US 20250250553A1

(19) **United States**

(12) **Patent Application Publication** (10) **Pub. No.: US 2025/0250553 A1**

Zhang et al. (43) **Pub. Date: Aug. 7, 2025**

(54) **ENGINEERING OF SYSTEMS, METHODS AND OPTIMIZED GUIDE COMPOSITIONS FOR SEQUENCE MANIPULATION**

(71) Applicants: **The Broad Institute, Inc.**, Cambridge, MA (US); **Massachusetts Institute of Technology**, Cambridge, MA (US); **President and Fellows of Harvard College**, Cambridge, MA (US)

(72) Inventors: **Feng Zhang**, Cambridge, MA (US); **Le Cong**, Cambridge, MA (US); **Patrick Hsu**, Cambridge, MA (US); **Fei Ran**, Cambridge, MA (US)

(73) Assignees: **The Broad Institute, Inc.**, Cambridge, MA (US); **Massachusetts Institute of Technology**, Cambridge, MA (US); **President and Fellows of Harvard College**, Cambridge, MA (US)

(21) Appl. No.: **19/025,692**

(22) Filed: **Jan. 16, 2025**

Related U.S. Application Data

(63) Continuation of application No. 15/230,025, filed on Aug. 5, 2016, which is a continuation of application No. 14/104,990, filed on Dec. 12, 2013, now abandoned.

(60) Provisional application No. 61/736,527, filed on Dec. 12, 2012, provisional application No. 61/748,427, filed on Jan. 2, 2013, provisional application No. 61/758,468, filed on Jan. 30, 2013, provisional application No. 61/769,046, filed on Feb. 25, 2013, provisional application No. 61/802,174, filed on Mar. 15, 2013, provisional application No. 61/791,409, filed on Mar. 15, 2013, provisional application No. 61/806,375, filed on Mar. 28, 2013, provisional application No. 61/814,263, filed on Apr. 20, 2013, provisional application No. 61/819,803, filed on May 6, 2013, provisional application No. 61/828,130, filed on May 28, 2013, provisional application No. 61/836,127, filed on Jun. 17, 2013, provisional application No. 61/835,931, filed on Jun. 17, 2013.

Publication Classification

(51) **Int. Cl.**

C12N 9/22	(2006.01)
C12N 9/16	(2006.01)
C12N 15/01	(2006.01)
C12N 15/10	(2006.01)
C12N 15/52	(2006.01)
C12N 15/63	(2006.01)
C12N 15/79	(2006.01)
C12N 15/85	(2006.01)
C12N 15/86	(2006.01)
C12N 15/90	(2006.01)
C12Q 1/6806	(2018.01)

(52) **U.S. Cl.**

CPC ***C12N 9/22*** (2013.01); ***C12N 9/16*** (2013.01); ***C12N 15/01*** (2013.01); ***C12N 15/52*** (2013.01); ***C12N 15/63*** (2013.01); ***C12N 15/85*** (2013.01); ***C12N 15/86*** (2013.01); ***C12N 15/902*** (2013.01); ***C12N 15/907*** (2013.01); ***C12Q 1/6806*** (2013.01); ***C12Y 301/00*** (2013.01); *C12N 15/1082* (2013.01); *C12N 15/79* (2013.01); *C12N 2310/20* (2017.05); *C12N 2800/10* (2013.01); *C12N 2810/50* (2013.01)

(57) **ABSTRACT**

The invention provides for systems, methods, and compositions for manipulation of sequences and/or activities of target sequences. Provided are vectors and vector systems, some of which encode one or more components of a CRISPR complex, as well as methods for the design and use of such vectors. Also provided are methods of directing CRISPR complex formation in eukaryotic cells and methods for selecting specific cells by introducing precise mutations utilizing the CRISPR-Cas system.

Specification includes a Sequence Listing.

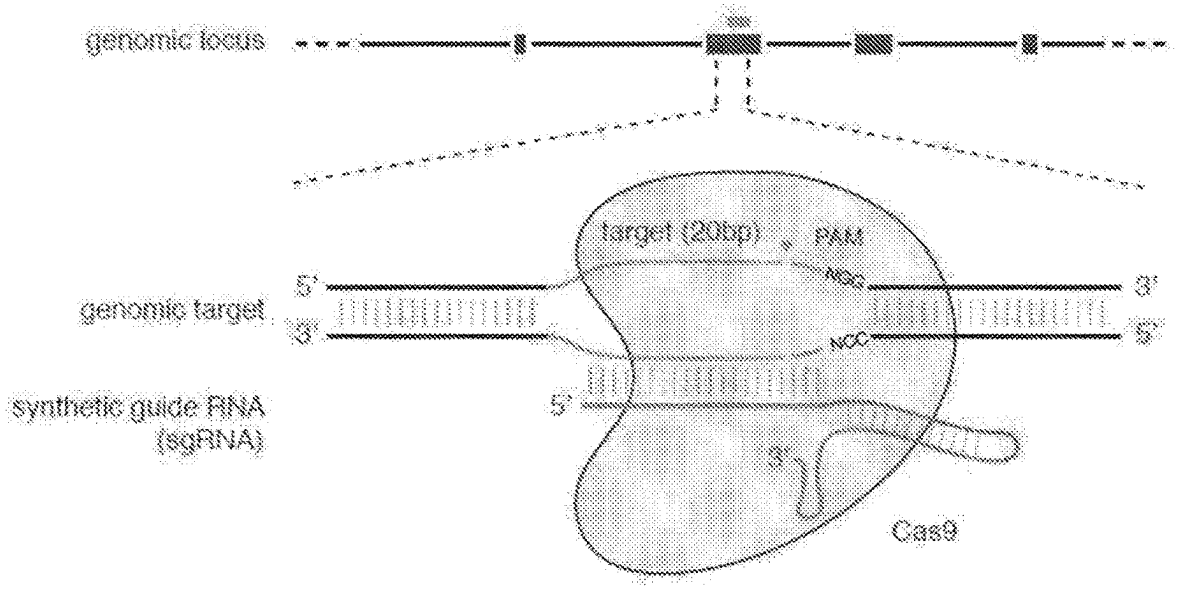

genomic locus
target (20bp)
PAM
NGG
NCC
genomic target
5'
3'
3'
5'
synthetic guide RNA
(sgRNA)
5'
3'
Cas9

FIG. 1

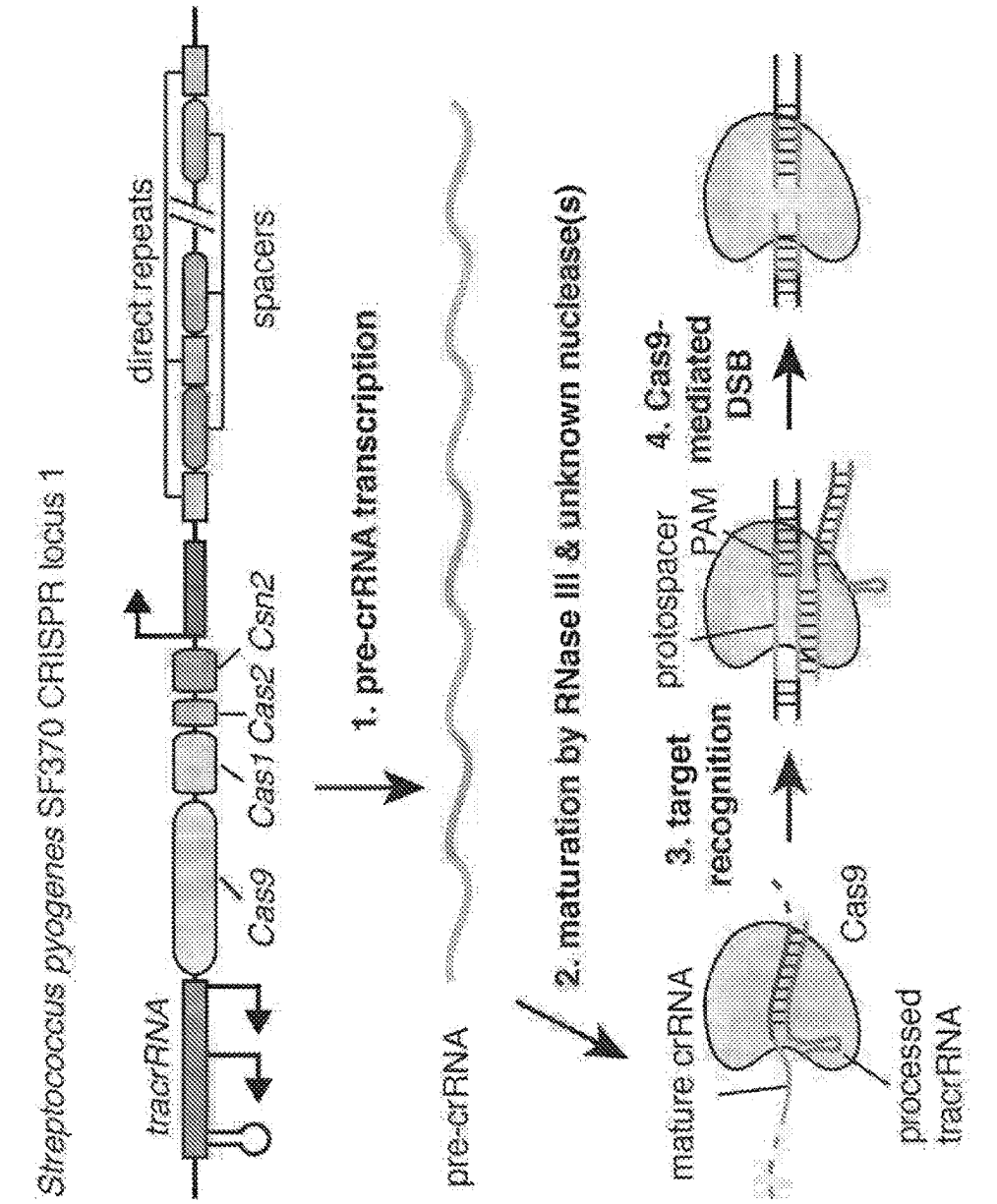

A
Streptococcus pyogenes SF370 CRISPR locus 1
tracrRNA
Cas9
Cas1
Cas2
Csn2
direct repeats
spacers
1. pre-crRNA transcription
pre-crRNA
2. maturation by RNase III & unknown nuclease(s)
mature crRNA
3. target recognition
protospacer
PAM
4. Cas9-mediated DSB
Cas9
processed tracrRNA

FIG. 2A

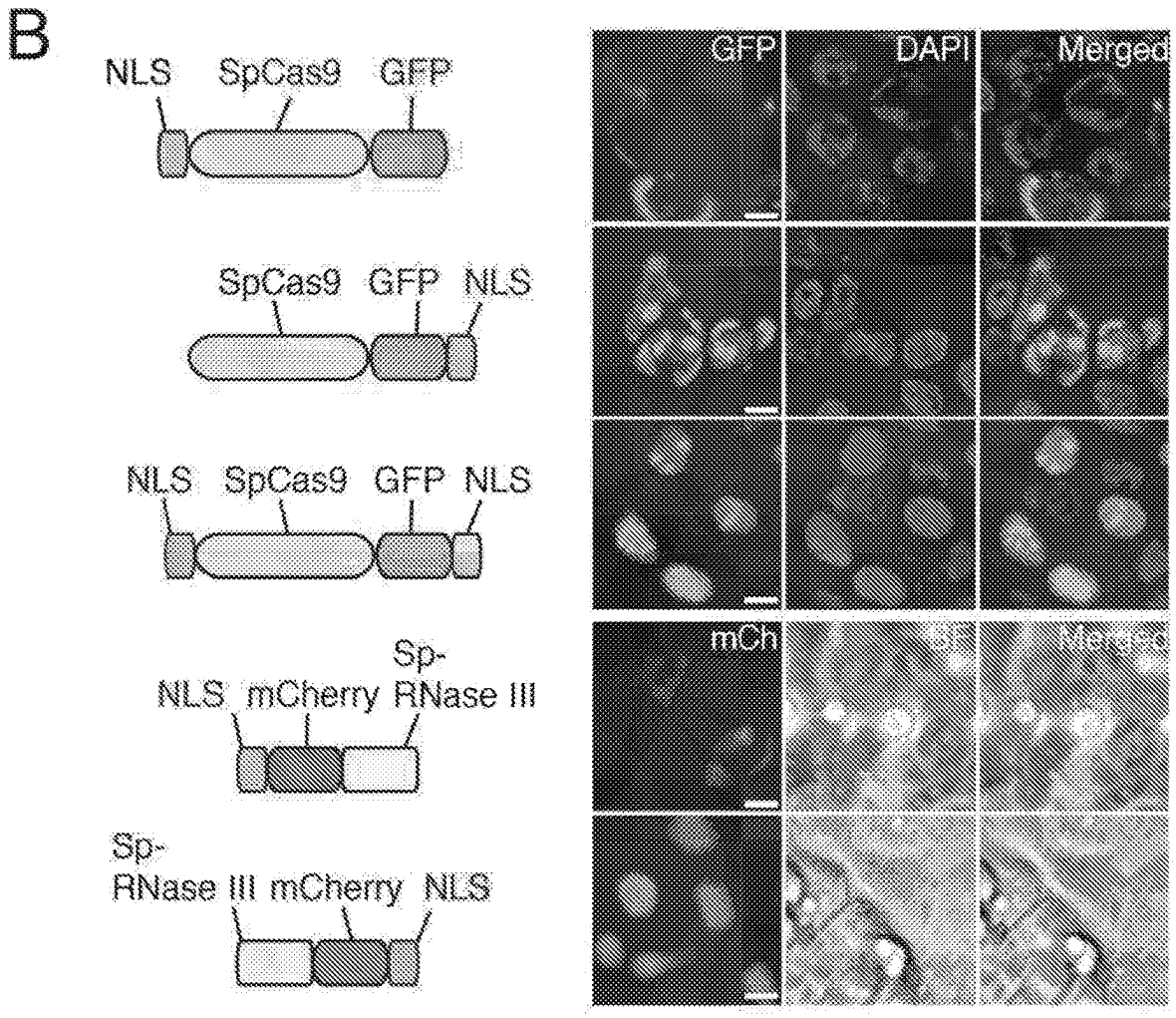

B
NLS SpCas9 GFP
SpCas9 GFP NLS
NLS SpCas9 GFP NLS
NLS mCherry Sp-RNase III
Sp-RNase III mCherry NLS
GFP
DAPI
Merged
mCh

FIG. 2B

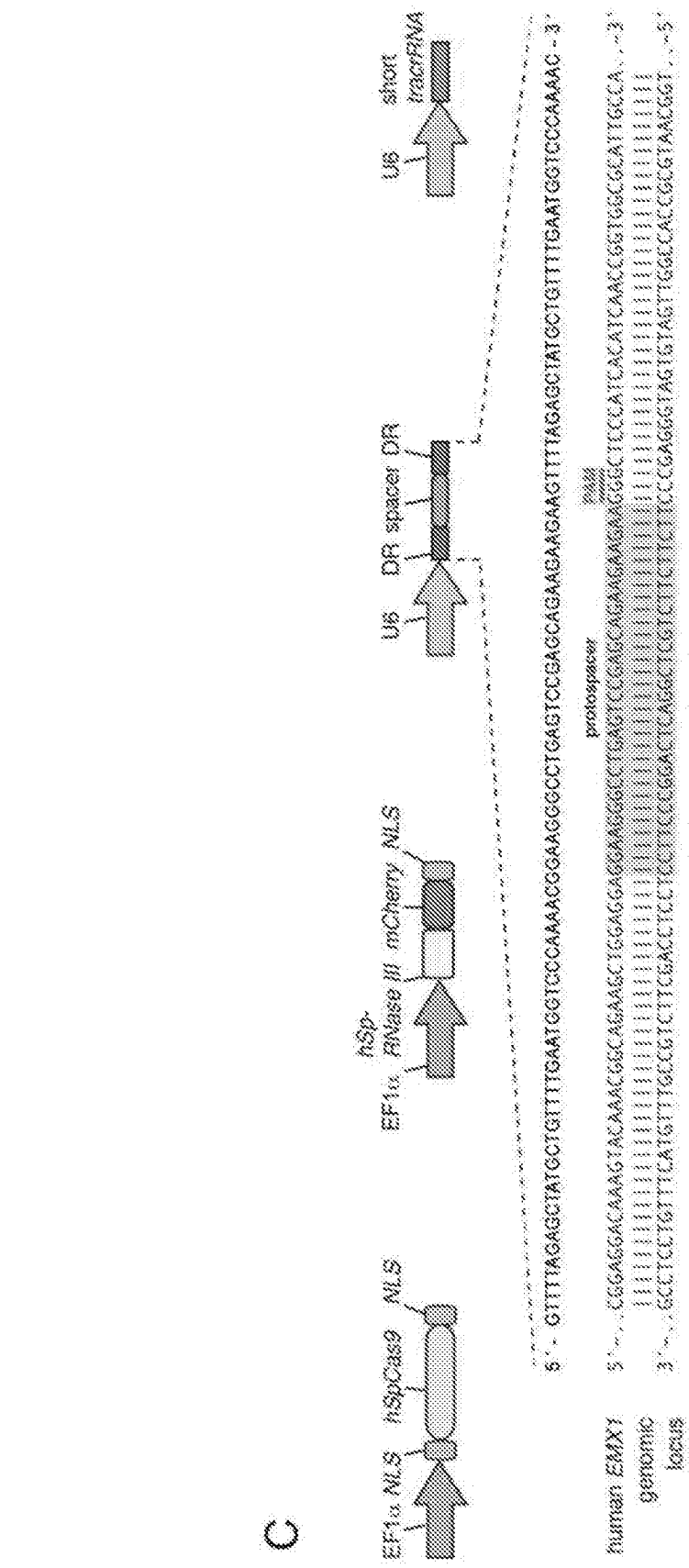

C
EF1α NLS hSpCas9 NLS
EF1α hSp RNase III mCherry NLS
U6 DR spacer DR
U6 short tracrRNA
5' - GTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAACGGAAGGCCCTGACTCCGAGCCAGAAGAAGAAGTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAAC - 3'
protospacer
PAM
human EMX1 genomic locus
5' -...CGGAGGACAAAGTACAAACGGCCAGAAGCTGGAGGAGGAAGAGCCTGACTCCGAGCCAGAAGAAGAAGGGCCTCCCACATCAACCGGTGGCGCATTGCCA...-3'
3' -...GCCTCCTGTTTCATGTTTGCCGGTCTTCGACCTCCTCCTTCTCGGACTCAGGCTCGGTCTTCTTCTTCCCGGAGGGTAGTGTAGTTGGCCACCGCGTAACGGT...-5'

FIG. 2C

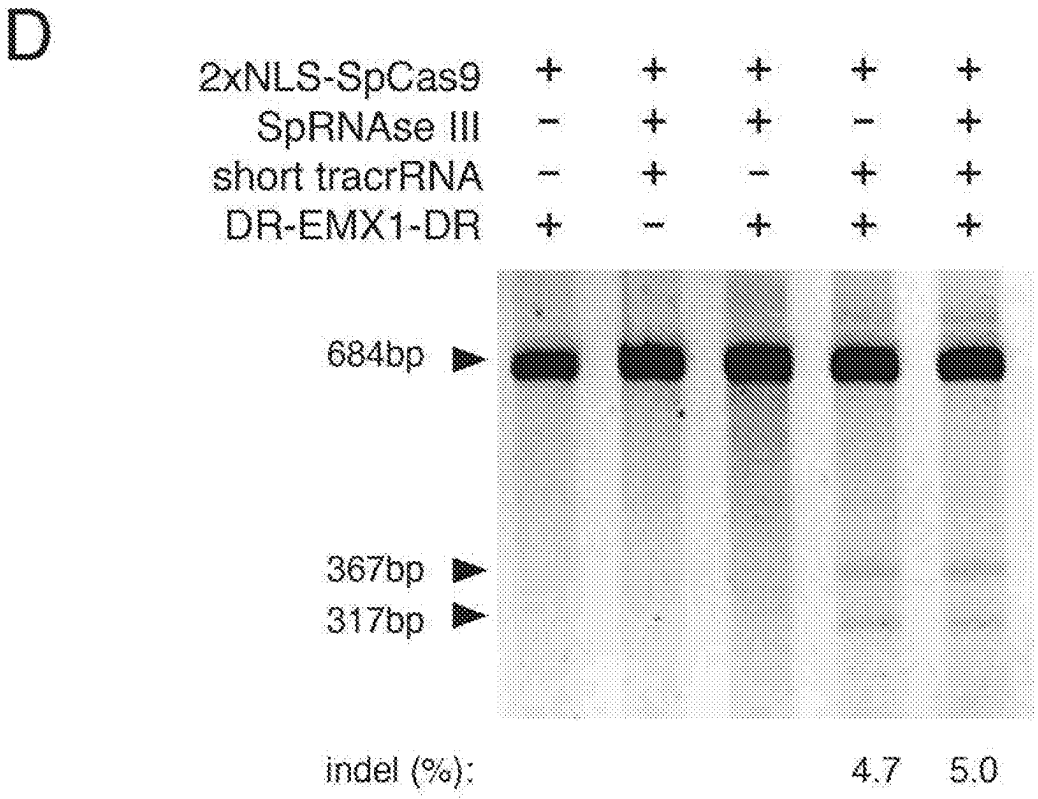

D
2xNLS-SpCas9 + + + + +
SpRNAse III − + + − +
short tracrRNA − + − + +
DR-EMX1-DR + − + + +
684bp
367bp
317bp
indel (%): 4.7 5.0

