plant cell, a non-mammalian vertebrate cell, an invertebrate cell, or a single cell eukaryotic organism.

37. The method of any one of claims 35 to 36, wherein the eukaryotic cell is *in vivo, ex vivo,* or *in vitro*.

FIG. 1

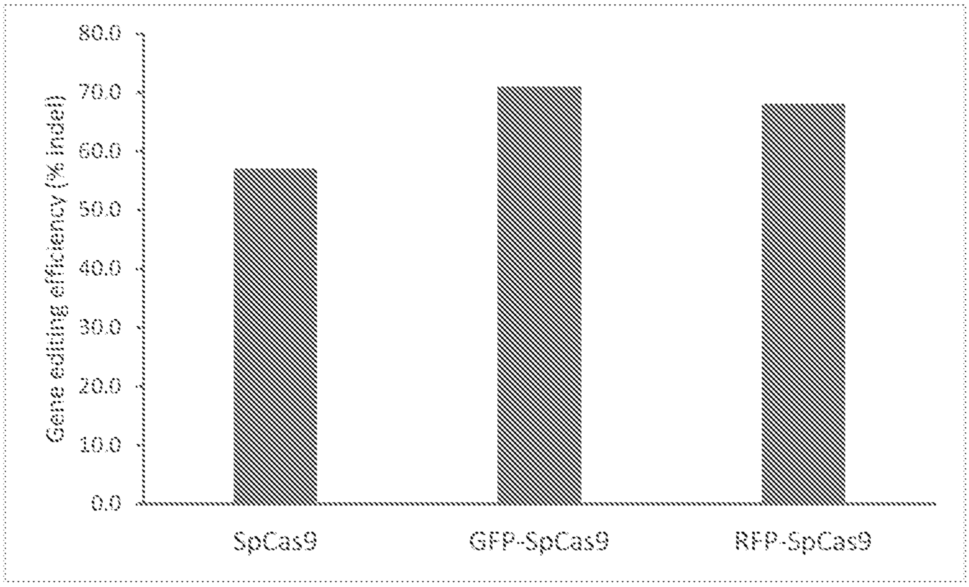

80.0
70.0
60.0
50.0
40.0
30.0
20.0
10.0
0.0
Gene editing efficiency (% indel)
SpCas9
GFP-SpCas9
RFP-SpCas9

FIG. 2A

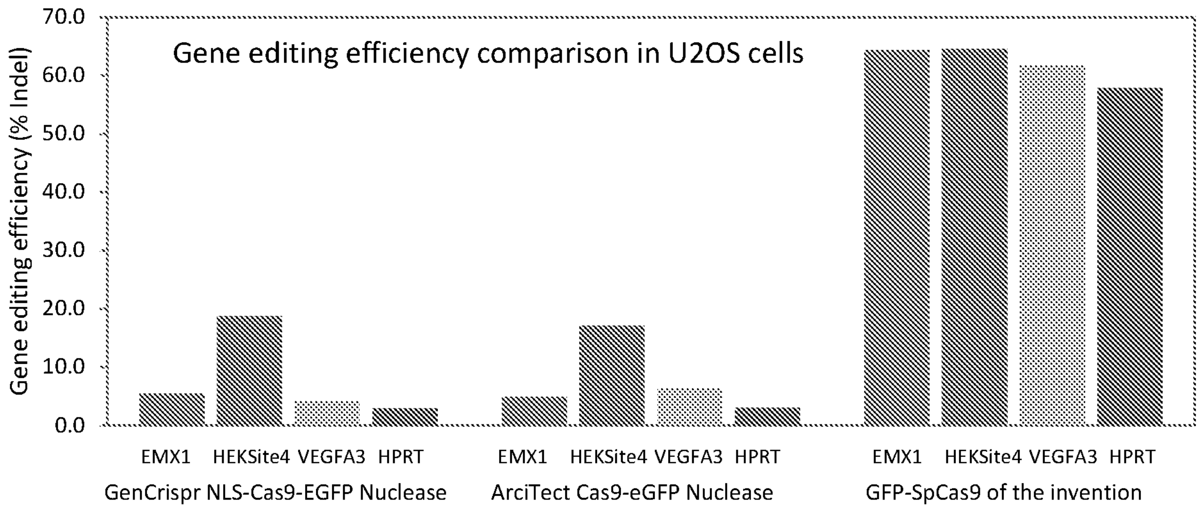
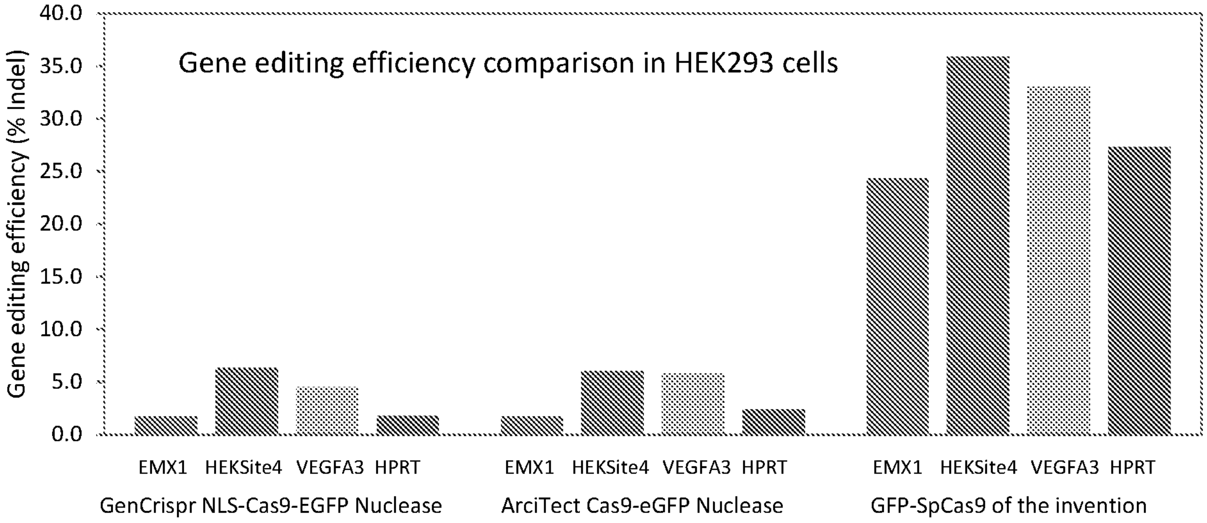