FIG. 2D

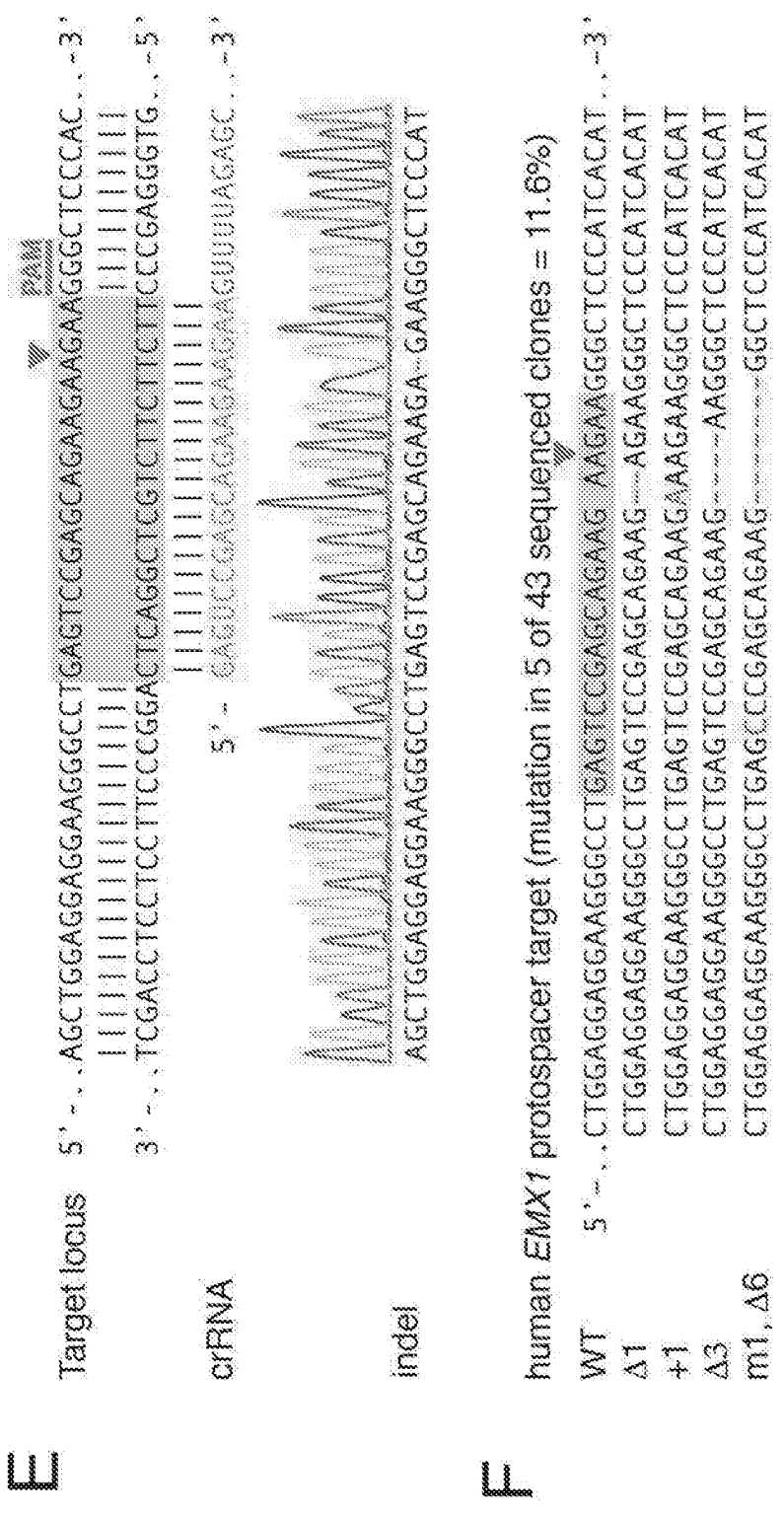

E
Target locus
PAM
5' - . . AGCTGGAGGAGGAAGGGCCTGAGTCCGAGCAGAAGAAGAAGGGCTCCCAC . . -3'
3' - . . TCGACCTCCTCCTTCCCGGACTCAGGCTCGTCTTCTTCTTCCCGAGGGTG . . -5'
crRNA
5' - CAGUCCGAGCAGAAGAAGUUUUAGAGC . . -3'
indel
AGCTGGAGGAGGAAGGGCCTGAGTCCGAGCAGAAGA-GAAGGGCTCCCAT
F
human EMX1 protospacer target (mutation in 5 of 43 sequenced clones = 11.6%)
WT 5' - . . CTGGAGGAGGAAGGGCCTGAGTCCGAGCAGAAG AAGAAGGGCTCCCATCACAT . . -3'
Δ1 CTGGAGGAGGAAGGGCCTGAGTCCGAGCAGAAG---AGAAGGGCTCCCATCACAT
+1 CTGGAGGAGGAAGGGCCTGAGTCCGAGCAGAAGAAAGAAGGGCTCCCATCACAT
Δ3 CTGGAGGAGGAAGGGCCTGAGTCCGAGCAGAAG-----AAGGGCTCCCATCACAT
m1, Δ6 CTGGAGGAGGAAGGGCCTGAGCCCGAGCAGAAG----------GGCTCCCATCACAT

FIG. 2E-F

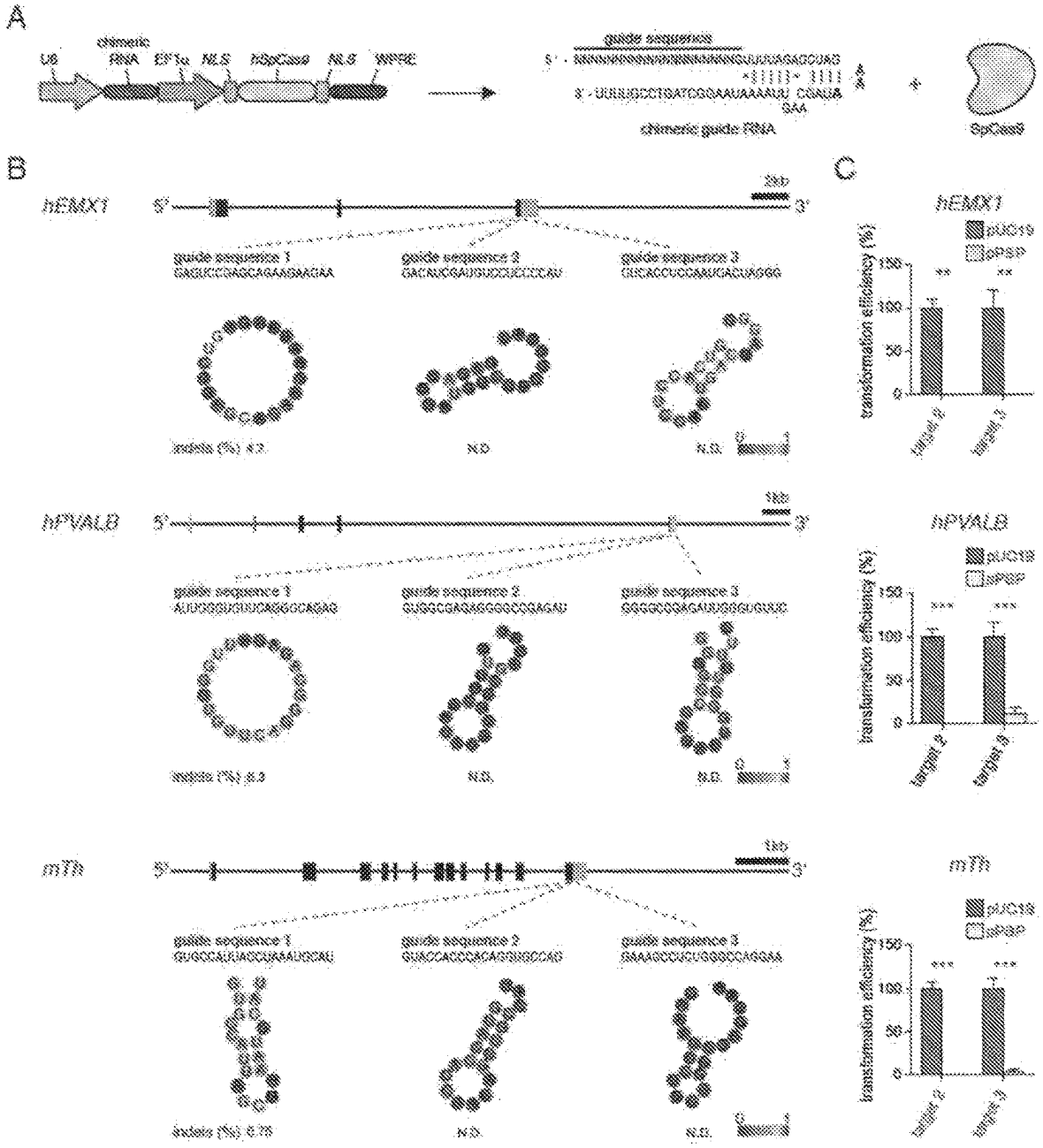
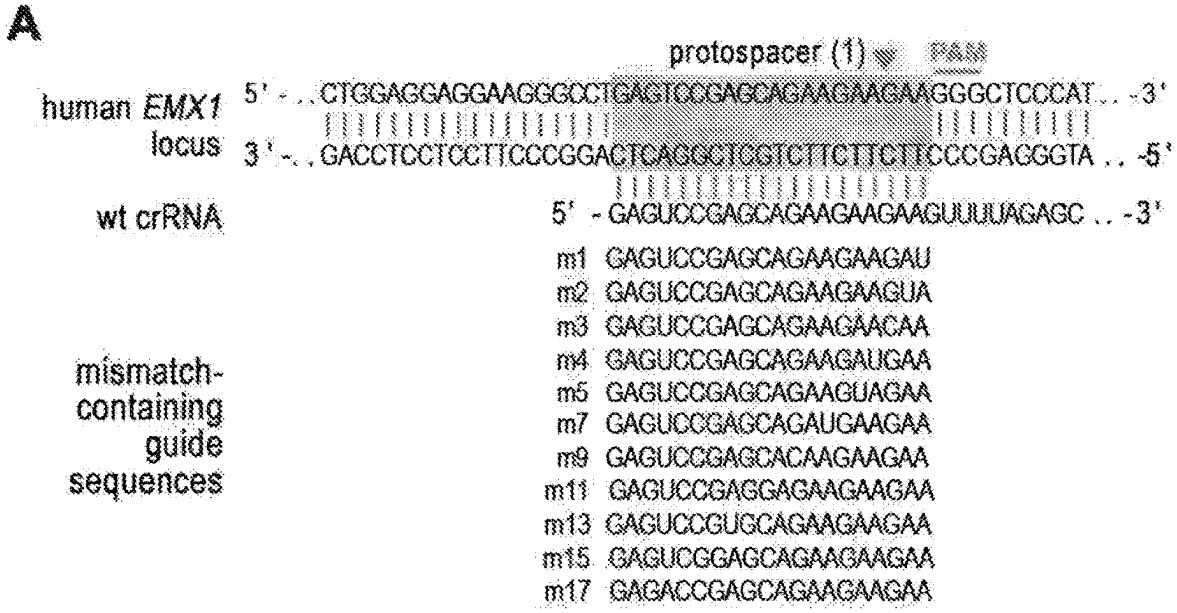
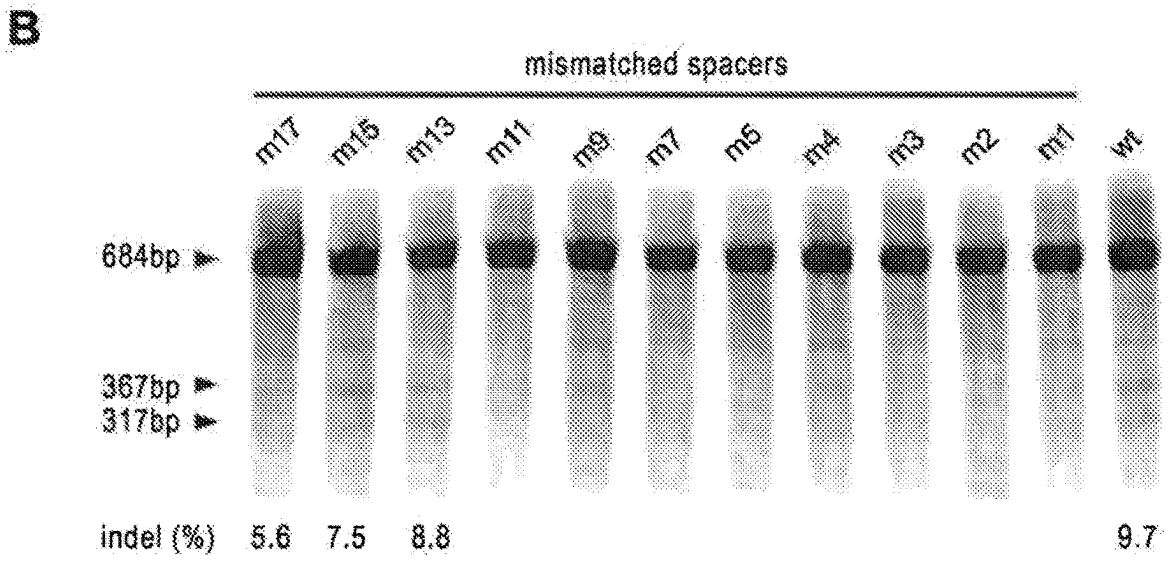

A
U6
chimeric RNA
EF1α
NLS
hSpCas9
NLS
WPRE
guide sequence
chimeric guide RNA
SpCas9
B
hEMX1
hPVALB
mTh
guide sequence 1
guide sequence 2
guide sequence 3
N.D.
C
hEMX1
hPVALB
mTh
pUC19
sPSP
transformation efficiency (%)
target 2
target 3

FIG. 3A-C


A
protospacer (1)
PAM
human EMX1 locus
5' - . . CTGGAGGAGGAAGGGCCTGAGTCCGAGCAGAAGAAGAAGGGCTCCCAT . . -3'
3' -. . GACCTCCTCCTTCCCGGACTCAGGCTCGTCTTCTTCTTCCCGAGGGTA . . -5'
wt crRNA
5' - GAGUCCGAGCAGAAGAAGAAGUUUUAGAGC . . -3'
mismatch-containing guide sequences
m1 GAGUCCGAGCAGAAGAAGAU
m2 GAGUCCGAGCAGAAGAAGUA
m3 GAGUCCGAGCAGAAGAACAA
m4 GAGUCCGAGCAGAAGAUGAA
m5 GAGUCCGAGCAGAAGUAGAA
m7 GAGUCCGAGCAGAUGAAGAA
m9 GAGUCCGAGCACAAGAAGAA
m11 GAGUCCGAGGAGAAGAAGAA
m13 GAGUCCGUGCAGAAGAAGAA
m15 GAGUCGGAGCAGAAGAAGAA
m17 GAGACCGAGCAGAAGAAGAA


B
mismatched spacers
m17
m15
m13
m11
m9
m7
m5
m4
m3
m2
m1
wt
684bp
367bp
317bp
indel (%)
5.6
7.5
8.8
9.7

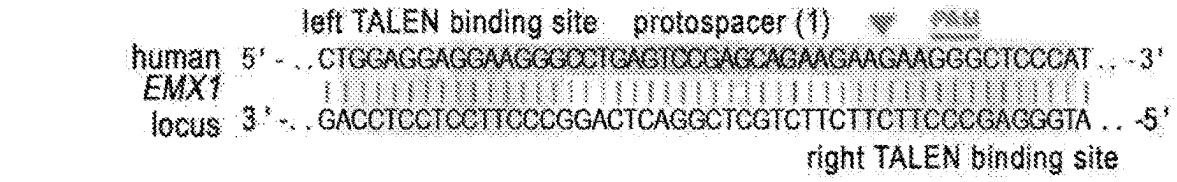

C
left TALEN binding site
protospacer (1)
PAM
human EMX1 locus
5' - . . CTGGAGGAGGAAGGGCCTGAGTCCGAGCAGAAGAAGAAGGGCTCCCAT . . -3'
3' -. . GACCTCCTCCTTCCCGGACTCAGGCTCGTCTTCTTCTTCCCGAGGGTA . . -5'
right TALEN binding site

FIG. 4A-C

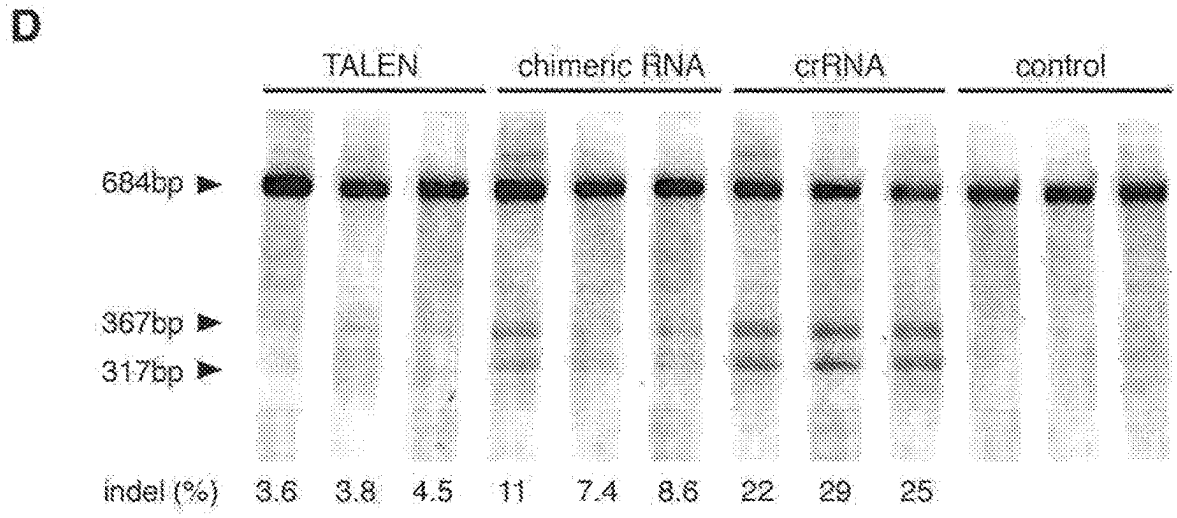

D
TALEN
chimeric RNA
crRNA
control
684bp
367bp
317bp
indel (%)
3.6
3.8
4.5
11
7.4
8.6
22
29
25

FIG. 4D

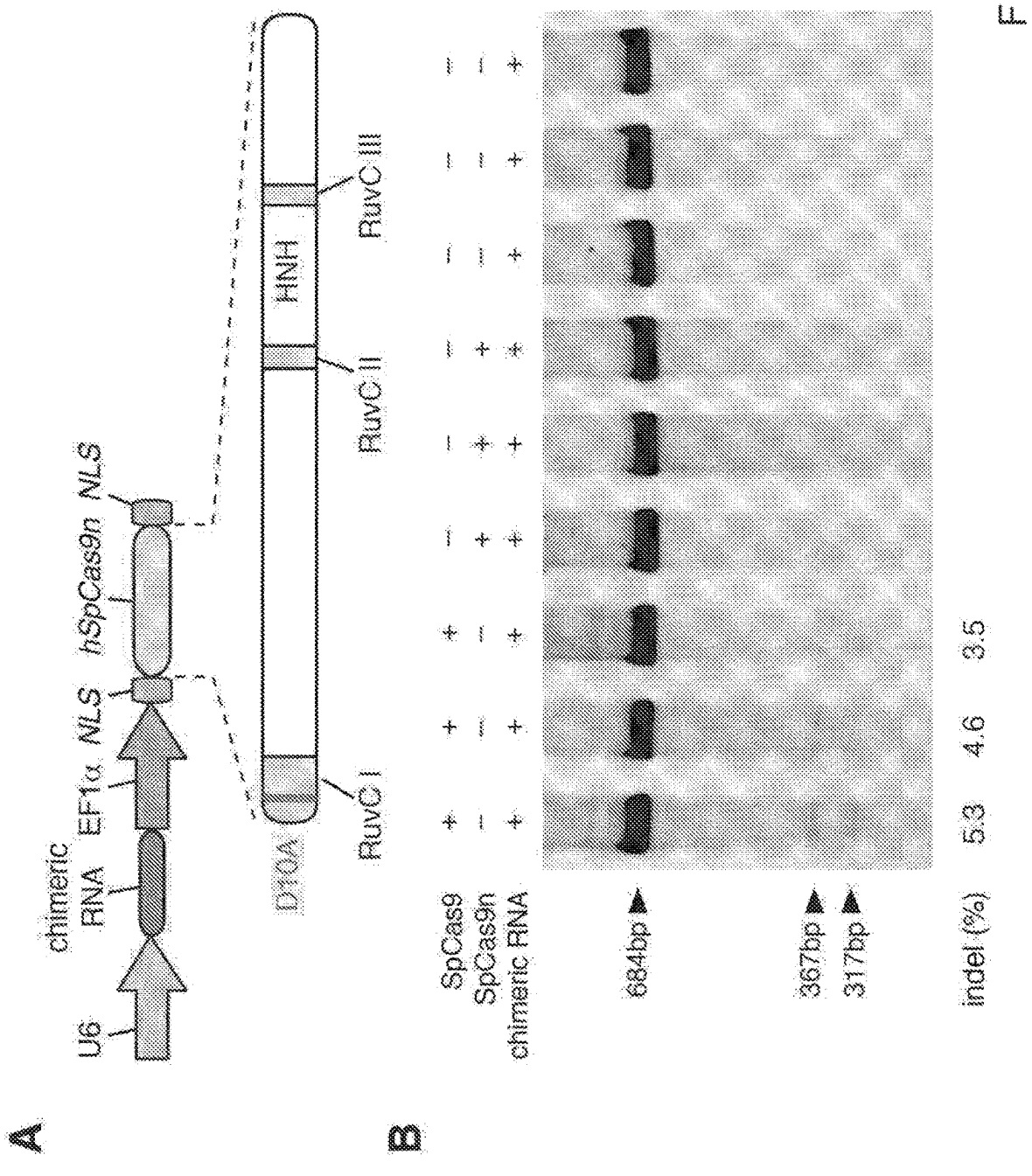

A
U6
chimeric RNA
EF1α
NLS
hSpCas9n
NLS
D10A
RuvC I
RuvC II
HNH
RuvC III
B
SpCas9
SpCas9n
chimeric RNA
684bp
367bp
317bp
indel (%) 5.3 4.6 3.5

FIG. 5A-B

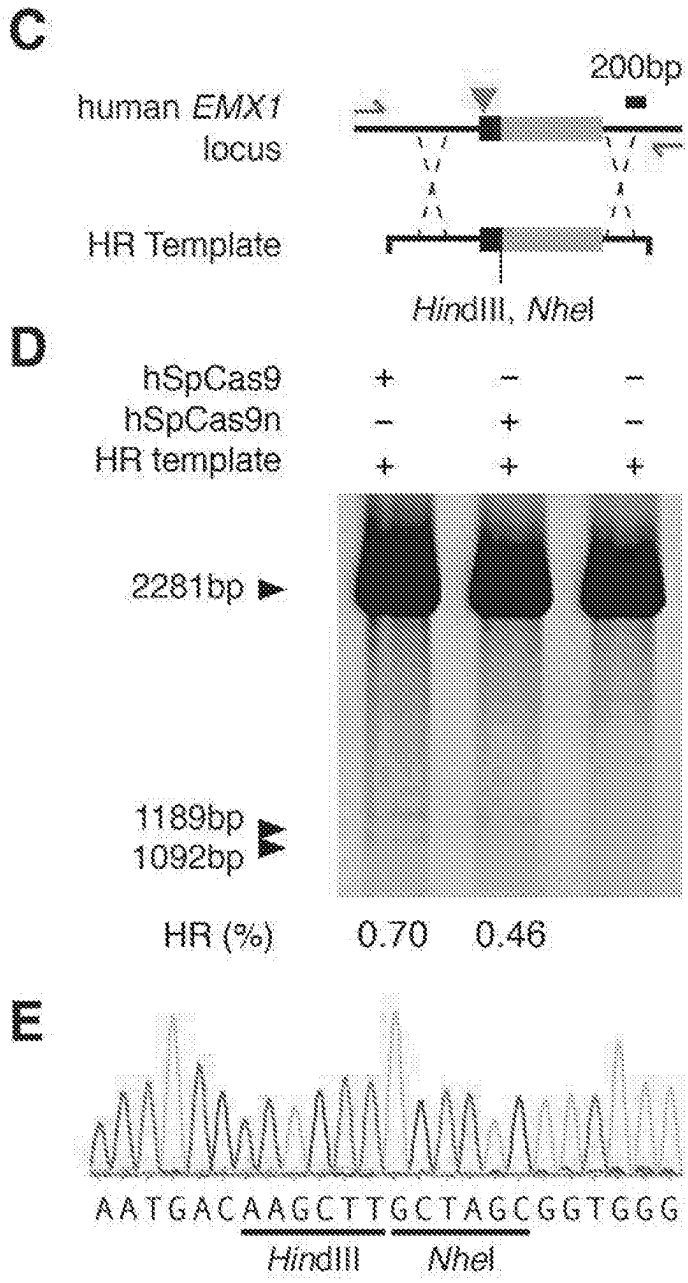

C
200bp
human EMX1 locus
HR Template
HindIII, NheI
D
hSpCas9
hSpCas9n
HR template
+ – –
– + –
+ + +
2281bp
1189bp
1092bp
HR (%)
0.70 0.46
E
AATGACAAGCTTGCTAGCGGTGGG
HindIII
NheI

FIG. 5C-E

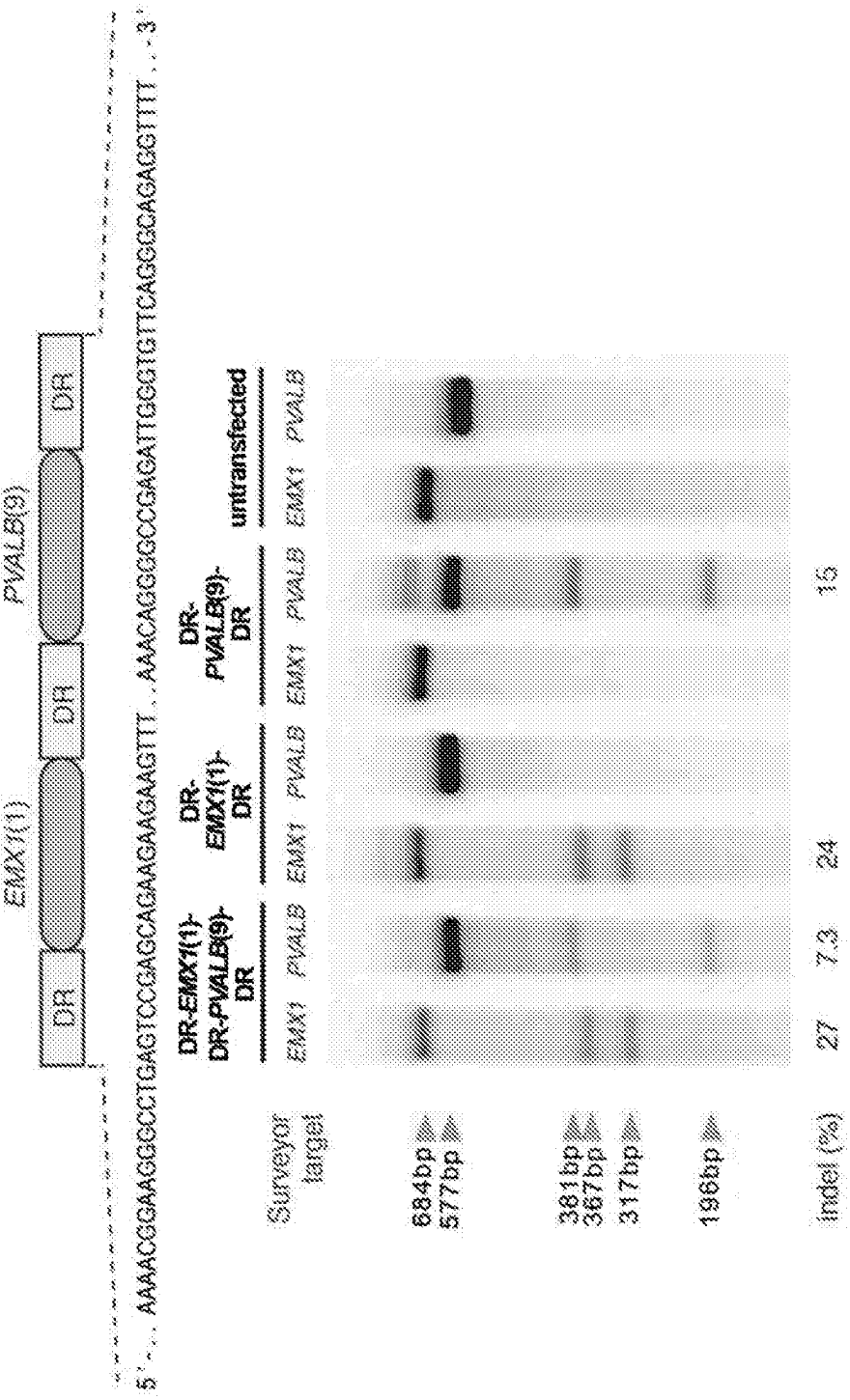

F
crRNA array design
EMX1(1)
PVALB(9)
DR
DR
DR
5' ... AAAACGGAAGGCCCTGAGTCCGAGCAGAAGAAGAAGTTT...AAACAGGGGCCGAGATTGGCTGTTCAGGGCAGAGGTTTT...3'
DR-EMX1(1)-DR-PVALB(9)-DR
DR-EMX1(1)-DR
DR-PVALB(9)-DR
untransfected
Surveyor target
EMX1 PVALB EMX1 PVALB EMX1 PVALB EMX1 PVALB
684bp
577bp
381bp
367bp
317bp
196bp
indel (%) 27 7.3 24 15

FIG. 5F

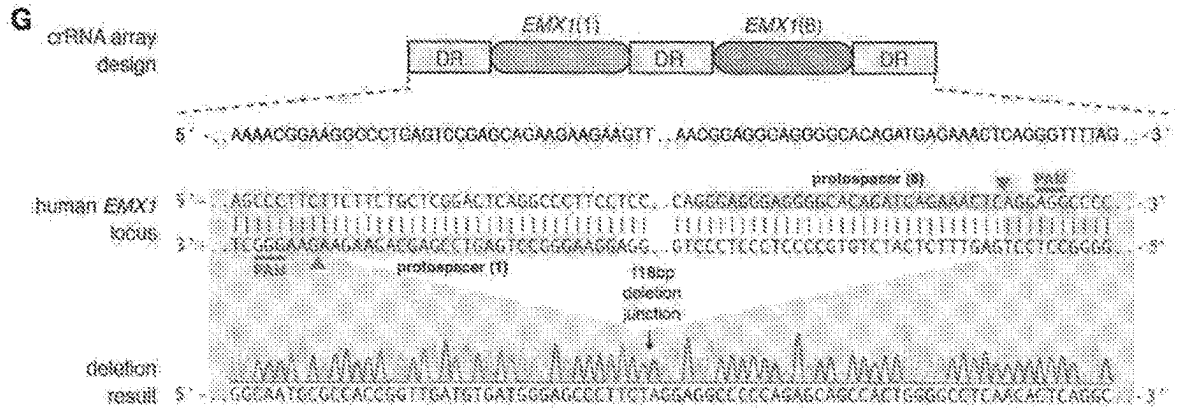
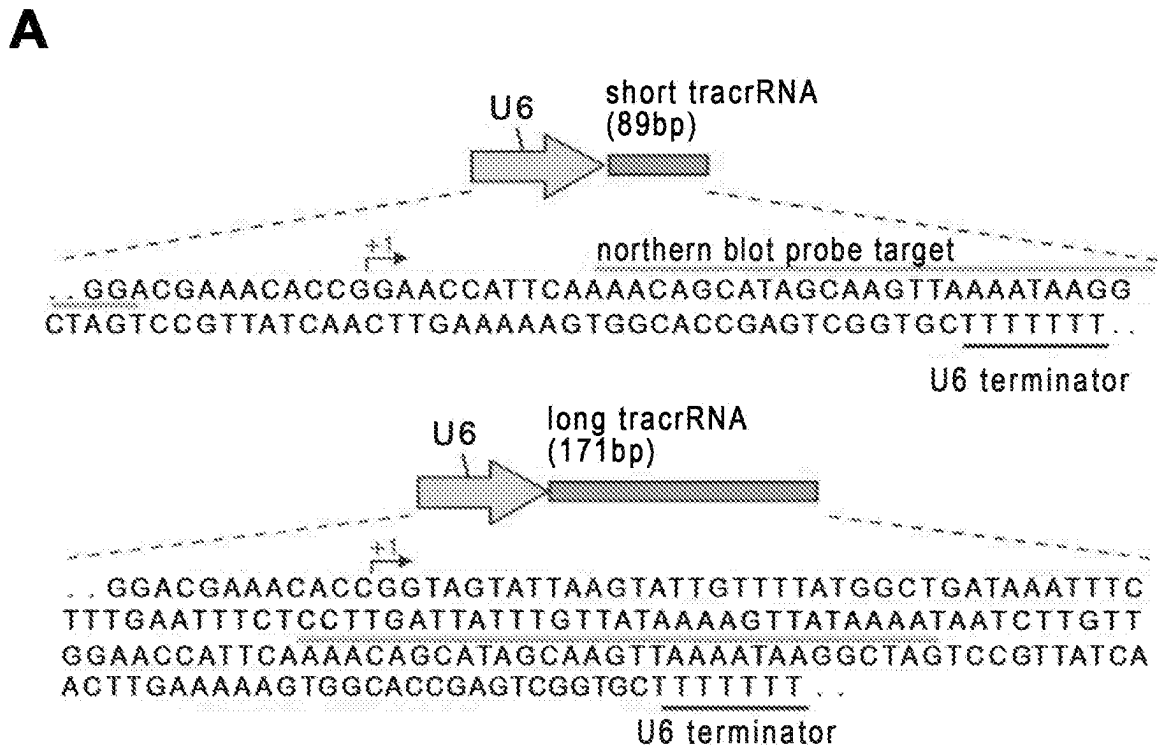

G
crRNA array design
EMX1(1)
EMX1(8)
DR
DR
DR
human EMX1 locus
protospacer (8)
PAM
protospacer (1)
1 bp deletion junction
deletion result

FIG. 5G


A
U6
short tracrRNA
(89bp)
+1
northern blot probe target
..GGACGAAACACCGGAACCATTCAAAACAGCATAGCAAGTTAAAATAAGG
CTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTTTTT..
U6 terminator
U6
long tracrRNA
(171bp)
+1
..GGACGAAACACCGGTAGTATTAAGTATTGTTTTATGGCTGATAAATTTC
TTTGAATTTCTCCTTGATTATTTGTTATAAAAGTTATAAAATAATCTTGTT
GGAACCATTCAAAACAGCATAGCAAGTTAAAATAAGGCTAGTCCGTTATCA
ACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTTTTT..
U6 terminator

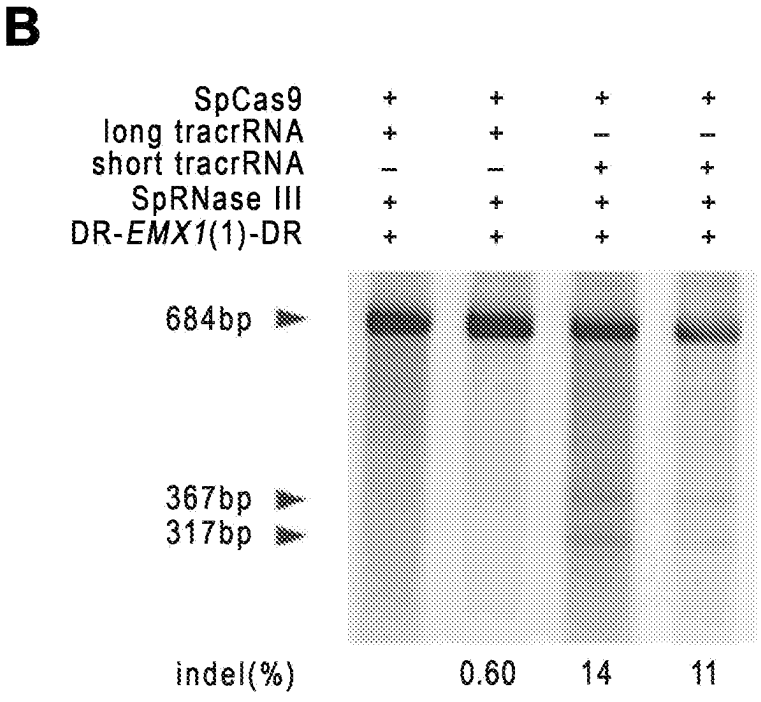

B
SpCas9 + + + +
long tracrRNA + + – –
short tracrRNA – – + +
SpRNase III + + + +
DR-EMX1(1)-DR + + + +
684bp
367bp
317bp
indel(%) 0.60 14 11

FIG. 6A-B

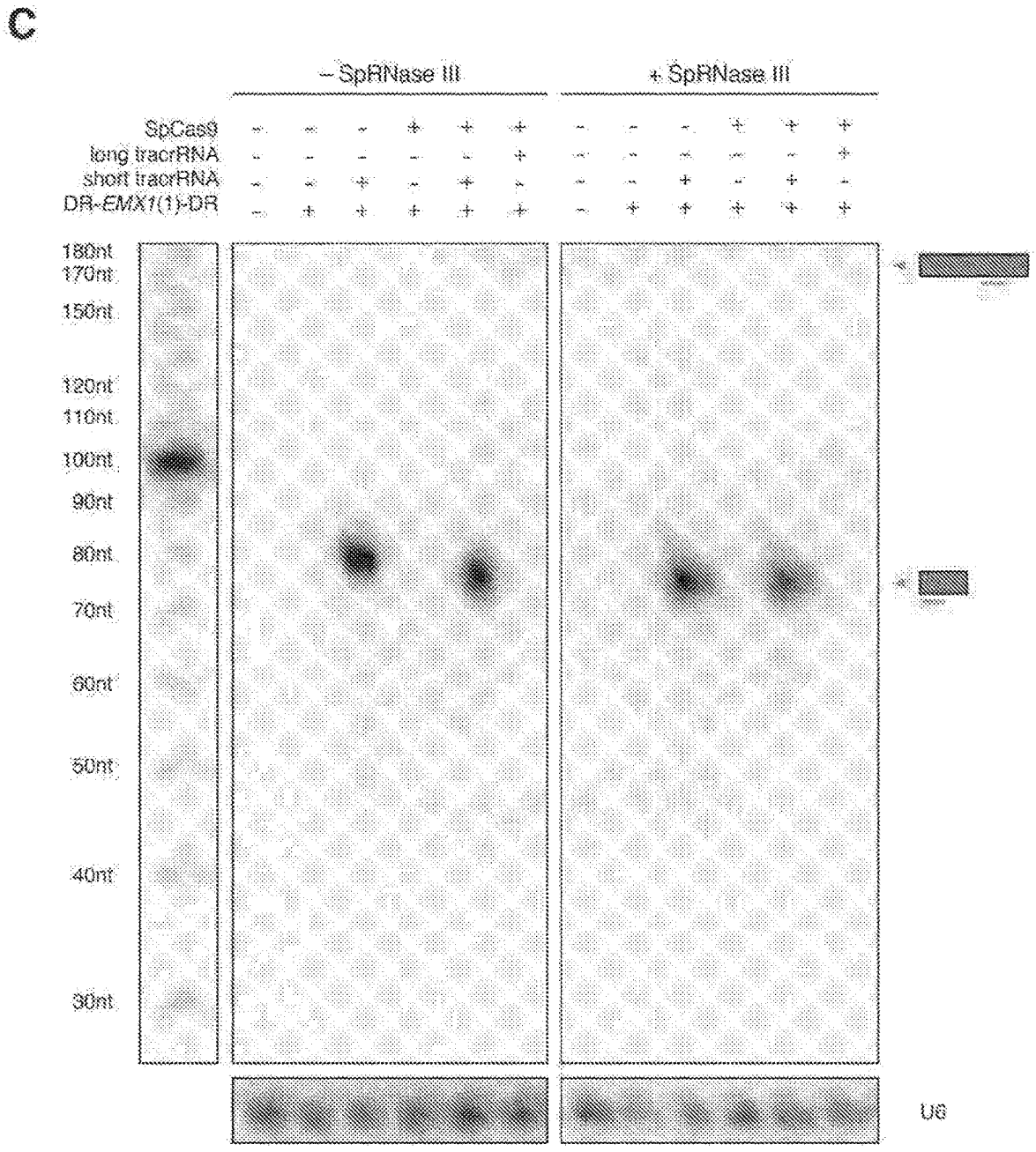

C
– SpRNase III
+ SpRNase III
SpCas9
long tracrRNA
short tracrRNA
DR-EMX1(1)-DR
180nt
170nt
150nt
120nt
110nt
100nt
90nt
80nt
70nt
60nt
50nt
40nt
30nt
U6

FIG. 6C

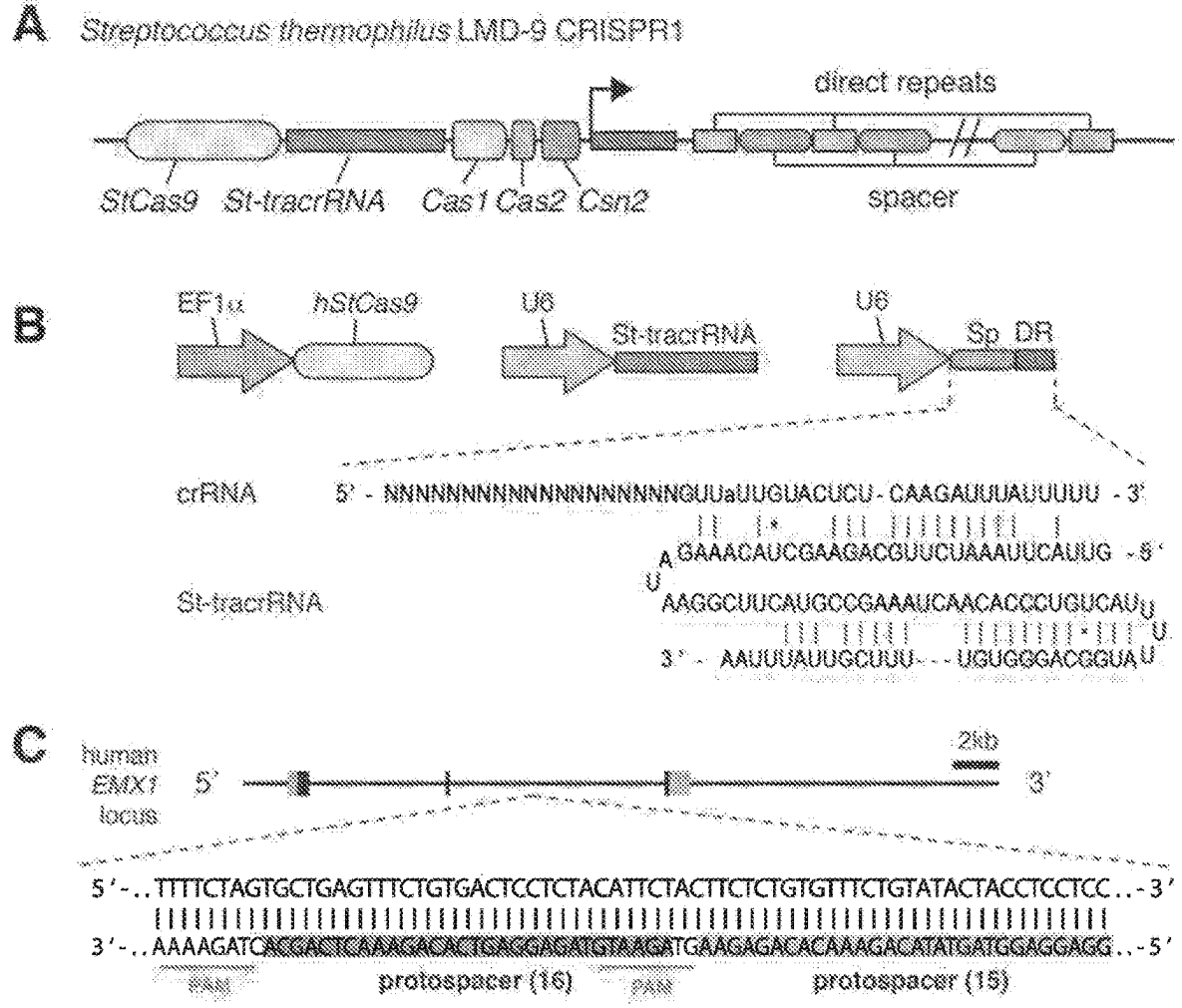

A
Streptococcus thermophilus LMD-9 CRISPR1
direct repeats
StCas9
St-tracrRNA
Cas1 Cas2 Csn2
spacer
B
EF1α
hStCas9
U6
St-tracrRNA
U6
Sp DR
crRNA
5' - NNNNNNNNNNNNNNNNNNNNGUUaUUGUACUCU - CAAGAUUUAUUUUU - 3'
A GAAACAUCGAAGACGUUCUAAAUUCAUUG - 5'
St-tracrRNA
U AAGGCUUCAUGCCGAAAUCAACACCCUGUCAU U
3' - AAUUUAUUGCUUU - - - UGUGGGACGGUA U
C
human
EMX1
locus
5'
2kb
3'
5'-..TTTTCTAGTGCTGAGTTTCTGTGACTCCTCTACATTCTACTTCTCTGTGTTTCTGTATACTACCTCCTCC..-3'
3'-..AAAAGATCACGACTCAAACACACTGACGACATGTAAGATGAAGAGACACAAAGACATATGATGGAGGACG..-5'
PAM
protospacer (16)
PAM
protospacer (15)

FIG. 7A-C

D

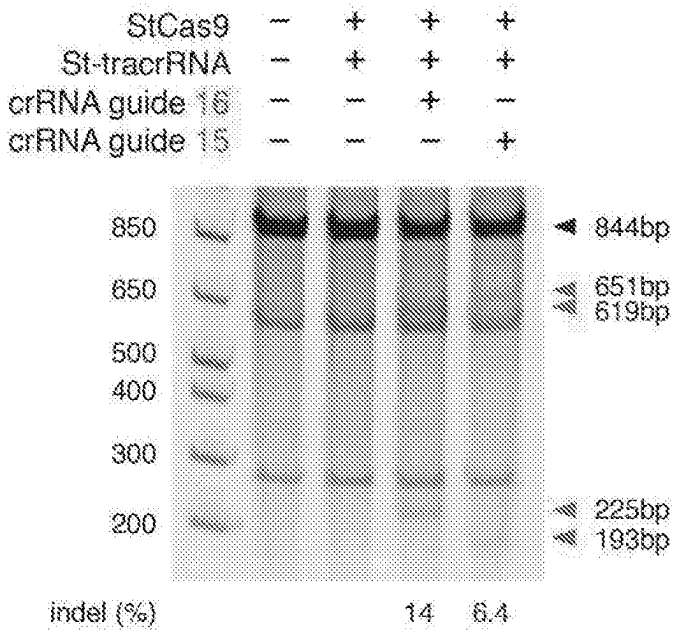

StCas9 – + + +
St-tracrRNA – + + +
crRNA guide 16 – – + –
crRNA guide 15 – – – +
850
650
500
400
300
200
844bp
651bp
619bp
225bp
193bp
indel (%)
14
6.4