Gene editing efficiency comparison in U2OS cells
Gene editing efficiency (% Indel)
70.0
60.0
50.0
40.0
30.0
20.0
10.0
0.0
EMX1
HEKSite4
VEGFA3
HPRT
GenCrispr NLS-Cas9-EGFP Nuclease
EMX1
HEKSite4
VEGFA3
HPRT
ArciTect Cas9-eGFP Nuclease
EMX1
HEKSite4
VEGFA3
HPRT
GFP-SpCas9 of the invention

FIG. 2B


Gene editing efficiency comparison in HEK293 cells
Gene editing efficiency (% Indel)
40.0
35.0
30.0
25.0
20.0
15.0
10.0
5.0
0.0
EMX1
HEKSite4
VEGFA3
HPRT
GenCrispr NLS-Cas9-EGFP Nuclease
EMX1
HEKSite4
VEGFA3
HPRT
ArciTect Cas9-eGFP Nuclease
EMX1
HEKSite4
VEGFA3
HPRT
GFP-SpCas9 of the invention

INTERNATIONAL SEARCH REPORT	International application No. PCT/US2020/018145

Box No. I Nucleotide and/or amino acid sequence(s) (Continuation of item 1.c of the first sheet)

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

   a. ☒ forming part of the international application as filed:

      ☒ in the form of an Annex C/ST.25 text file.

      □ on paper or in the form of an image file.

   b. □ furnished together with the international application under PCT Rule 13*ter*.1(a) for the purposes of international search only in the form of an Annex C/ST.25 text file.

   c. □ furnished subsequent to the international filing date for the purposes of international search only:

      □ in the form of an Annex C/ST.25 text file (Rule 13*ter*.1(a)).

      □ on paper or in the form of an image file (Rule 13*ter*.1(b) and Administrative Instructions, Section 713).

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

3. Additional comments:

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2020/018145

A. CLASSIFICATION OF SUBJECT MATTER
INV. C12N9/22 C12N15/10
ADD.

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
C12N C40B C07K

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

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

EPO-Internal, Sequence Search, WPI Data, BIOSIS

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X Y A	WO 2014/186686 A2 (TWO BLADES FOUNDATION [US]) 20 November 2014 (2014-11-20) sequence 4	1-13,16, 18-37 14,15 17
X Y A	WO 2017/131150 A1 (UNIV TOKYO [JP]) 3 August 2017 (2017-08-03) sequence 24 figure 7a	1-13,16, 18-37 14,15 17

-/--

☒ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

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

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

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

"&" document member of the same patent family

Date of the actual completion of the international search 14 May 2020	Date of mailing of the international search report 28/05/2020
Name and mailing address of the ISA/ European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Authorized officer Niebuhr-Ebel, K

1

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	BAOHUI CHEN ET AL: "Dynamic Imaging of Genomic Loci in Living Human Cells by an Optimized CRISPR/Cas System", CELL, vol. 155, no. 7, 1 December 2013 (2013-12-01), pages 1479-1491, XP055181416, ISSN: 0092-8674, DOI: 10.1016/j.cell.2013.12.001	1-13,16, 18-37
A	abstract page 1481, left-hand column, paragraph 2; figure 1B -----	14,15,17
Y	G W Platt ET AL: "GSP:BFZ61813, Fusion protein preparation related rigid linker", , 10 January 2019 (2019-01-10), XP055695105, Retrieved from the Internet: URL:http://ibis.internal.epo.org/exam/dbfetch.jsp?id=GSP:BFZ61813 [retrieved on 2020-05-13] sequence -----	14,15
Y	DATABASE Geneseq [Online] 1 November 2018 (2018-11-01), "Fusion protein constructing linker peptide, SEQ ID 385.", XP55695103, retrieved from EBI accession no. GSP:BFR18056 Database accession no. BFR18056 sequence -----	14,15

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/US2020/018145

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2014186686 A2	20-11-2014	NONE	
WO 2017131150 A1	03-08-2017	JP WO2017131150 A1 WO 2017131150 A1	29-11-2018 03-08-2017