FIG. 7D

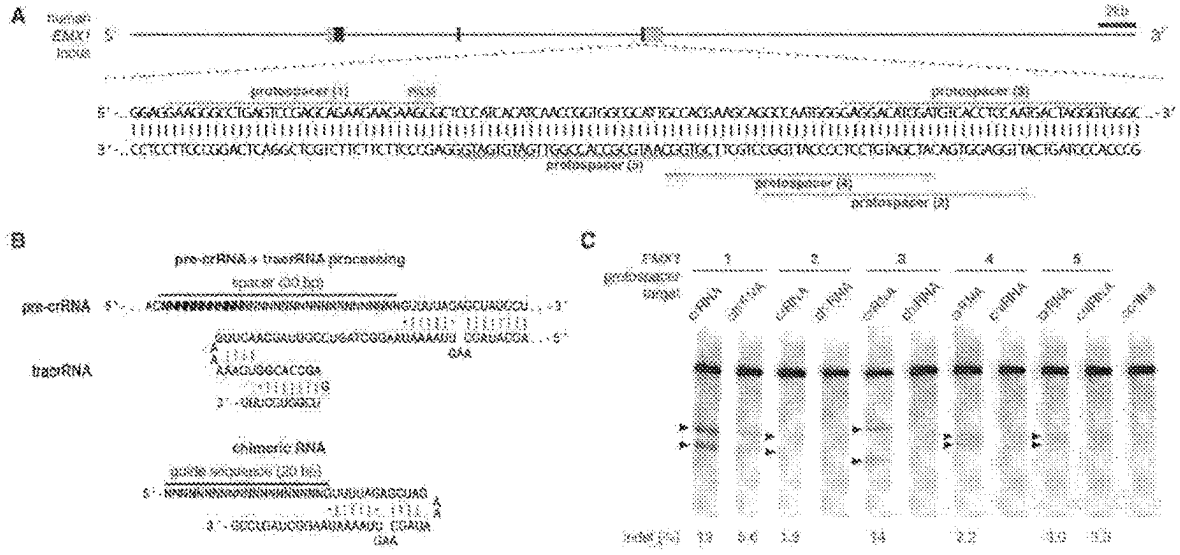

A
B
pre-crRNA + tracrRNA processing
pre-crRNA
tracrRNA
chimeric RNA
C
FANX1
1
2
3
4
5
crRNA
ctRNA
cERNA
crFRNA
crERNA
crFRNA
crFRNA
crFRNA
crFRNA
crFRNA
crRNA

FIG. 8A-C

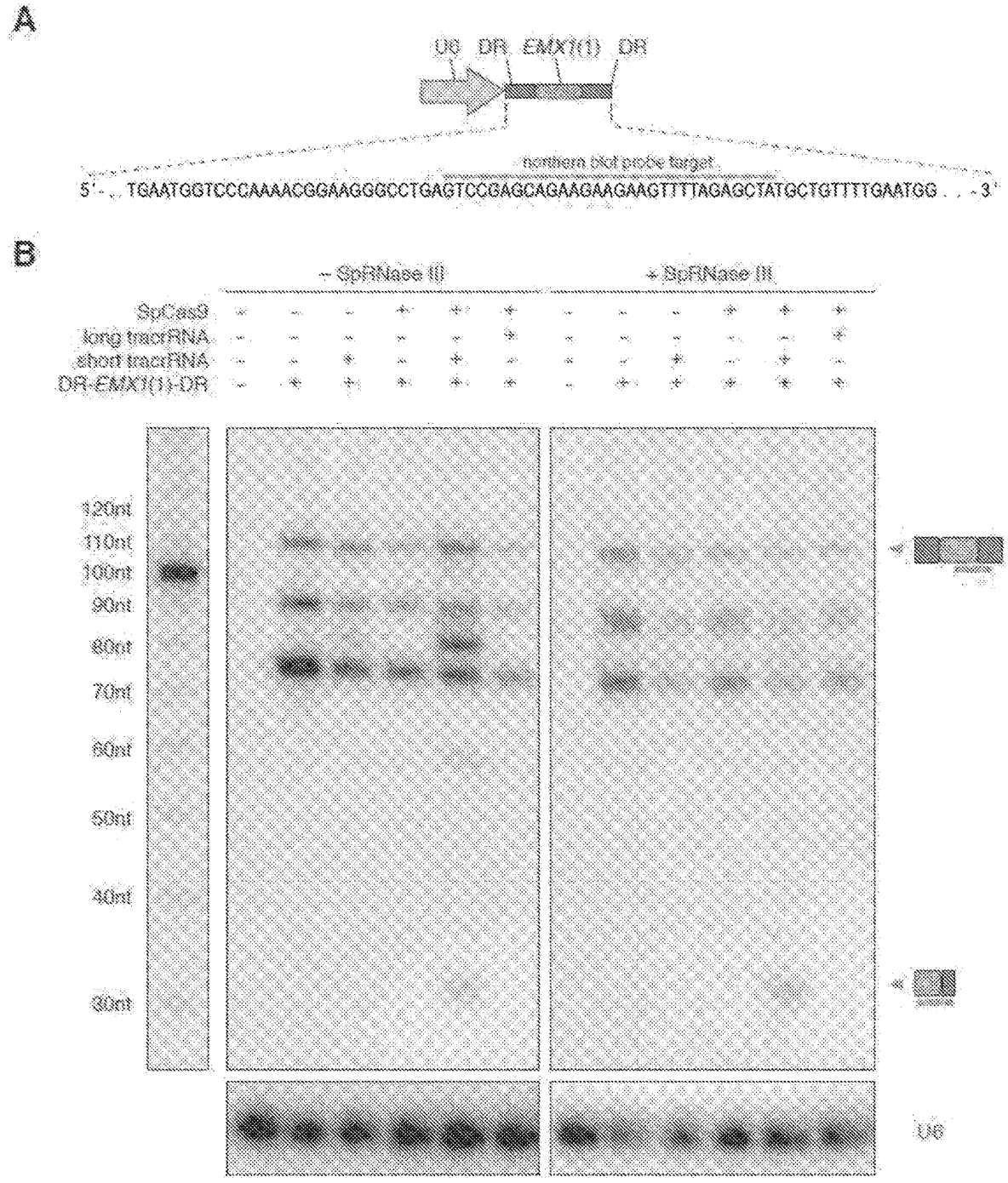

A
U6
DR
EMX1(1)
DR
northern blot probe target
5' - . . TGAATGGTCCCAAAACGGAAGGGCCTGAGTCCGAGCAGAAGAAGAAGTTTTAGAGCTATGCTGTTTTGAATGG . . - 3'
B
- SpRNase III
+ SpRNase III
SpCas9
long tracrRNA
short tracrRNA
DR-EMX1(1)-DR
120nt
110nt
100nt
90nt
80nt
70nt
60nt
50nt
40nt
30nt
U6

FIG. 9A-B

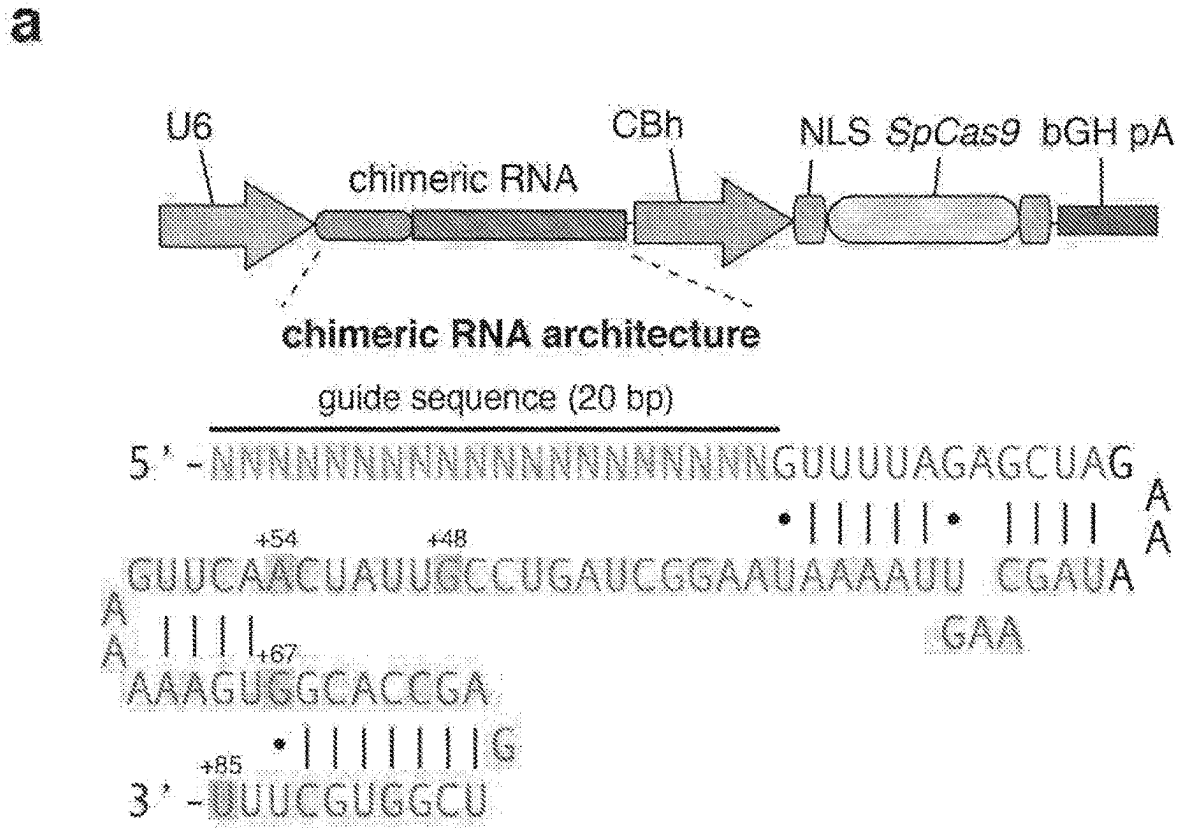

a
U6
CBh
NLS SpCas9 bGH pA
chimeric RNA
chimeric RNA architecture
guide sequence (20 bp)
5' -NNNNNNNNNNNNNNNNNNNNGUUUUAGAGCUAG
A
A
+54
+48
GUUCAACUAUUGCCUGAUCGGAAUAAAAUU CGAUA
A
A
GAA
+67
AAAGUGGCACCGA
G
+85
3' -UUUCGUGGCU

FIG. 10A

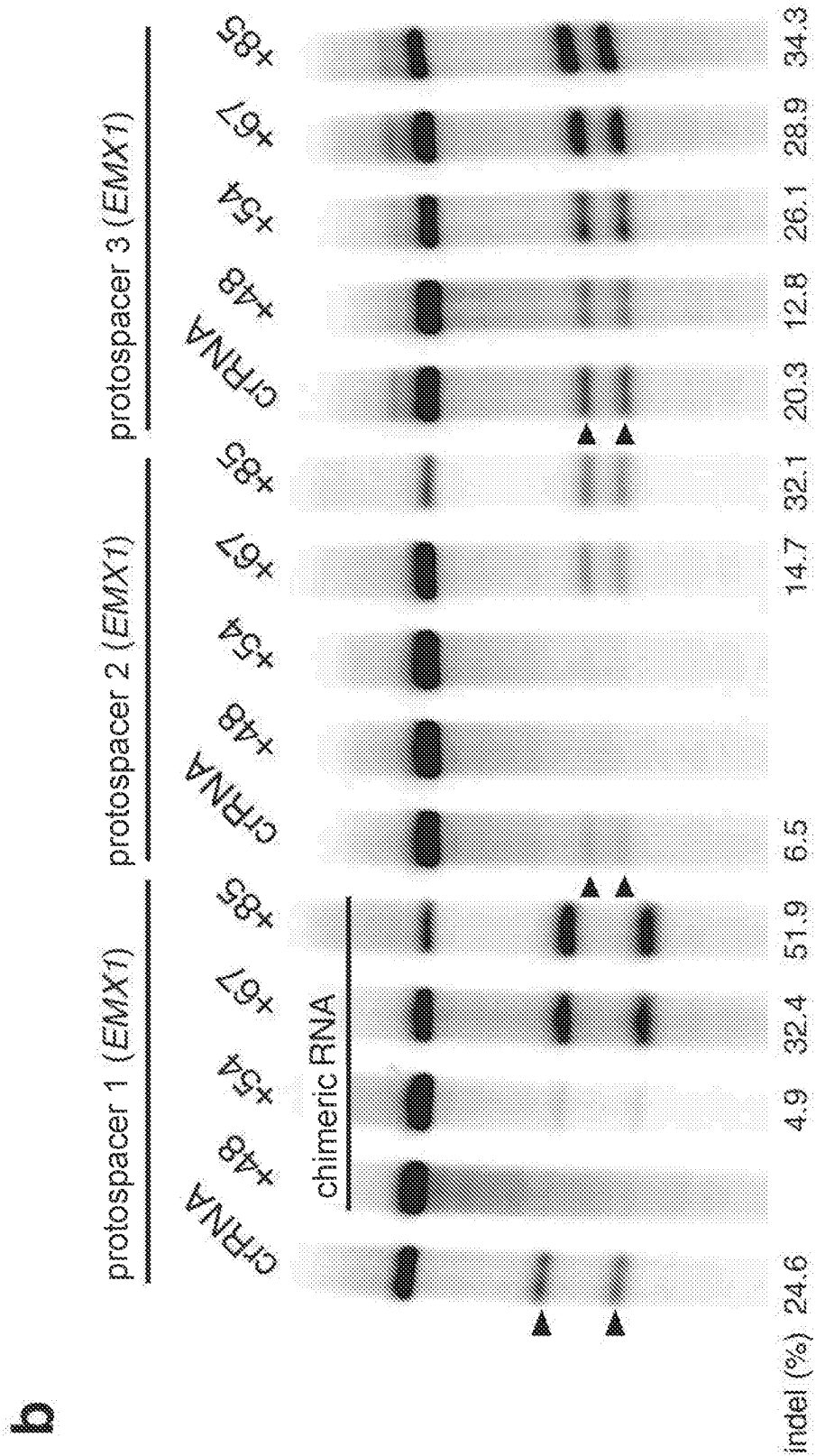

b
protospacer 1 (EMX1)
protospacer 2 (EMX1)
protospacer 3 (EMX1)
crRNA
+48
+54
+67
+85
chimeric RNA
indel (%) 24.6 4.9 32.4 51.9 6.5 14.7 32.1 20.3 12.8 26.1 28.9 34.3

FIG. 10B

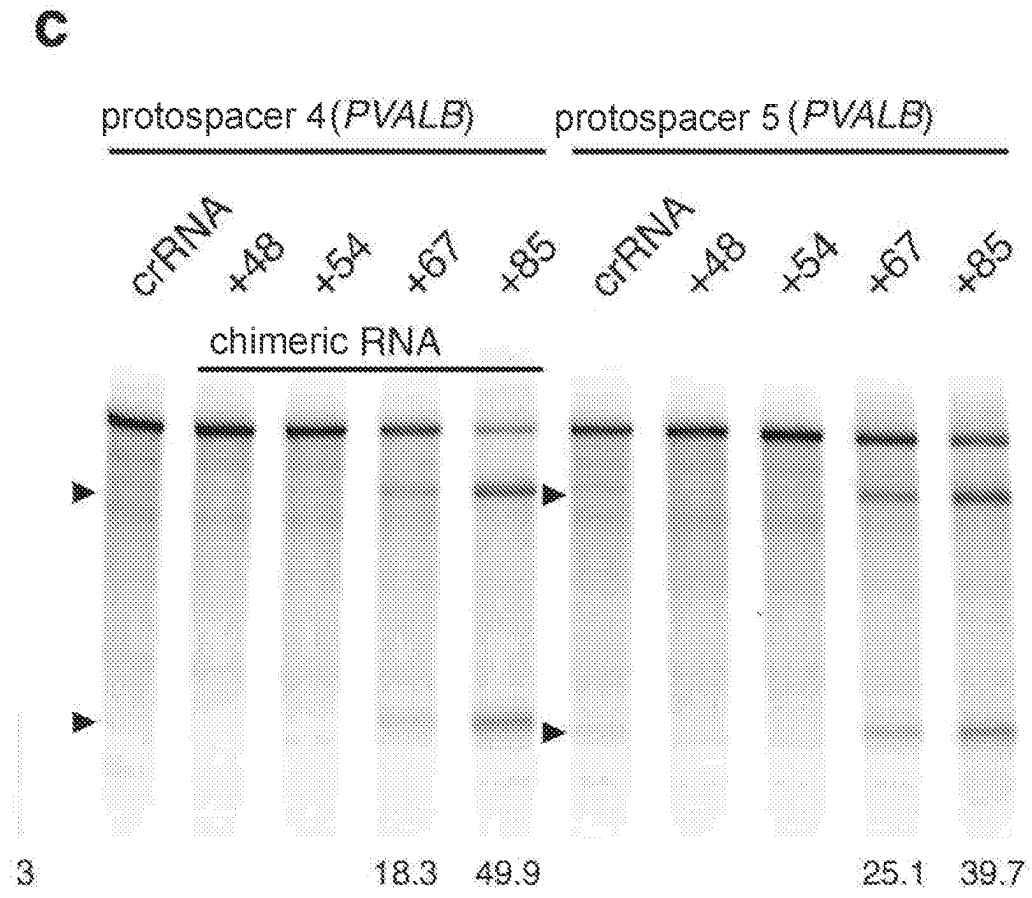

c
protospacer 4 (PVALB)
protospacer 5 (PVALB)
crRNA
+48
+54
+67
+85
crRNA
+48
+54
+67
+85
chimeric RNA
3
18.3
49.9
25.1
39.7

FIG. 10C

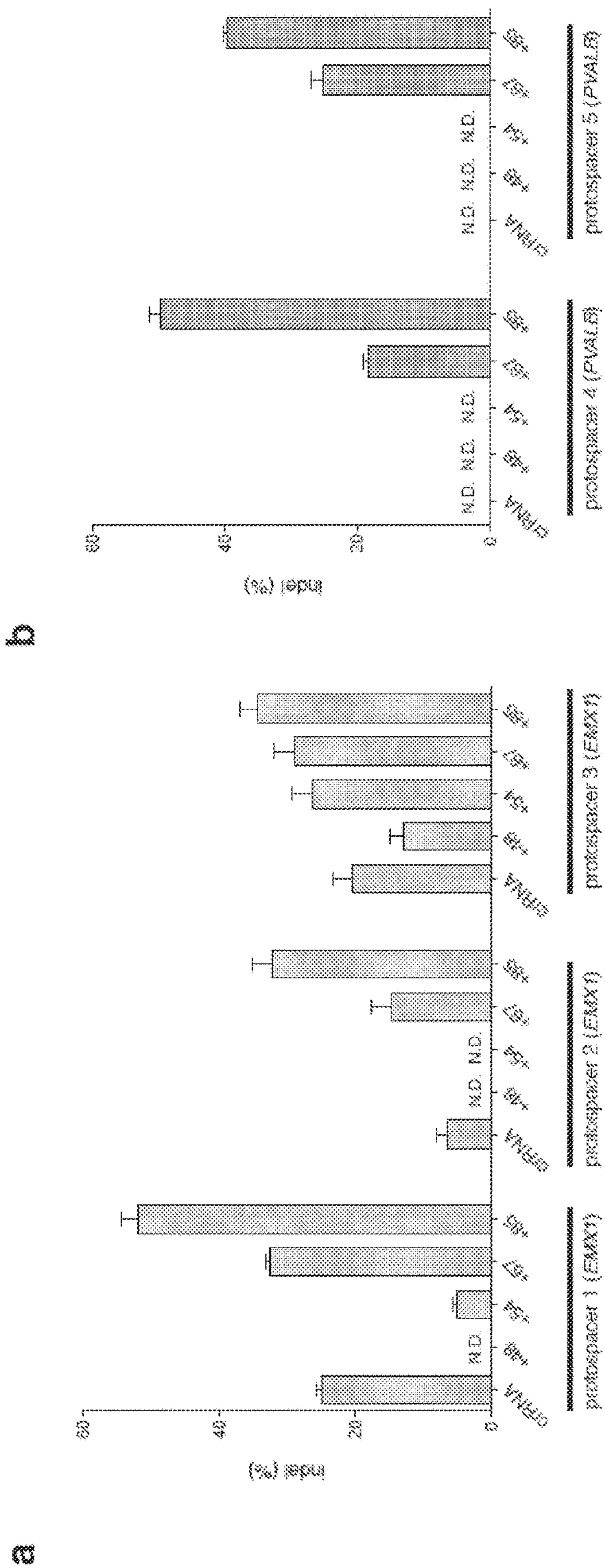

a
b
Indel (%)
60
40
20
0
crRNA
+48
+54
+57
+85
N.D.
protospacer 1 (EMX1)
protospacer 2 (EMX1)
protospacer 3 (EMX1)
protospacer 4 (PVALB)
protospacer 5 (PVALB)

FIG. 11A-B

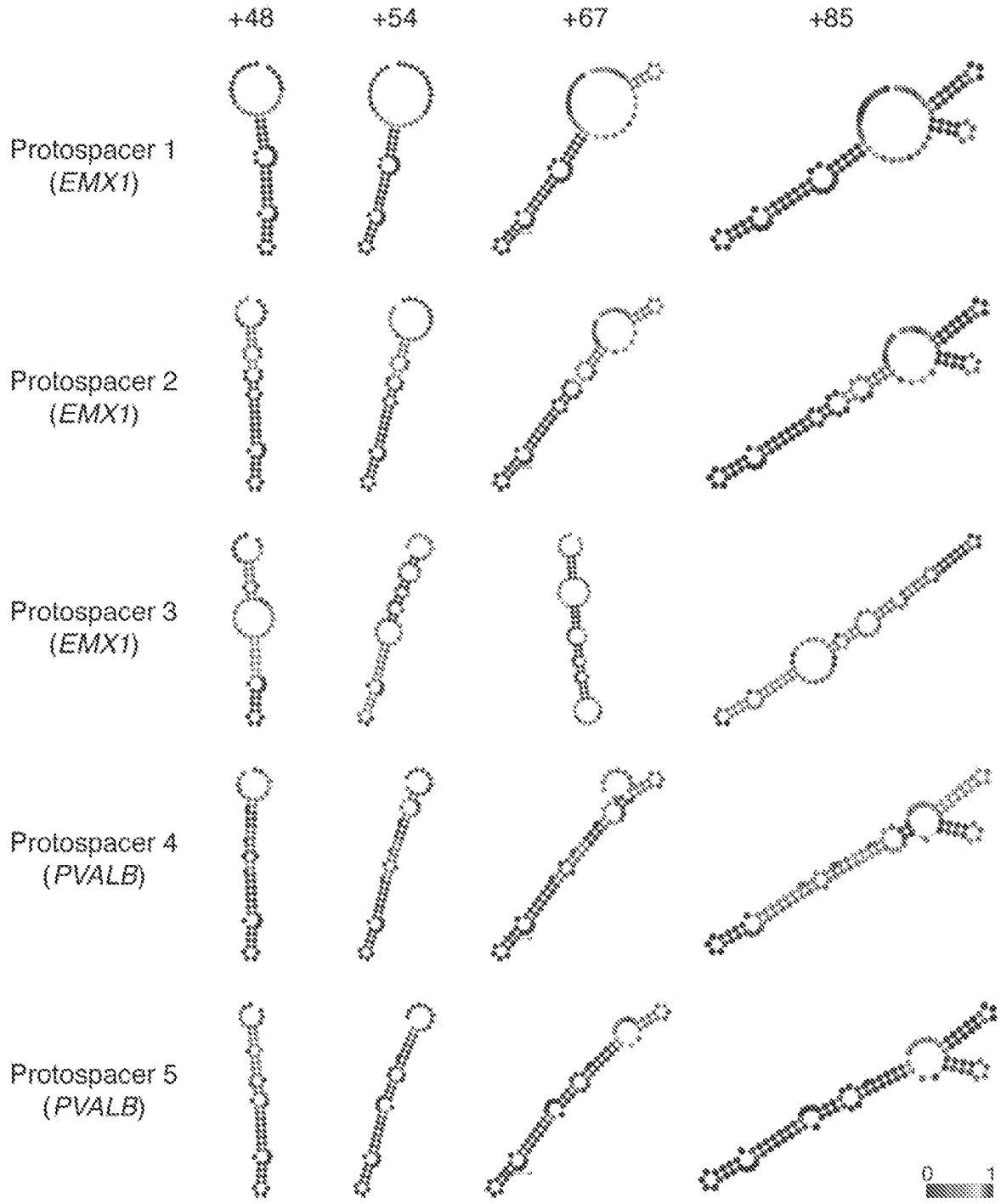

+48
+54
+67
+85
Protospacer 1 (EMX1)
Protospacer 2 (EMX1)
Protospacer 3 (EMX1)
Protospacer 4 (PVALB)
Protospacer 5 (PVALB)
0
1

FIG. 12

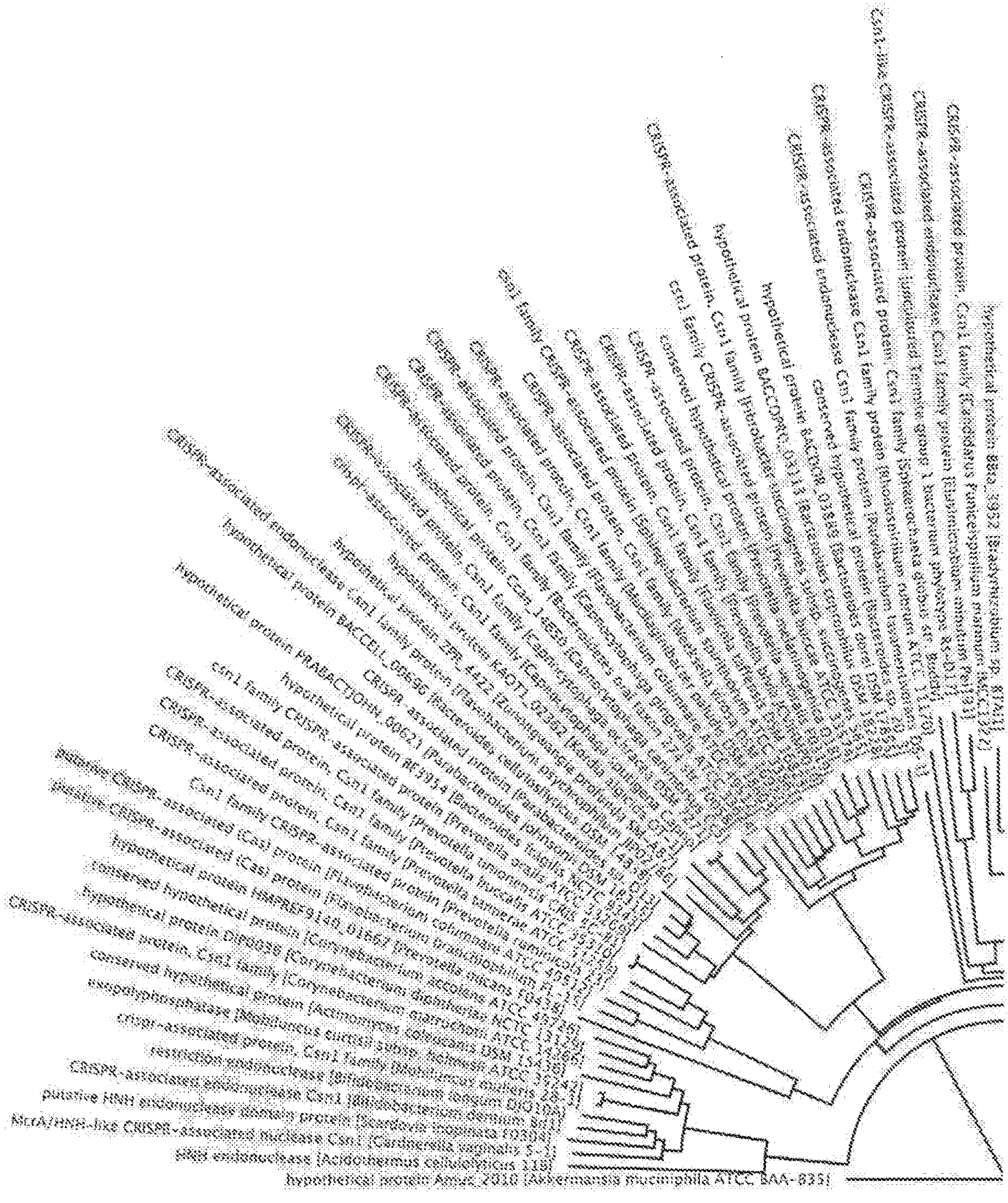


FIG. 13A

FIG. 13B

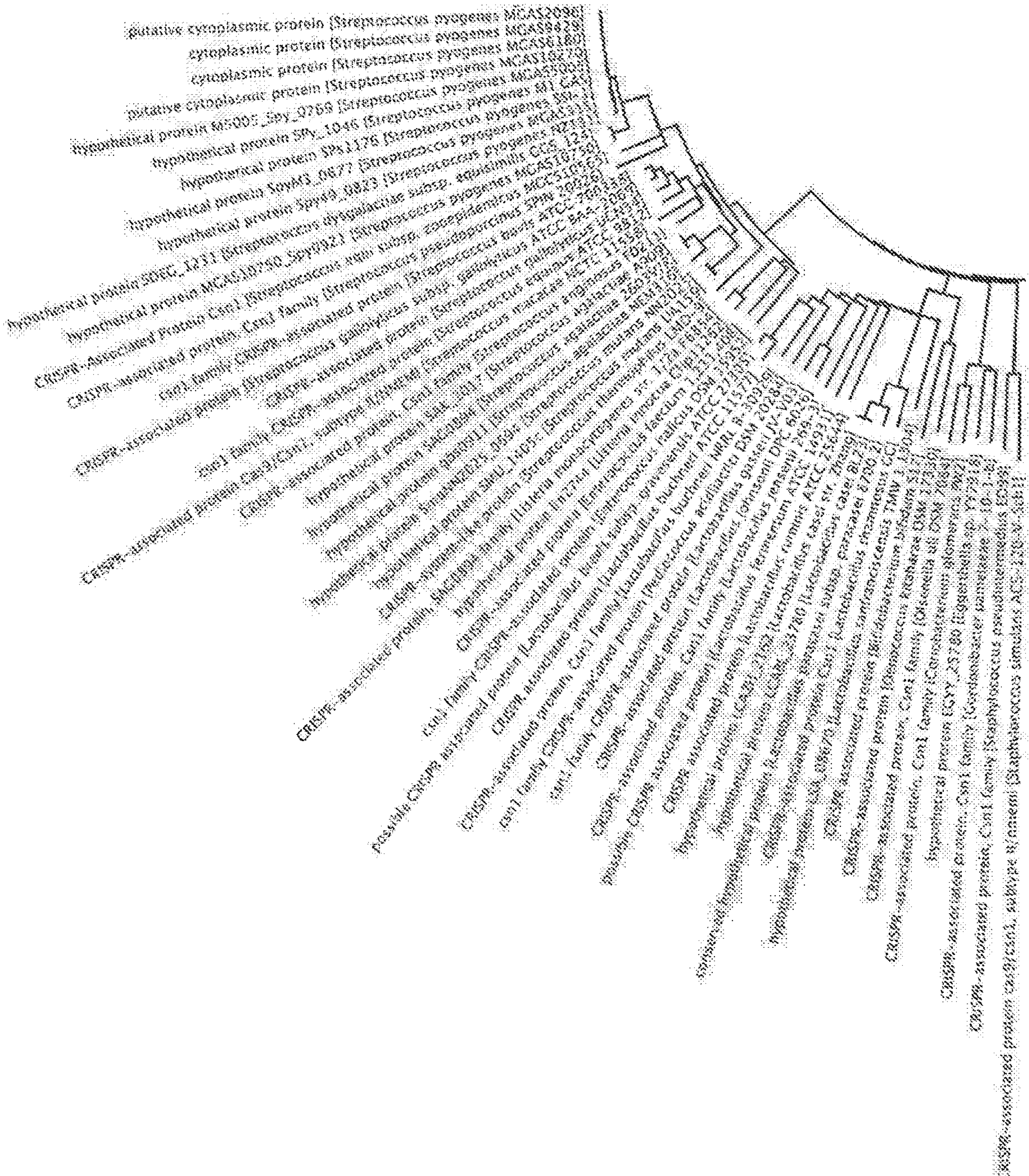

putative cytoplasmic protein [Streptococcus pyogenes MGAS2096]
cytoplasmic protein [Streptococcus pyogenes MGAS9429]
cytoplasmic protein [Streptococcus pyogenes MGAS6180]
cytoplasmic protein [Streptococcus pyogenes MGAS10270]

FIG. 13C

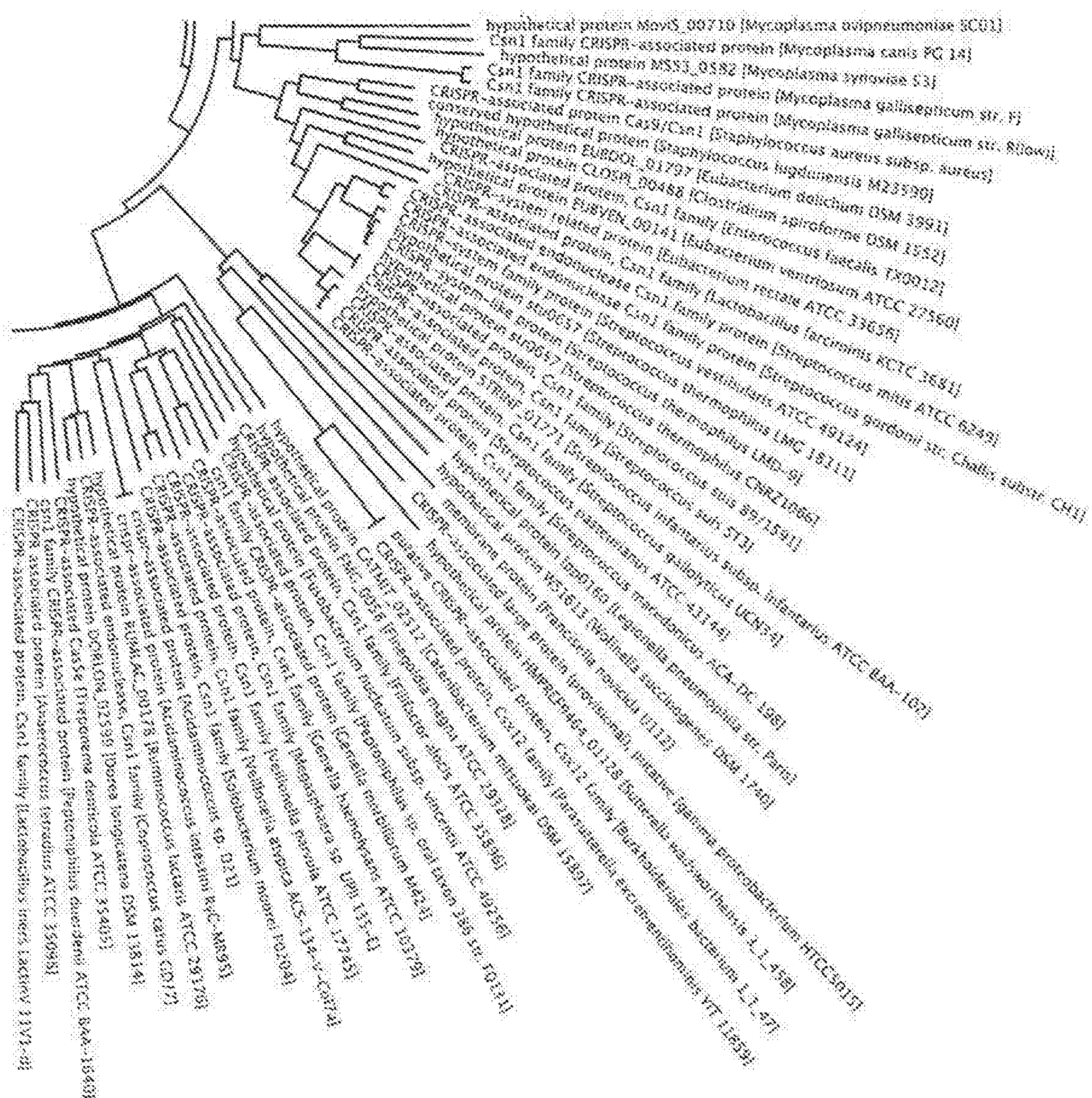

hypothetical protein MoviS_00710 [Mycoplasma ovipneumoniae SC01]
Csn1 family CRISPR-associated protein [Mycoplasma canis PG 14]
hypothetical protein MSS3_0582 [Mycoplasma synoviae 53]
Csn1 family CRISPR-associated protein [Mycoplasma gallisepticum str. F]
CRISPR-associated protein [Mycoplasma gallisepticum str. R(low)]
conserved hypothetical protein Cas9/Csn1 [Staphylococcus aureus subsp. aureus]
hypothetical protein [Staphylococcus lugdunensis M23590]
hypothetical protein EUBDOL_01797 [Eubacterium dolichum DSM 3991]
hypothetical protein CLOSPI_00468 [Clostridium spiroforme DSM 1552]
CRISPR-associated protein, Csn1 family [Enterococcus faecalis TX0012]
hypothetical protein EUBVEN_00141 [Eubacterium ventriosum ATCC 27560]
CRISPR-system related protein [Eubacterium rectale ATCC 33656]
CRISPR-associated protein, Csn1 family [Lactobacillus farciminis KCTC 3681]
CRISPR-associated endonuclease Csn1 family protein [Streptococcus mitis ATCC 6249]
CRISPR-system family protein [Streptococcus gordonii str. Challis substr. CH1]
hypothetical protein stu0657 [Streptococcus vestibularis ATCC 49124]
CRISPR-system-like protein [Streptococcus thermophilus LMG 18311]
CRISPR-associated protein str0657 [Streptococcus thermophilus LMD-9]

FIG. 13D

FIG. 14A

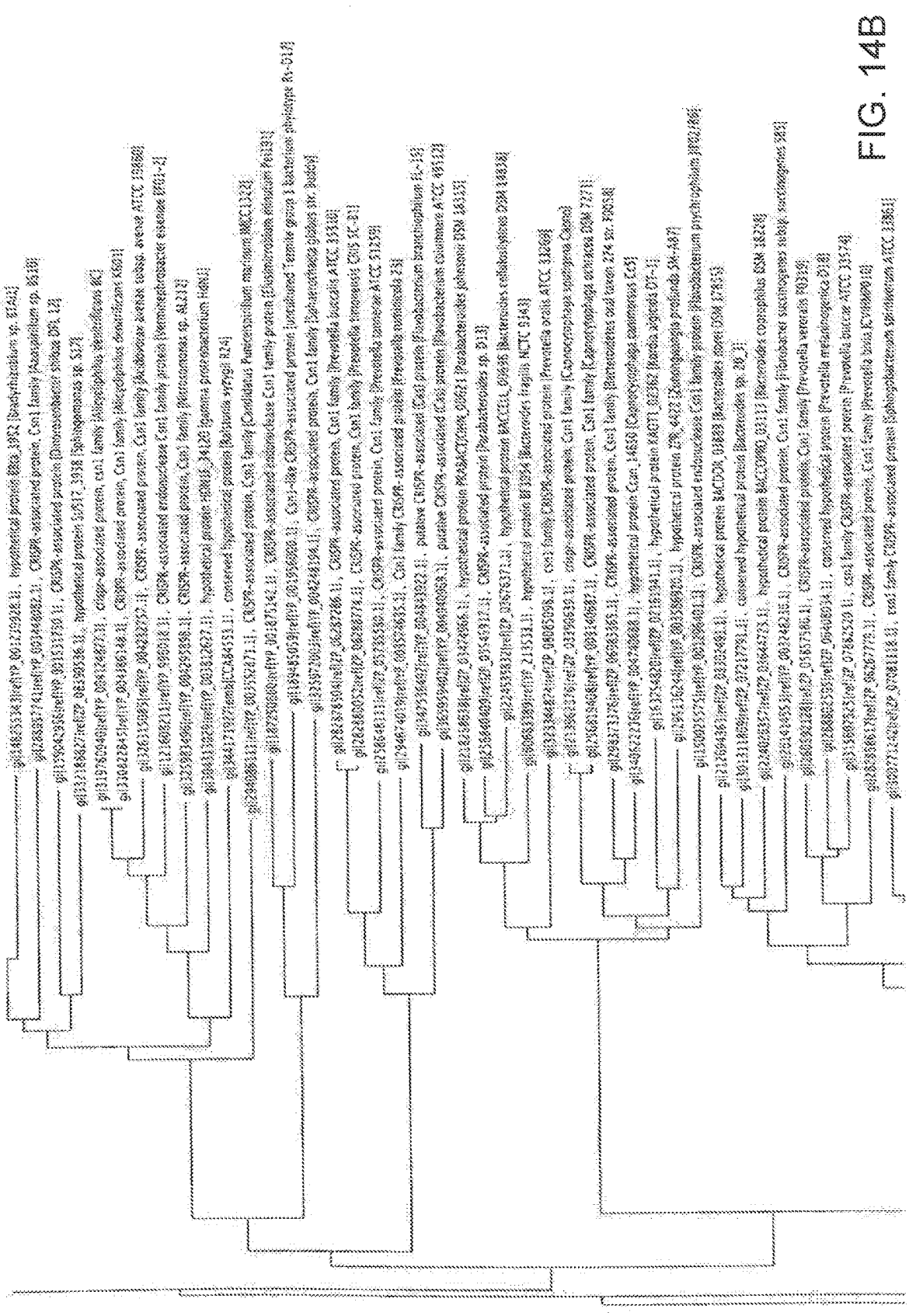


FIG. 14B

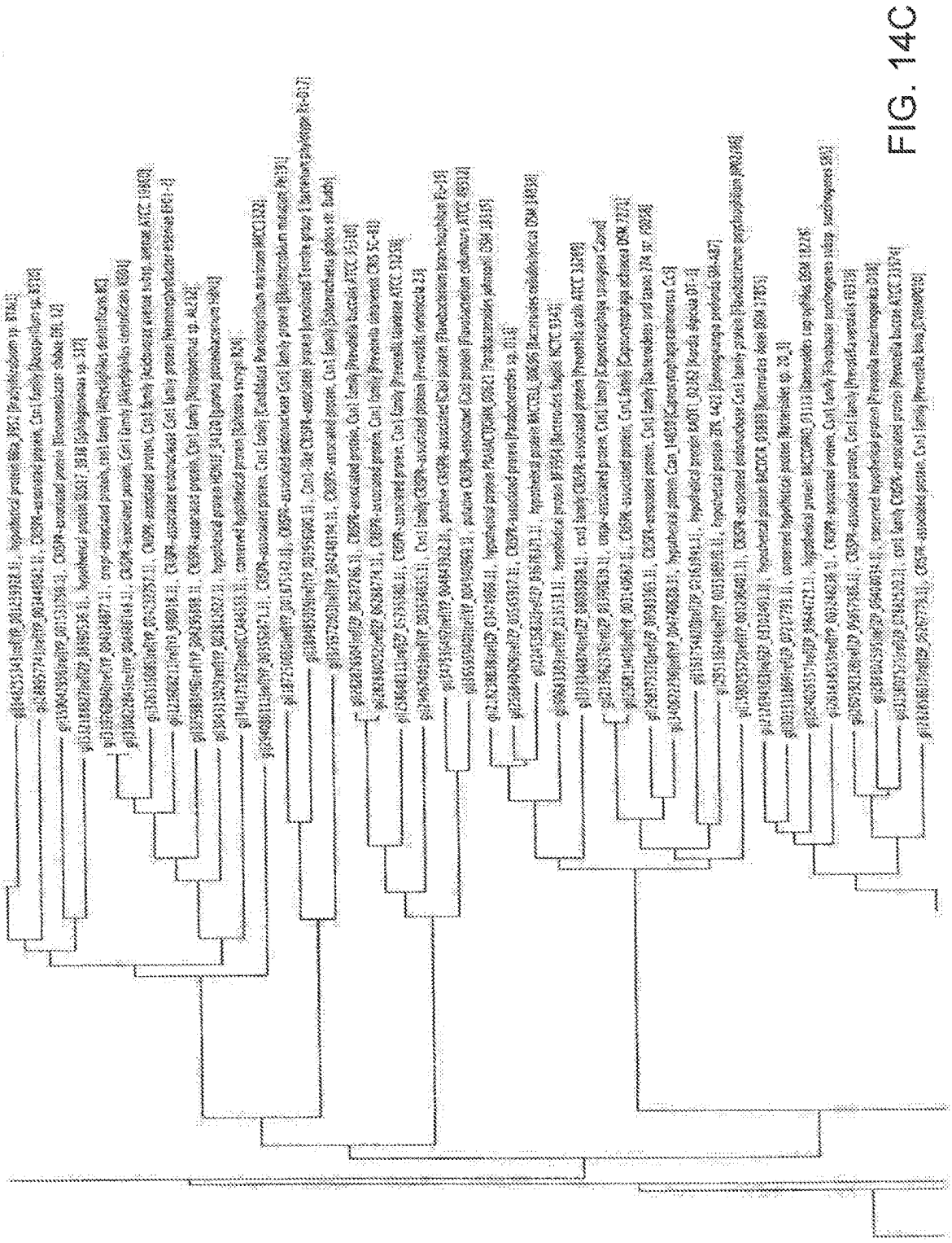


FIG. 14C

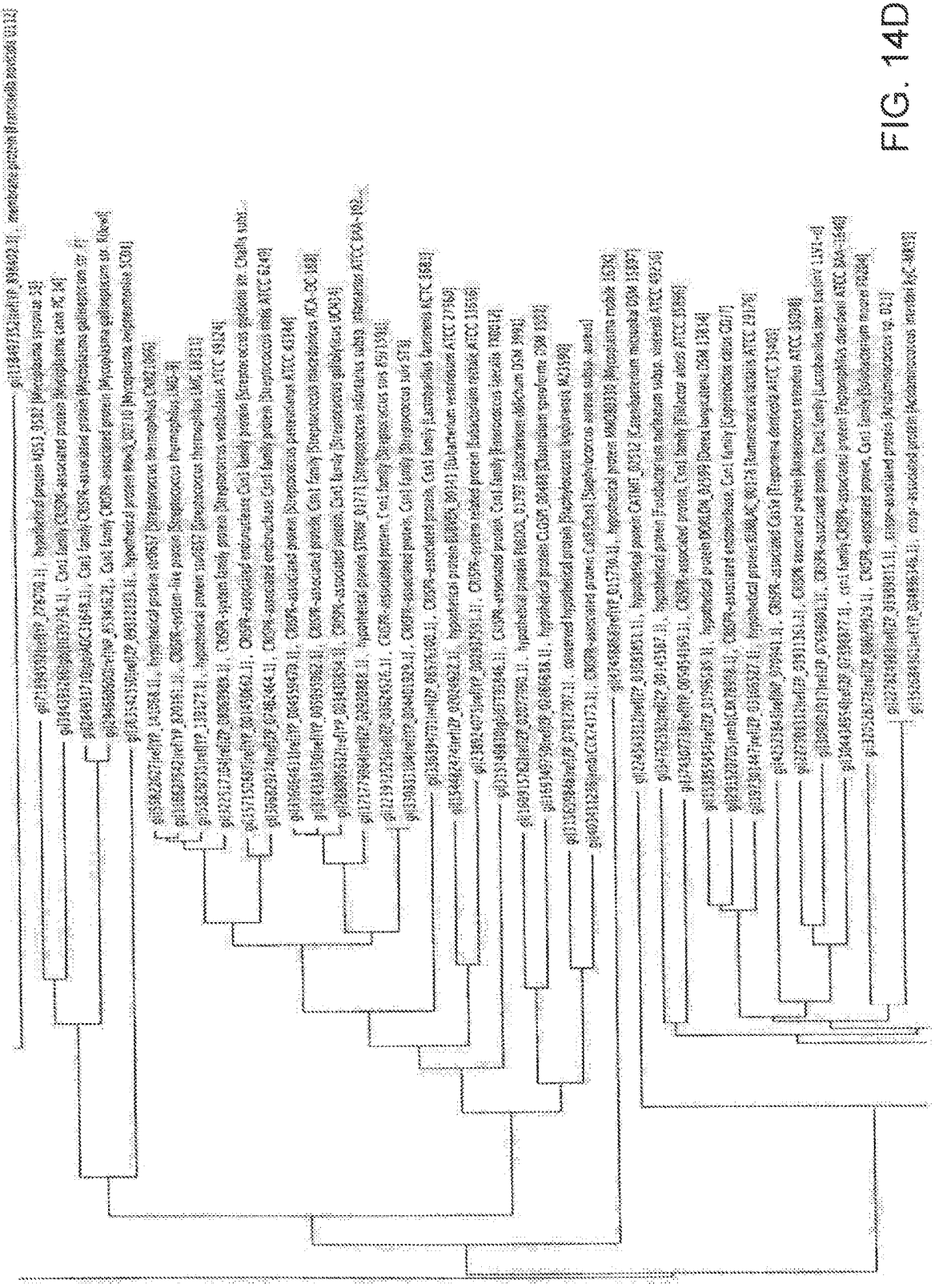


FIG. 14D

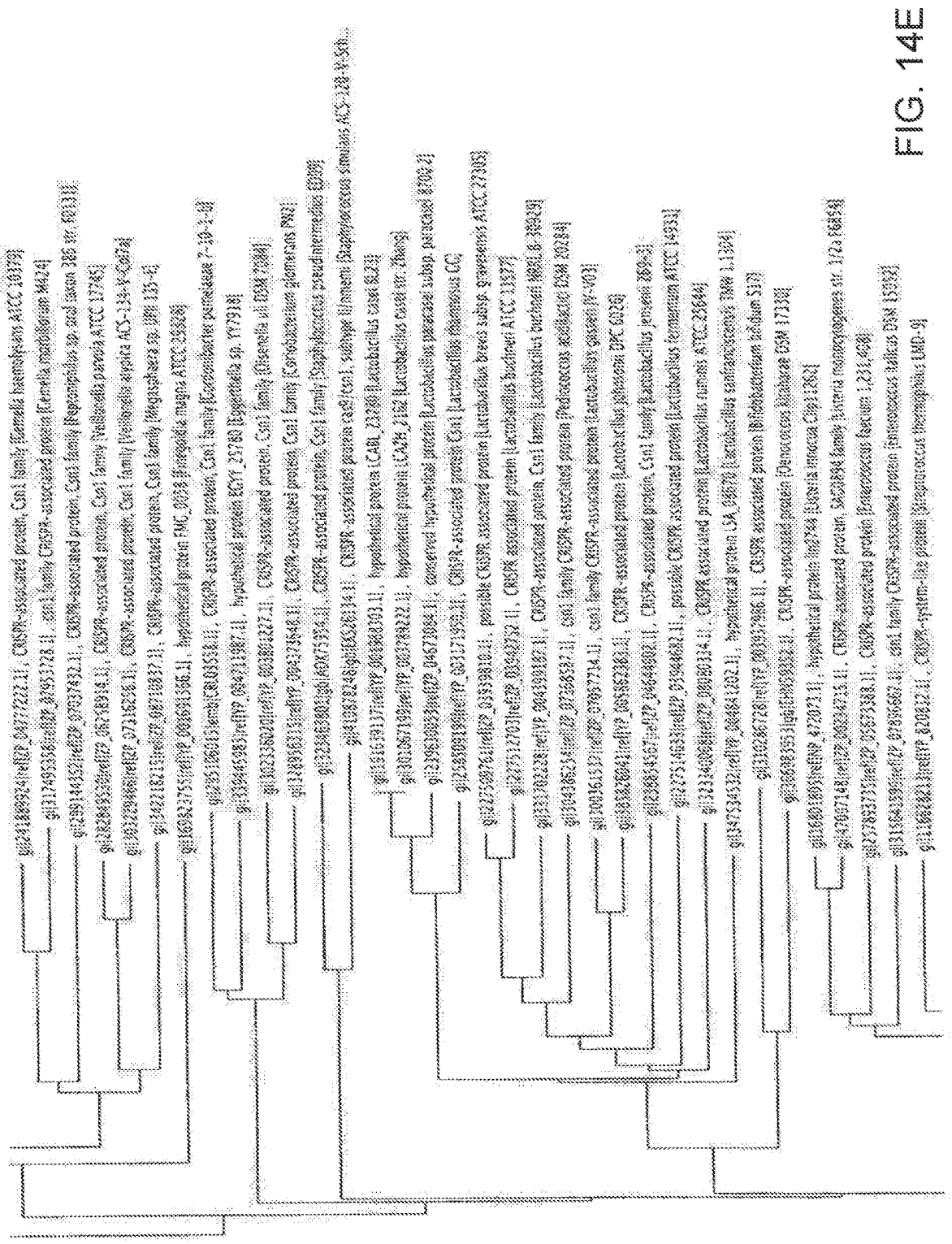


FIG. 14E

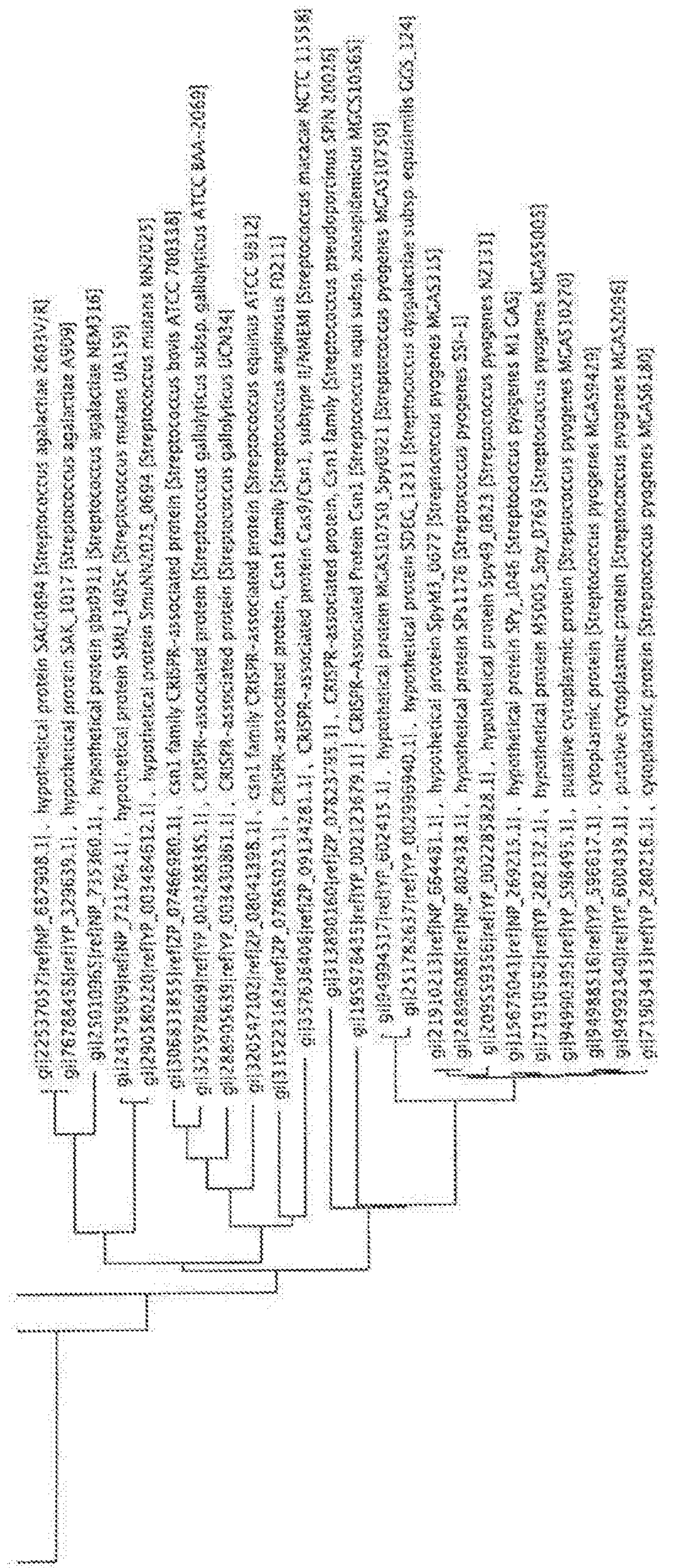


FIG. 14F

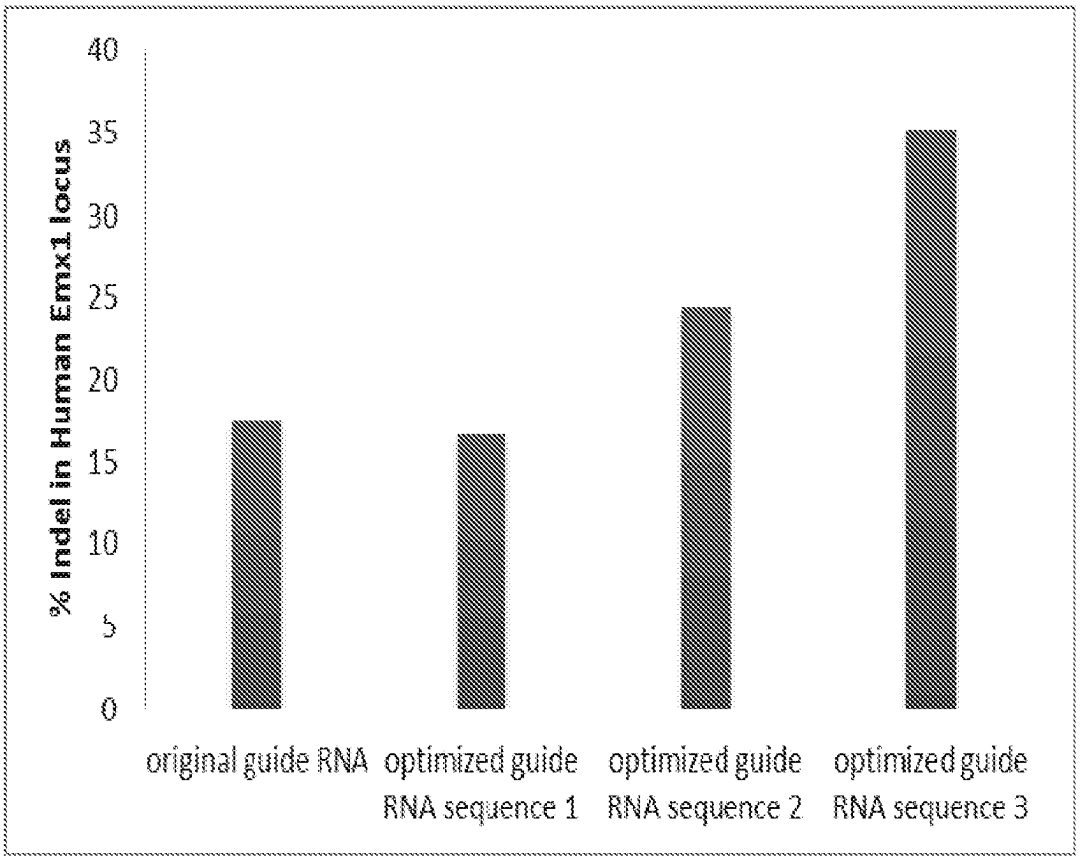
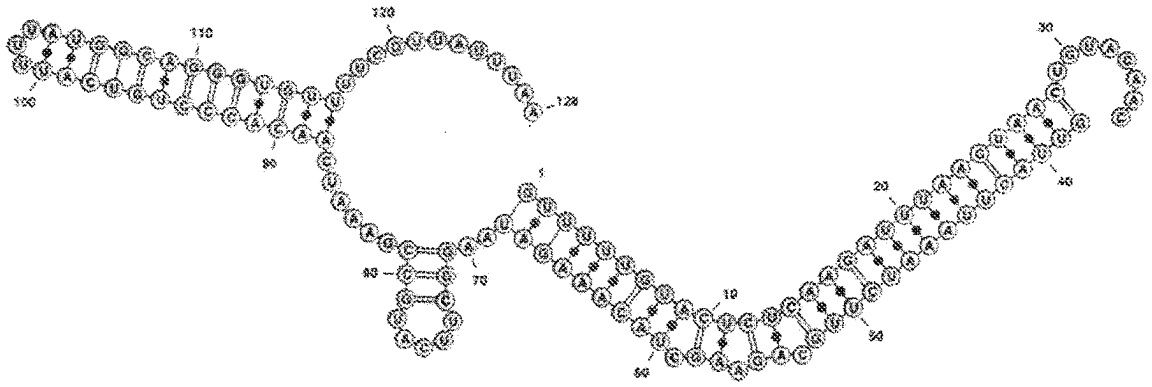

% Indel in Human Emx1 locus
40
35
30
25
20
15
10
5
0
original guide RNA
optimized guide RNA sequence 1
optimized guide RNA sequence 2
optimized guide RNA sequence 3

FIG. 15

S. thermophilus CRISPR1 RNA cofold of tracrRNA with direct repeat



```
                                                5'-NNNNNNNNNNNNNNNNNNNNgttttgta-ct-ct-caagatttaagtaactgtacaac-3'
                                                   •|•||||||  ||  ||  ||||||||||||||||
3'-aatttattgcttttgtgggacggtatttactgtcccacaactaaagccgtacttcggaatagaaacatcgaagacgttctaaattcattg-5'
```

chimeric guide RNA design

```
                                                5'-NNNNNNNNNNNNNNNNNNNNgttttgta-ct-ct-caagatttaagg 3'
                                                   •|•||||||  ||  ||  ||||||||||
3'-aatttatgcttttgtgggacggtatttactgtcccacaactaaagccgtacttcggaatagaaacatcgaagacgttctaaatca 3'

                                                5'-NNNNNNNNNNNNNNNNNNNNgttttgta-ct-ct-caagatttag 3'
                                                   •|•||||||  ||  ||  |||||||||
3'-aatttattgcttttgtgggacggtatttactgtcccacaactaaagccgtacttcggaatagaaacatcgaagacgttctaaata 3'

                                                5'-NNNNNNNNNNNNNNNNNNNNgttttgta-ct-ct-caagatg 3'
                                                   •|•||||||  ||  ||  ||||||
3'-aatttattgcttttgtgggacggtatttactgtcccacaactaaagccgtacttcggaatagaaacatcgaagacgttctaa 3'

                                                5'-NNNNNNNNNNNNNNNNNNNNgttttgta-ct-ctg 3'
                                                   •|•||||||  ||  ||
3'-aatttattgcttttgtgggacggtatttactgtcccacaactaaagccgtacttcggaatagaaacatcgaagaa 3'

                                                5'-NNNNNNNNNNNNNNNNNNNNgttttgta-ctg 3'
                                                   •|•||||||  ||
3'-aatttattgcttttgtgggacggtatttactgtcccacaactaaagccgtacttcggaatagaaacatcgaa 3'

                                                5'-NNNNNNNNNNNNNNNNNNNNgttttgta-ct-ct-caagatttag 3'
                                                   •|•||||||  ||  ||  |||||||||
                                       3'-tattttactgtcccacaactaaagccgtacttcggaatagaaacatcgaagacgttctaaata 3'

                                                5'-NNNNNNNNNNNNNNNNNNNNgttttgta-ct-ct-caagatttag 3'
                                                   •|•||||||  ||  ||  |||||||||
                                          3'-agccgtacttcggaatagaaacatcgaagacgttctaaata 3'

                                                5'-NNNNNNNNNNNNNNNNNNNNgttttgta-ct-ct-caagatttag 3'
                                                   •|•||||||  ||  ||  |||||||||
                                             3'-cttcggaatagaaacatcgaagacgttctaaata 3'

                                                5'-NNNNNNNNNNNNNNNNNNNNgttttgta-ct-ct-caagatttag 3'
                                                   •|•||||||  ||  ||  |||||||||
                                                3'-aatagaaacatcgaagacgttctaaata 3'
```

FIG. 16

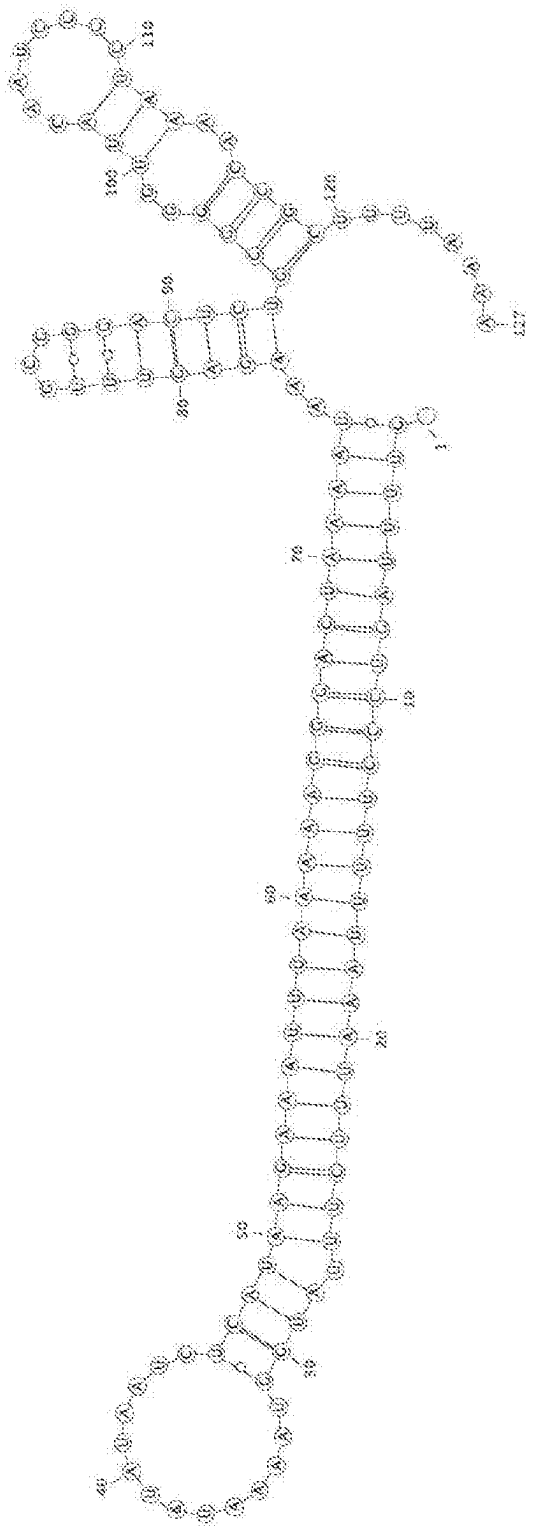
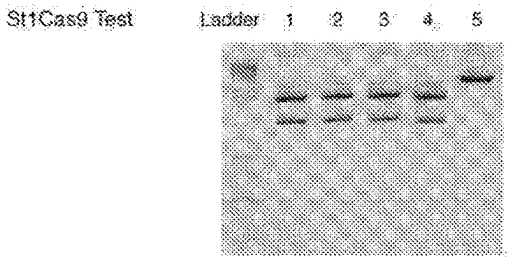


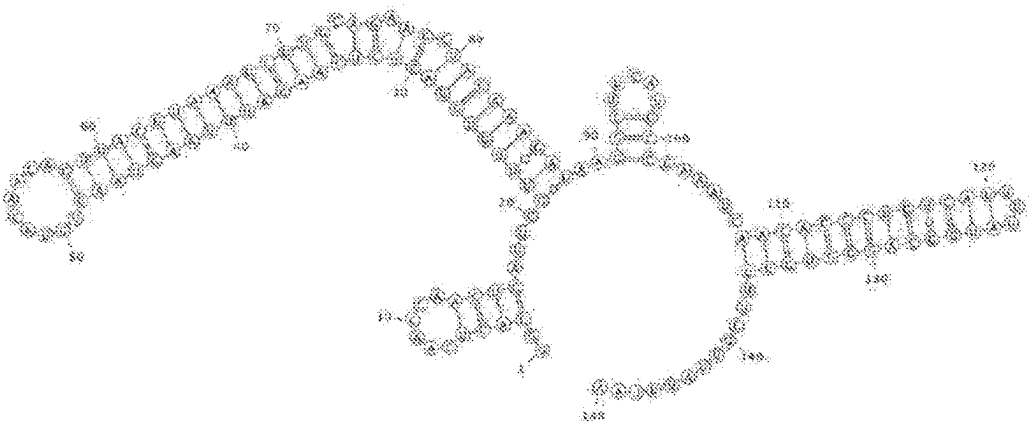
FIG. 17



Lane 1: full tracrRNA + full spacer_DR

full tracrRNA: GUUACUUAAAUCUUGCAGAAGCUACAAAGAUAAGGCUUCAUGCCGAAAUCAACACCCUGUCAUUUUAUGGCAGGGUGUUUUCGUUAUUUAA

full spacer_DR: GGGACUCAACCAAGUCAUUCGUUUUUGUACUCUCAAGAUUUAAGUAACUGUACAAC



Lane 2: mature tracrRNA + mature spacr_DR

mature tracrRNA: GUUGCAGAAGCUACAAAGAUAAGGCUUCAUGCCGAAAUCAACACCCUGUCAUUUUAUGGCAGGGUGUUUU

mature spacer_DR: GGGACUCAACCAAGUCAUUCGUUUUUGUACUCUCAAGAUUUA

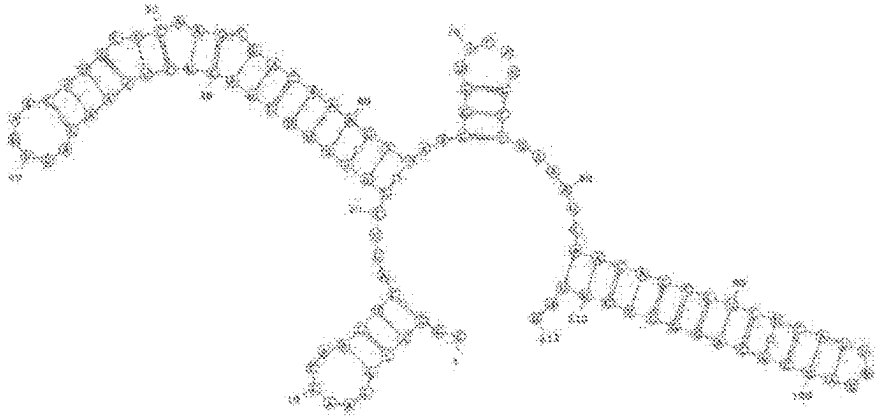
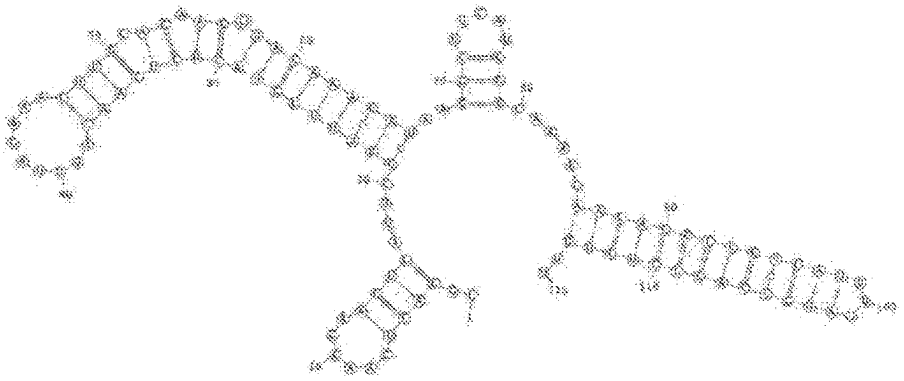


FIG 18A

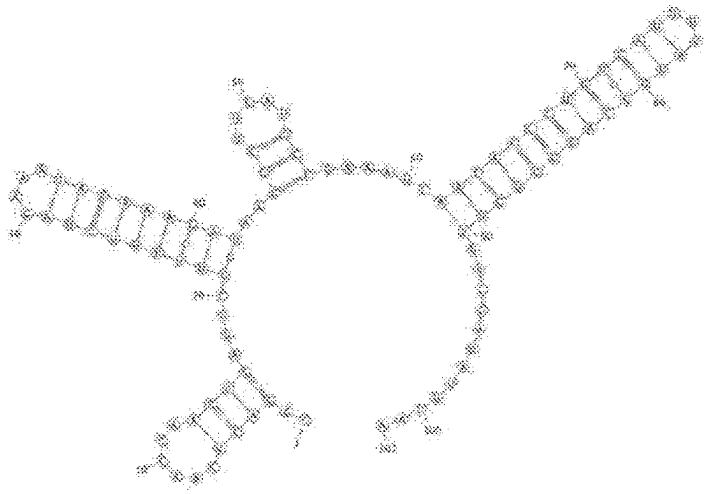
Lane 3: chimeric guideRNA design 1 (1-22 DR, 12-81 tracrRNA)

sequence: GGGACUCAACCAAGUCAUUCGUUUUUGUAGUCUCAAGAUUAGAAACUUGCAGAAGCUACAAAGAUAAGGCUUCAUGCCGAAAUCAACACCCUGUCAUUUUAUGGCAGGGUGUUUU



Lane 4: chimeric guide RNA design 2 (1-9 DR, 20-91 tracrRNA)

sequence: GGGACUCAACCAAGUCAUUCGUUUUUGUAGAAAUACAAAGAUAAGGCUUCAUGCCGAAAUCAACACCCUGUCAUUUUAUGGCAGGGUGUUUUCGUUAUUUAA



Lane 5: chimeric guide RNA design 3 (1-9 DR, 20-46 tracrRNA)

sequence: GGGACUCAACCAAGUCAUUCGUUUUUGUAGAAAUACAAAGAUAAGGCUUCAUGCCGA

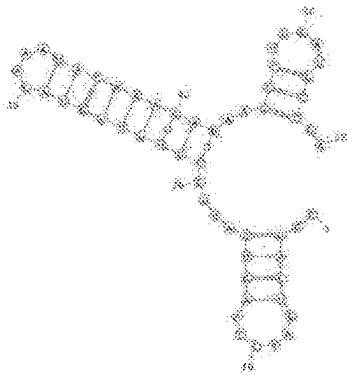


FIG. 18B

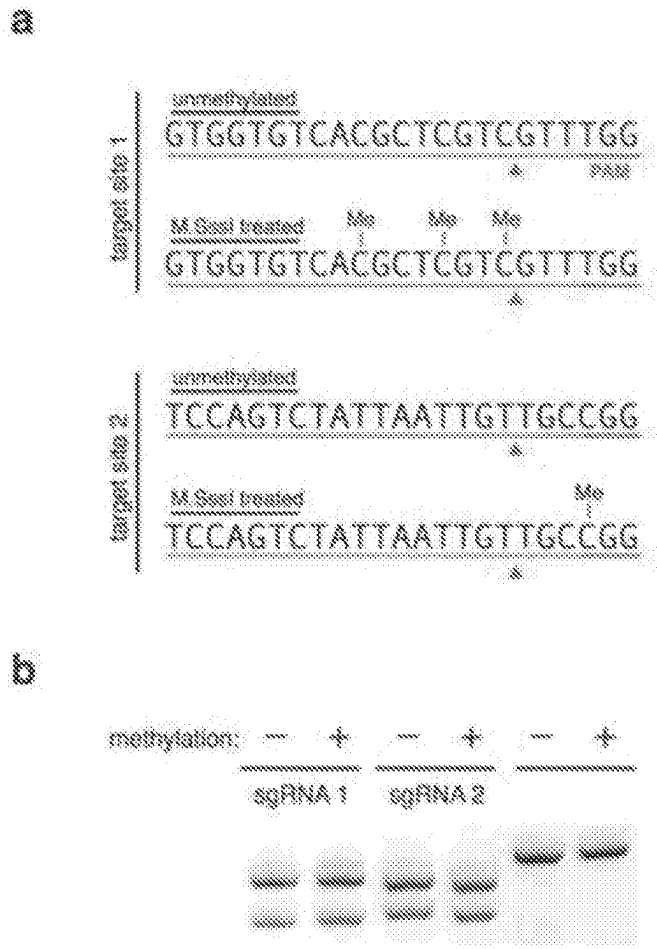

a
target site 1
unmethylated
GTGGTGTCACGCTCGTCGTTTGG
PAM
M.SssI treated
GTGGTGTCACGCTCGTCGTTTGG
target site 2
unmethylated
TCCAGTCTATTAATTGTTGCCGG
M.SssI treated
TCCAGTCTATTAATTGTTGCCGG
b
methylation: – + – + – +
sgRNA 1
sgRNA 2

FIG. 19A-B

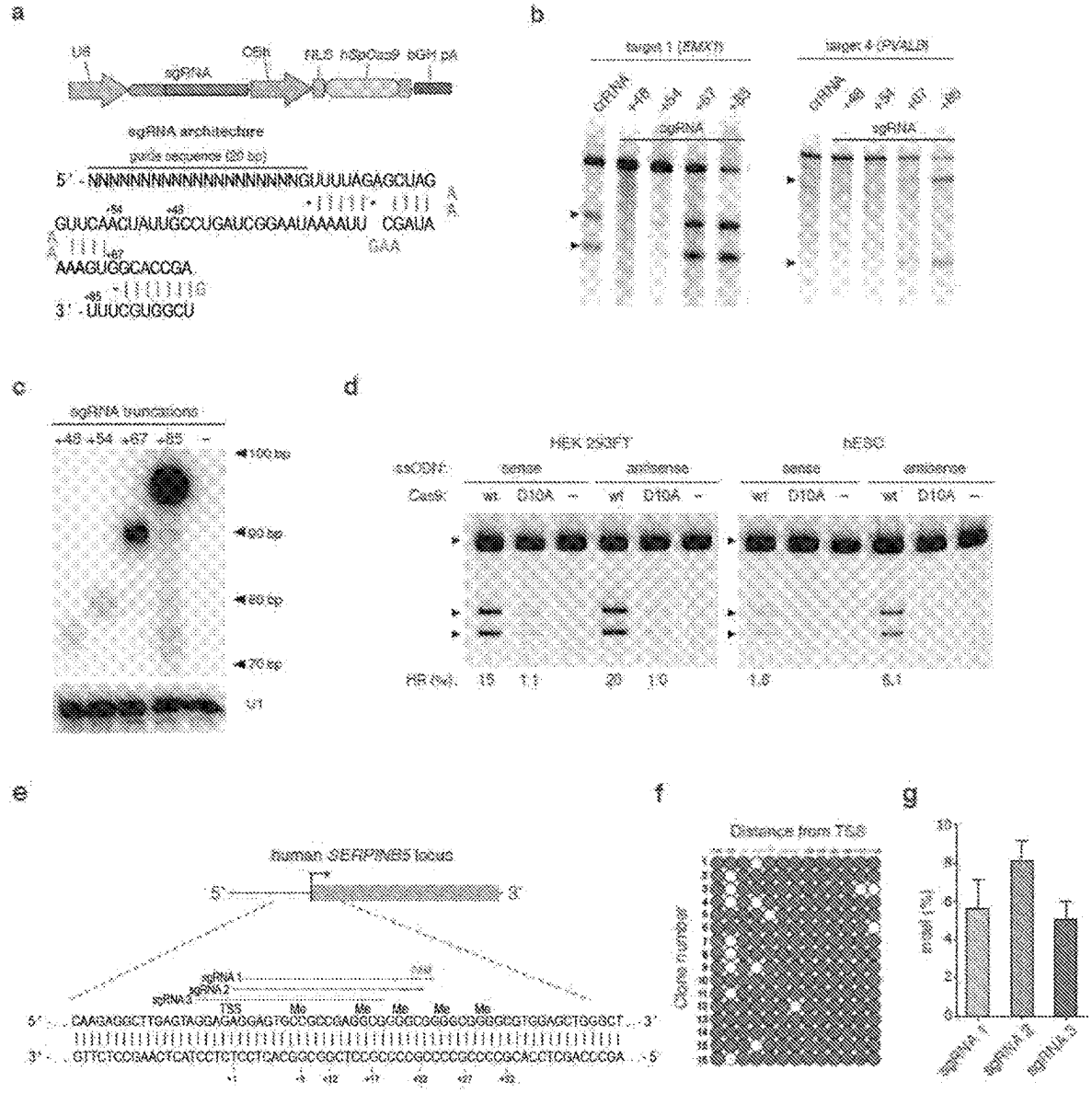

a
U6
sgRNA
CBh
nLS
nSpCas9
bGH pA
sgRNA architecture
guide sequence (20 bp)
5' - NNNNNNNNNNNNNNNNNNNNGUUUUAGAGCUAG
GUUCAACUAUUGCCUGAUCGGAAUAAAAUU CGAUA
GAA
AAAGUGGCACCGA
3' - UUUCGUGGCU
b
target 1 (RMX1)
target 4 (PVALB)
crRNA
sgRNA
c
sgRNA truncations
+48 +54 +67 +85 –
100 bp
80 bp
60 bp
40 bp
U1
d
HEK 293FT
hESC
sgRNA:
sense
antisense
Cas9:
wt
D10A
–
HR (%):
15
1.1
20
1.0
1.6
8.1
e
human SERPINB5 locus
5'
3'
sgRNA 1
sgRNA 2
sgRNA 3
TSS
Me
5' . . . CAAGAGGCTTGAGTAGGAGAGGAGTGCCGCCGAGGCGCGCGCGGGCGCGGCGTGGAGCTGGGCT . . . 3'
3' . . . GTTCTCCGAACTCATCCTCTCCTCACGGCGGCTCCGCGCGCCCGCGCGCGCACCTCGACCCGA . . . 5'
f
Distance from TSS
Cluster number
g
indel (%)
sgRNA 1
sgRNA 2
sgRNA 3

FIG. 20A-G

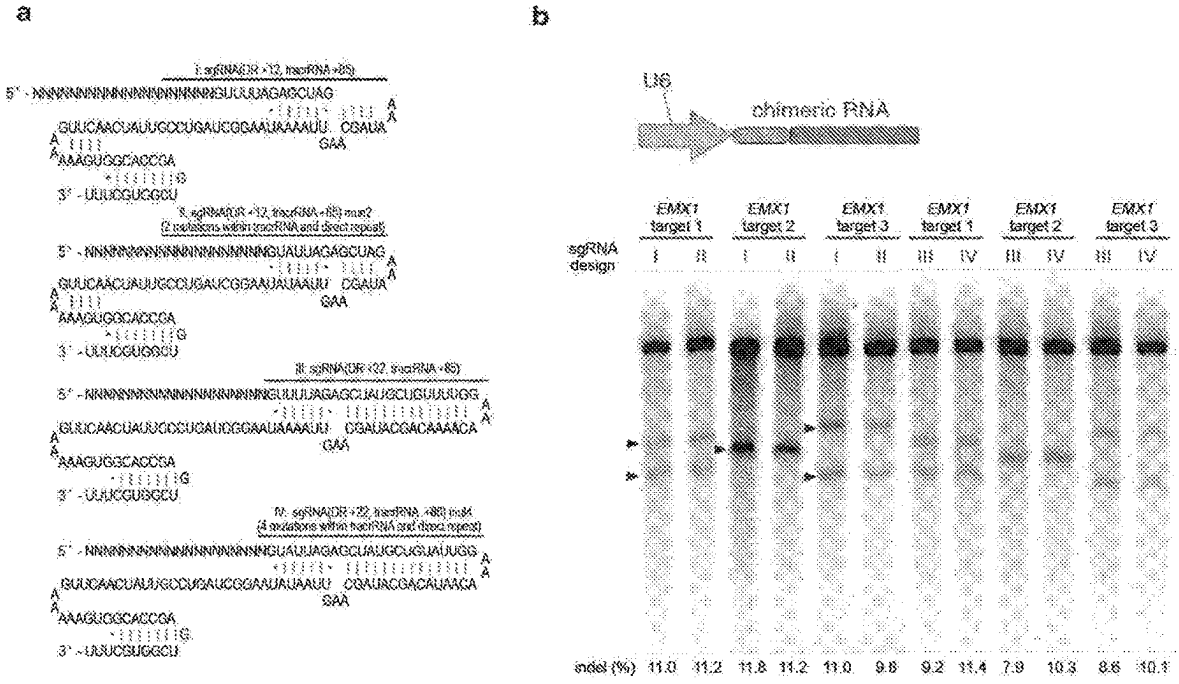
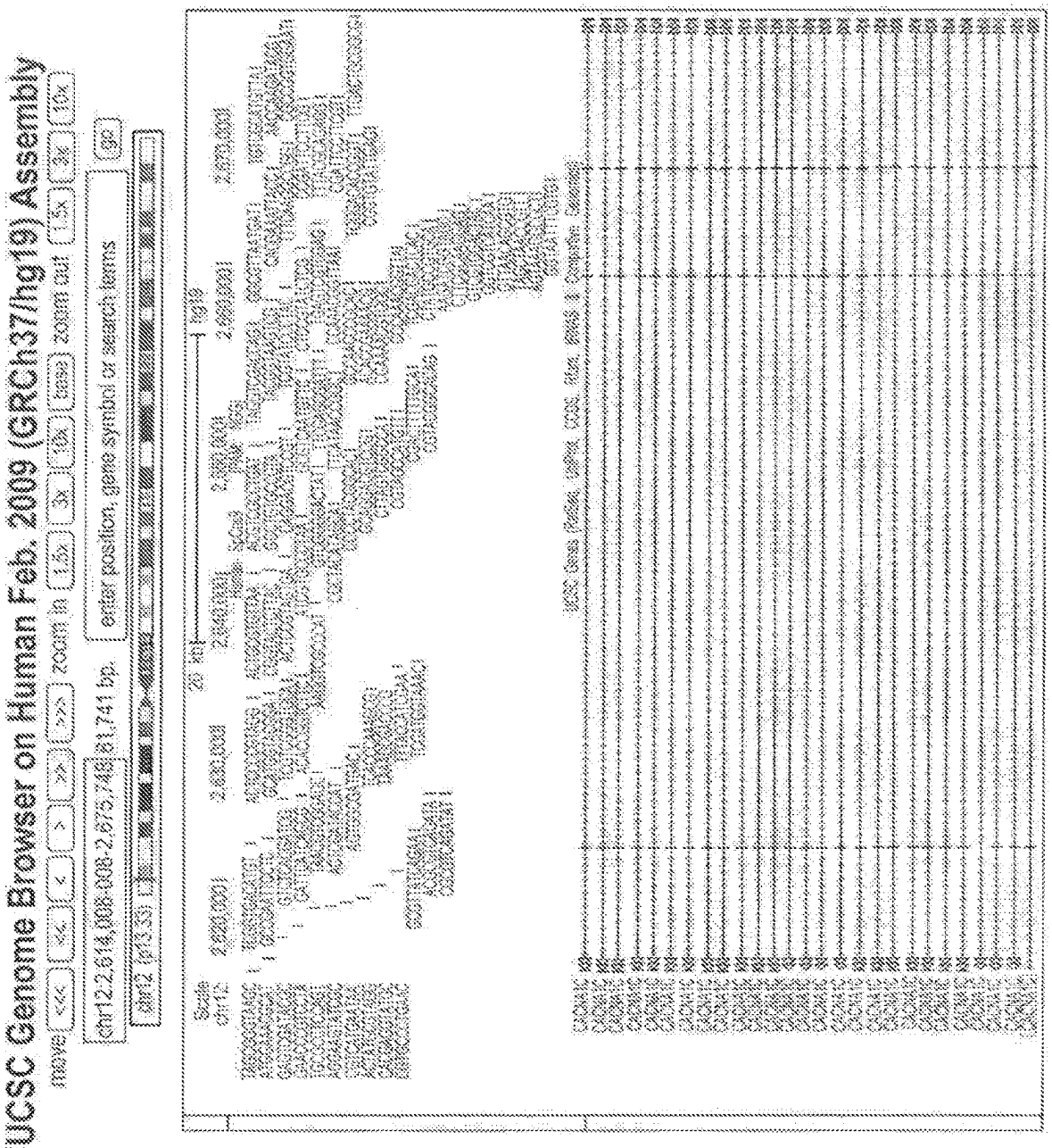
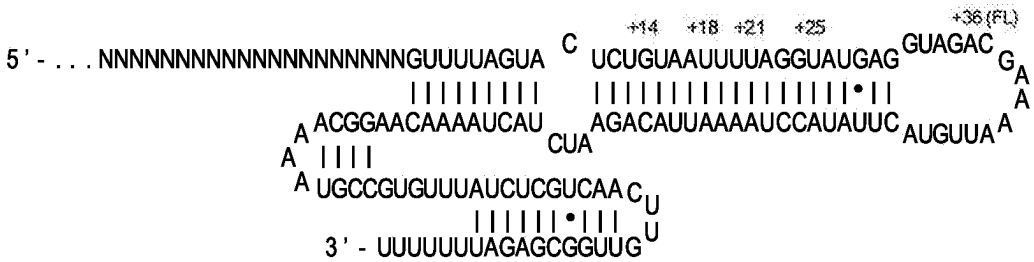

a
b
chimeric RNA
EMX1 target 1
EMX1 target 2
EMX1 target 3
EMX1 target 1
EMX1 target 2
EMX1 target 3
sgRNA design
indel (%) 11.0 11.2 11.8 11.2 11.0 9.8 9.2 11.4 7.9 10.3 8.6 10.1

FIG. 21A-B

FIG. 22


UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly
move
zoom in
zoom out
chr12:2,814,008-2,675,748
81,741 bp.
enter position, gene symbol or search terms
go
chr12 (p13.33)
hg19
Scale
chr12
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)

A


+14
+18
+21
+25
+36 (FL)
5' - . . . NNNNNNNNNNNNNNNNNNNNGUUUUAGUA C UCUGUAAUUUUAGGUAUGAG GUAGAC G A
A ACGGAACAAAAUCAU CUA AGACAUUAAAAUCCAUAUUC AUGUUA A
A UGCCGUGUUUAUCUCGUCAA C U
3' - UUUUUUUAGAGCGGUUG U

B

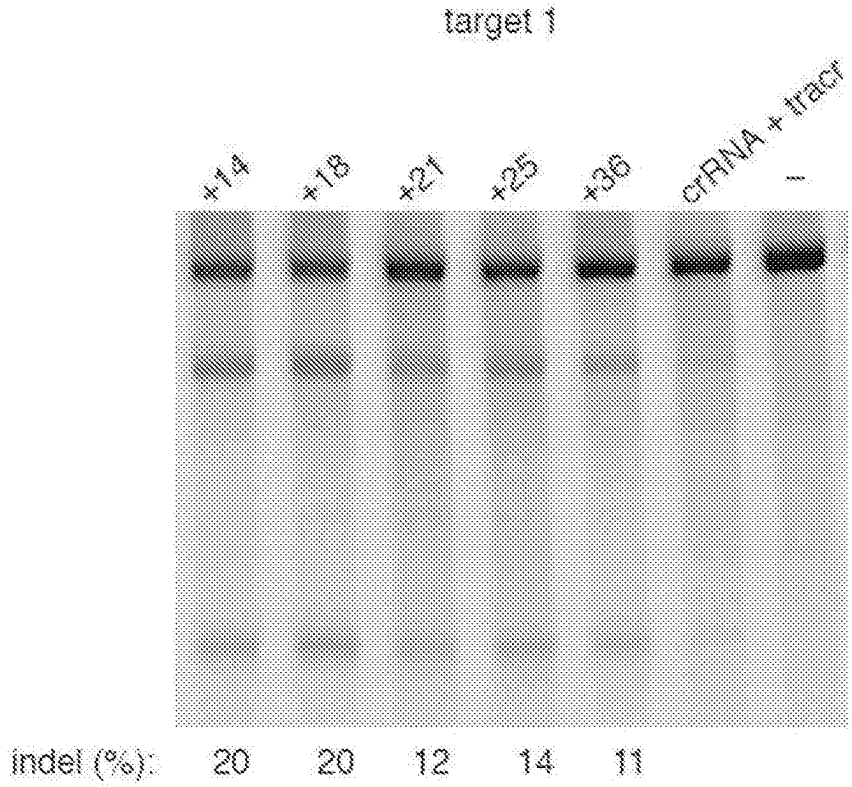

target 1
+14
+18
+21
+25
+36
crRNA + tracr
–
indel (%): 20 20 12 14 11

FIG. 23A-B

ENGINEERING OF SYSTEMS, METHODS AND OPTIMIZED GUIDE COMPOSITIONS FOR SEQUENCE MANIPULATION

RELATED APPLICATIONS AND INCORPORATION BY REFERENCE

[0001] This application is a continuation of U.S. patent application Ser. No. 15/230,025, filed Aug. 5, 2016, which is a continuation of U.S. application Ser. No. 14/104,990, filed Dec. 12, 2013 and which claims priority to U.S. provisional patent application 61/836,127 entitled ENGINEERING OF SYSTEMS, METHODS AND OPTIMIZED COMPOSITIONS FOR SEQUENCE MANIPULATION filed on Jun. 17, 2013. This application also claims priority to U.S. provisional patent applications 61/758,468; 61/769,046; 61/802,174; 61/806,375; 61/814,263; 61/819,803 and 61/828,130 each entitled ENGINEERING AND OPTIMIZATION OF SYSTEMS, METHODS AND COMPOSITIONS FOR SEQUENCE MANIPULATION, filed on Jan. 30, 2013; Feb. 25, 2013; Mar. 15, 2013; Mar. 28, 2013; Apr. 20, 2013; May 6, 2013 and May 28, 2013 respectively. Priority is also claimed to U.S. provisional patent applications 61/736,527 and 61/748,427, both entitled SYSTEMS METHODS AND COMPOSITIONS FOR SEQUENCE MANIPULATION filed on Dec. 12, 2012 and Jan. 2, 2013, respectively. Priority is also claimed to U.S. provisional patent applications 61/791,409 and 61/835,931 both entitled BI-2011/008/44790.02.2003 and BI-2011/008/44790.03.2003 filed on Mar. 15, 2013 and Jun. 17, 2013 respectively.

[0002] Reference is also made to U.S. provisional patent applications 61/835,936, 61/836,101, 61/836,080, 61/836,123 and 61/835,973 each filed Jun. 17, 2013.

[0003] The foregoing applications, and all documents cited therein or during their prosecution ("appln cited documents") and all documents cited or referenced in the appln cited documents, and all documents cited or referenced herein ("herein cited documents"), and all documents cited or referenced in herein cited documents, together with any manufacturer's instructions, descriptions, product specifications, and product sheets for any products mentioned herein or in any document incorporated by reference herein, are hereby incorporated herein by reference, and may be employed in the practice of the invention. More specifically, all referenced documents are incorporated by reference to the same extent as if each individual document was specifically and individually indicated to be incorporated by reference.

STATEMENT AS TO FEDERALLY SPONSORED RESEARCH

[0004] This invention was made with government support under grant number DP1MH100706 awarded by the National Institutes of Health. The government has certain rights in the invention.

FIELD OF THE INVENTION

[0005] The present invention generally relates to systems, methods and compositions used for the control of gene expression involving sequence targeting, such as genome perturbation or gene-editing, that may use vector systems related to Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) and components thereof.

SEQUENCE LISTING

[0006] The instant application contains a Sequence Listing which has been submitted electronically in XML format and is hereby incorporated by reference in its entirety. Said XML copy, created on Apr. 16, 2025, is named 114203-6217_SL and is 347,699 bytes in size.

BACKGROUND OF THE INVENTION

[0007] Recent advances in genome sequencing techniques and analysis methods have significantly accelerated the ability to catalog and map genetic factors associated with a diverse range of biological functions and diseases. Precise genome targeting technologies are needed to enable systematic reverse engineering of causal genetic variations by allowing selective perturbation of individual genetic elements, as well as to advance synthetic biology, biotechnological, and medical applications. Although genome-editing techniques such as designer zinc fingers, transcription activator-like effectors (TALEs), or homing meganucleases are available for producing targeted genome perturbations, there remains a need for new genome engineering technologies that are affordable, easy to set up, scalable, and amenable to targeting multiple positions within the eukaryotic genome.

SUMMARY OF THE INVENTION

[0008] There exists a pressing need for alternative and robust systems and techniques for sequence targeting with a wide array of applications. This invention addresses this need and provides related advantages. The CRISPR/Cas or the CRISPR-Cas system (both terms are used interchangeably throughout this application) does not require the generation of customized proteins to target specific sequences but rather a single Cas enzyme can be programmed by a short RNA molecule to recognize a specific DNA target, in other words the Cas enzyme can be recruited to a specific DNA target using said short RNA molecule. Adding the CRISPR-Cas system to the repertoire of genome sequencing techniques and analysis methods may significantly simplify the methodology and accelerate the ability to catalog and map genetic factors associated with a diverse range of biological functions and diseases. To utilize the CRISPR-Cas system effectively for genome editing without deleterious effects, it is critical to understand aspects of engineering and optimization of these genome engineering tools, which are aspects of the claimed invention.

[0009] In one aspect, the invention provides a vector system comprising one or more vectors. In some embodiments, the system comprises: (a) a first regulatory element operably linked to a tracr mate sequence and one or more insertion sites for inserting one or more guide sequences upstream of the tracr mate sequence, wherein when expressed, the guide sequence directs sequence-specific binding of a CRISPR complex to a target sequence in a cell, e.g., eukaryotic cell, wherein the CRISPR complex comprises a CRISPR enzyme complexed with (1) the guide sequence that is hybridized to the target sequence, and (2) the tracr mate sequence that is hybridized to the tracr sequence; and (b) a second regulatory element operably linked to an enzyme-coding sequence encoding said CRISPR enzyme comprising a nuclear localization sequence; wherein components (a) and (b) are located on the same or different vectors of the system. In some embodiments, component (a) further comprises the tracr sequence

downstream of the tracr mate sequence under the control of the first regulatory element. In some embodiments, component (a) further comprises two or more guide sequences operably linked to the first regulatory element, wherein when expressed, each of the two or more guide sequences direct sequence specific binding of a CRISPR complex to a different target sequence in a eukaryotic cell. In some embodiments, the system comprises the tracr sequence under the control of a third regulatory element, such as a polymerase III promoter. In some embodiments, the tracr sequence exhibits at least 50%, 60%, 70%, 80%, 90%, 95%, or 99% of sequence complementarity along the length of the tracr mate sequence when optimally aligned.

[0010] In some embodiments, the CRISPR complex comprises one or more nuclear localization sequences of sufficient strength to drive accumulation of said CRISPR complex in a detectable amount in the nucleus of a eukaryotic cell. Without wishing to be bound by theory, it is believed that a nuclear localization sequence is not necessary for CRISPR complex activity in eukaryotes, but that including such sequences enhances activity of the system, especially as to targeting nucleic acid molecules in the nucleus. In some embodiments, the CRISPR enzyme is a type II CRISPR system enzyme. In some embodiments, the CRISPR enzyme is a Cas9 enzyme. In some embodiments, the Cas9 enzyme is *S. pneumoniae, S. pyogenes*, or *S. thermophilus* Cas9, and may include mutated Cas9 derived from these organisms. The enzyme may be a Cas9 homolog or ortholog. In some embodiments, the CRISPR enzyme is codon-optimized for expression in a eukaryotic cell. In some embodiments, the CRISPR enzyme directs cleavage of one or two strands at the location of the target sequence. In some embodiments, the CRISPR enzyme lacks DNA strand cleavage activity. In some embodiments, the first regulatory element is a polymerase III promoter. In some embodiments, the second regulatory element is a polymerase II promoter. In some embodiments, the guide sequence is at least 15, 16, 17, 18, 19, 20, 25 nucleotides, or between 10-30, or between 15-25, or between 15-20 nucleotides in length.

[0011] Aspects of the invention comprehend one or more of the guide, tracr and tracr mate sequences are modified to improve stability in the CRISPR-Cas system chiRNA or CRISPR enzyme system. In an embodiment of the invention, the modification may comprise sequence optimization. In another embodiment, the modification may comprise reduction in polyT sequences in the tracr and/or tracr mate sequence. The invention provides that one or more Ts present in a poly-T sequence of the relevant wild type sequence have been substituted with a non-T nucleotide. The modified sequence may not comprise any polyT sequence having more than 4 contiguous Ts. In a further aspect of the invention the modification may comprise altering loops and/or hairpins. Embodiments of the invention encompass providing a minimum of two hairpins in the guide sequence or providing a hairpin formed by complementation between the tracr and tracr mate sequence or providing one or more further hairpin(s) at the 3' end of the tracrRNA sequence. Further embodiments of the invention encompass providing one or more additional hairpin(s) added to the 3' of the guide sequence or extending the 5' end of the guide sequence or providing one or more hairpins in the 5' end of the guide sequence. In a preferred embodiment, the modification comprises two hairpins or three hairpins or at most five hairpins.

[0012] In general, and throughout this specification, the term "vector" refers to a nucleic acid molecule capable of transporting another nucleic acid to which it has been linked. Vectors include, but are not limited to, nucleic acid molecules that are single-stranded, double-stranded, or partially double-stranded; nucleic acid molecules that comprise one or more free ends, no free ends (e.g. circular); nucleic acid molecules that comprise DNA, RNA, or both; and other varieties of polynucleotides known in the art. One type of vector is a "plasmid," which refers to a circular double stranded DNA loop into which additional DNA segments can be inserted, such as by standard molecular cloning techniques. Another type of vector is a viral vector, wherein virally-derived DNA or RNA sequences are present in the vector for packaging into a virus (e.g. retroviruses, replication defective retroviruses, adenoviruses, replication defective adenoviruses, and adeno-associated viruses). Viral vectors also include polynucleotides carried by a virus for transfection into a host cell. Certain vectors are capable of autonomous replication in a host cell into which they are introduced (e.g. bacterial vectors having a bacterial origin of replication and episomal mammalian vectors). Other vectors (e.g., non-episomal mammalian vectors) are integrated into the genome of a host cell upon introduction into the host cell, and thereby are replicated along with the host genome. Moreover, certain vectors are capable of directing the expression of genes to which they are operatively-linked. Such vectors are referred to herein as "expression vectors." Common expression vectors of utility in recombinant DNA techniques are often in the form of plasmids.

[0013] Recombinant expression vectors can comprise a nucleic acid of the invention in a form suitable for expression of the nucleic acid in a host cell, which means that the recombinant expression vectors include one or more regulatory elements, which may be selected on the basis of the host cells to be used for expression, that is operatively-linked to the nucleic acid sequence to be expressed. Within a recombinant expression vector, "operably linked" is intended to mean that the nucleotide sequence of interest is linked to the regulatory element(s) in a manner that allows for expression of the nucleotide sequence (e.g. in an in vitro transcription/translation system or in a host cell when the vector is introduced into the host cell).

[0014] The term "regulatory element" is intended to include promoters, enhancers, internal ribosomal entry sites (IRES), and other expression control elements (e.g. transcription termination signals, such as polyadenylation signals and poly-U sequences). Such regulatory elements are described, for example, in Goeddel, GENE EXPRESSION TECHNOLOGY: METHODS IN ENZYMOLOGY 185, Academic Press, San Diego, Calif. (1990). Regulatory elements include those that direct constitutive expression of a nucleotide sequence in many types of host cell and those that direct expression of the nucleotide sequence only in certain host cells (e.g., tissue-specific regulatory sequences). A tissue-specific promoter may direct expression primarily in a desired tissue of interest, such as muscle, neuron, bone, skin, blood, specific organs (e.g. liver, pancreas), or particular cell types (e.g. lymphocytes). Regulatory elements may also direct expression in a temporal-dependent manner, such as in a cell-cycle dependent or developmental stage-dependent manner, which may or may not also be tissue or cell-type specific. In some embodiments, a vector comprises one or more pol III promoter (e.g. 1, 2, 3, 4, 5, or more pol

III promoters), one or more pol II promoters (e.g. 1, 2, 3, 4, 5, or more pol II promoters), one or more pol I promoters (e.g. 1, 2, 3, 4, 5, or more pol I promoters), or combinations thereof. Examples of pol III promoters include, but are not limited to, U6 and H1 promoters. Examples of pol II promoters include, but are not limited to, the retroviral Rous sarcoma virus (RSV) LTR promoter (optionally with the RSV enhancer), the cytomegalovirus (CMV) promoter (optionally with the CMV enhancer) [see, e.g., Boshart et al, Cell, 41:521-530 (1985)], the SV40 promoter, the dihydrofolate reductase promoter, the 3-actin promoter, the phosphoglycerol kinase (PGK) promoter, and the EF1α promoter. Also encompassed by the term "regulatory element" are enhancer elements, such as WPRE; CMV enhancers; the R-U5' segment in LTR of HTLV-I (Mol. Cell. Biol., Vol. 8(1), p. 466-472, 1988); SV40 enhancer; and the intron sequence between exons 2 and 3 of rabbit β-globin (Proc. Natl. Acad. Sci. USA., Vol. 78(3), p. 1527-31, 1981). It will be appreciated by those skilled in the art that the design of the expression vector can depend on such factors as the choice of the host cell to be transformed, the level of expression desired, etc. A vector can be introduced into host cells to thereby produce transcripts, proteins, or peptides, including fusion proteins or peptides, encoded by nucleic acids as described herein (e.g., clustered regularly interspersed short palindromic repeats (CRISPR) transcripts, proteins, enzymes, mutant forms thereof, fusion proteins thereof, etc.).

[0015] Advantageous vectors include lentiviruses and adeno-associated viruses, and types of such vectors can also be selected for targeting particular types of cells.

[0016] In one aspect, the invention provides a vector comprising a regulatory element operably linked to an enzyme-coding sequence encoding a CRISPR enzyme comprising one or more nuclear localization sequences. In some embodiments, said regulatory element drives transcription of the CRISPR enzyme in a eukaryotic cell such that said CRISPR enzyme accumulates in a detectable amount in the nucleus of the eukaryotic cell. In some embodiments, the regulatory element is a polymerase II promoter. In some embodiments, the CRISPR enzyme is a type II CRISPR system enzyme. In some embodiments, the CRISPR enzyme is a Cas9 enzyme. In some embodiments, the Cas9 enzyme is *S. pneumoniae, S. pyogenes* or *S. thermophilus* Cas9, and may include mutated Cas9 derived from these organisms. In some embodiments, the CRISPR enzyme is codon-optimized for expression in a eukaryotic cell. In some embodiments, the CRISPR enzyme directs cleavage of one or two strands at the location of the target sequence. In some embodiments, the CRISPR enzyme lacks DNA strand cleavage activity.

[0017] In one aspect, the invention provides a CRISPR enzyme comprising one or more nuclear localization sequences of sufficient strength to drive accumulation of said CRISPR enzyme in a detectable amount in the nucleus of a eukaryotic cell. In some embodiments, the CRISPR enzyme is a type II CRISPR system enzyme. In some embodiments, the CRISPR enzyme is a Cas9 enzyme. In some embodiments, the Cas9 enzyme is *S. pneumoniae, S. pyogenes* or *S. thermophilus* Cas9, and may include mutated Cas9 derived from these organisms. The enzyme may be a Cas9 homolog or ortholog. In some embodiments, the CRISPR enzyme lacks the ability to cleave one or more strands of a target sequence to which it binds.

[0018] In one aspect, the invention provides a eukaryotic host cell comprising (a) a first regulatory element operably linked to a tracr mate sequence and one or more insertion sites for inserting one or more guide sequences upstream of the tracr mate sequence, wherein when expressed, the guide sequence directs sequence-specific binding of a CRISPR complex to a target sequence in a eukaryotic cell, wherein the CRISPR complex comprises a CRISPR enzyme complexed with (1) the guide sequence that is hybridized to the target sequence, and (2) the tracr mate sequence that is hybridized to the tracr sequence; and/or (b) a second regulatory element operably linked to an enzyme-coding sequence encoding said CRISPR enzyme comprising a nuclear localization sequence. In some embodiments, the host cell comprises components (a) and (b). In some embodiments, component (a), component (b), or components (a) and (b) are stably integrated into a genome of the host eukaryotic cell. In some embodiments, component (a) further comprises the tracr sequence downstream of the tracr mate sequence under the control of the first regulatory element. In some embodiments, component (a) further comprises two or more guide sequences operably linked to the first regulatory element, wherein when expressed, each of the two or more guide sequences direct sequence specific binding of a CRISPR complex to a different target sequence in a eukaryotic cell. In some embodiments, the eukaryotic host cell further comprises a third regulatory element, such as a polymerase III promoter, operably linked to said tracr sequence. In some embodiments, the tracr sequence exhibits at least 50%, 60%, 70%, 80%, 90%, 95%, or 99% of sequence complementarity along the length of the tracr mate sequence when optimally aligned. In some embodiments, the CRISPR enzyme comprises one or more nuclear localization sequences of sufficient strength to drive accumulation of said CRISPR enzyme in a detectable amount in the nucleus of a eukaryotic cell. In some embodiments, the CRISPR enzyme is a type II CRISPR system enzyme. In some embodiments, the CRISPR enzyme is a Cas9 enzyme. In some embodiments, the Cas9 enzyme is *S. pneumoniae, S. pyogenes* or *S. thermophilus* Cas9, and may include mutated Cas9 derived from these organisms. The enzyme may be a Cas9 homolog or ortholog. In some embodiments, the CRISPR enzyme is codon-optimized for expression in a eukaryotic cell. In some embodiments, the CRISPR enzyme directs cleavage of one or two strands at the location of the target sequence. In some embodiments, the CRISPR enzyme lacks DNA strand cleavage activity. In some embodiments, the first regulatory element is a polymerase III promoter. In some embodiments, the second regulatory element is a polymerase II promoter. In some embodiments, the guide sequence is at least 15, 16, 17, 18, 19, 20, 25 nucleotides, or between 10-30, or between 15-25, or between 15-20 nucleotides in length. In an aspect, the invention provides a non-human eukaryotic organism; preferably a multicellular eukaryotic organism, comprising a eukaryotic host cell according to any of the described embodiments. In other aspects, the invention provides a eukaryotic organism; preferably a multicellular eukaryotic organism, comprising a eukaryotic host cell according to any of the described embodiments. The organism in some embodiments of these aspects may be an animal; for example a mammal. Also, the organism may be an arthropod such as an insect. The organism also may be a plant. Further, the organism may be a fungus.

[0019] In one aspect, the invention provides a kit comprising one or more of the components described herein. In some embodiments, the kit comprises a vector system and instructions for using the kit. In some embodiments, the vector system comprises (a) a first regulatory element operably linked to a tracr mate sequence and one or more insertion sites for inserting one or more guide sequences upstream of the tracr mate sequence, wherein when expressed, the guide sequence directs sequence-specific binding of a CRISPR complex to a target sequence in a eukaryotic cell, wherein the CRISPR complex comprises a CRISPR enzyme complexed with (1) the guide sequence that is hybridized to the target sequence, and (2) the tracr mate sequence that is hybridized to the tracr sequence; and/or (b) a second regulatory element operably linked to an enzyme-coding sequence encoding said CRISPR enzyme comprising a nuclear localization sequence. In some embodiments, the kit comprises components (a) and (b) located on the same or different vectors of the system. In some embodiments, component (a) further comprises the tracr sequence downstream of the tracr mate sequence under the control of the first regulatory element. In some embodiments, component (a) further comprises two or more guide sequences operably linked to the first regulatory element, wherein when expressed, each of the two or more guide sequences direct sequence specific binding of a CRISPR complex to a different target sequence in a eukaryotic cell. In some embodiments, the system further comprises a third regulatory element, such as a polymerase III promoter, operably linked to said tracr sequence. In some embodiments, the tracr sequence exhibits at least 50%, 60%, 70%, 80%, 90%, 95%, or 99% of sequence complementarity along the length of the tracr mate sequence when optimally aligned. In some embodiments, the CRISPR enzyme comprises one or more nuclear localization sequences of sufficient strength to drive accumulation of said CRISPR enzyme in a detectable amount in the nucleus of a eukaryotic cell. In some embodiments, the CRISPR enzyme is a type II CRISPR system enzyme. In some embodiments, the CRISPR enzyme is a Cas9 enzyme. In some embodiments, the Cas9 enzyme is *S. pneumoniae, S. pyogenes* or *S. thermophilus* Cas9, and may include mutated Cas9 derived from these organisms. The enzyme may be a Cas9 homolog or ortholog. In some embodiments, the CRISPR enzyme is codon-optimized for expression in a eukaryotic cell. In some embodiments, the CRISPR enzyme directs cleavage of one or two strands at the location of the target sequence. In some embodiments, the CRISPR enzyme lacks DNA strand cleavage activity. In some embodiments, the first regulatory element is a polymerase III promoter. In some embodiments, the second regulatory element is a polymerase II promoter. In some embodiments, the guide sequence is at least 15, 16, 17, 18, 19, 20, 25 nucleotides, or between 10-30, or between 15-25, or between 15-20 nucleotides in length.

[0020] In one aspect, the invention provides a method of modifying a target polynucleotide in a eukaryotic cell. In some embodiments, the method comprises allowing a CRISPR complex to bind to the target polynucleotide to effect cleavage of said target polynucleotide thereby modifying the target polynucleotide, wherein the CRISPR complex comprises a CRISPR enzyme complexed with a guide sequence hybridized to a target sequence within said target polynucleotide, wherein said guide sequence is linked to a tracr mate sequence which in turn hybridizes to a tracr sequence. In some embodiments, said cleavage comprises cleaving one or two strands at the location of the target sequence by said CRISPR enzyme. In some embodiments, said cleavage results in decreased transcription of a target gene. In some embodiments, the method further comprises repairing said cleaved target polynucleotide by homologous recombination with an exogenous template polynucleotide, wherein said repair results in a mutation comprising an insertion, deletion, or substitution of one or more nucleotides of said target polynucleotide. In some embodiments, said mutation results in one or more amino acid changes in a protein expressed from a gene comprising the target sequence. In some embodiments, the method further comprises delivering one or more vectors to said eukaryotic cell, wherein the one or more vectors drive expression of one or more of: the CRISPR enzyme, the guide sequence linked to the tracr mate sequence, and the tracr sequence. In some embodiments, said vectors are delivered to the eukaryotic cell in a subject. In some embodiments, said modifying takes place in said eukaryotic cell in a cell culture. In some embodiments, the method further comprises isolating said eukaryotic cell from a subject prior to said modifying. In some embodiments, the method further comprises returning said eukaryotic cell and/or cells derived therefrom to said subject.

[0021] In one aspect, the invention provides a method of modifying expression of a polynucleotide in a eukaryotic cell. In some embodiments, the method comprises allowing a CRISPR complex to bind to the polynucleotide such that said binding results in increased or decreased expression of said polynucleotide; wherein the CRISPR complex comprises a CRISPR enzyme complexed with a guide sequence hybridized to a target sequence within said polynucleotide, wherein said guide sequence is linked to a tracr mate sequence which in turn hybridizes to a tracr sequence. In some embodiments, the method further comprises delivering one or more vectors to said eukaryotic cells, wherein the one or more vectors drive expression of one or more of: the CRISPR enzyme, the guide sequence linked to the tracr mate sequence, and the tracr sequence.

[0022] In one aspect, the invention provides a method of generating a model eukaryotic cell comprising a mutated disease gene. In some embodiments, a disease gene is any gene associated with an increase in the risk of having or developing a disease. In some embodiments, the method comprises (a) introducing one or more vectors into a eukaryotic cell, wherein the one or more vectors drive expression of one or more of: a CRISPR enzyme, a guide sequence linked to a tracr mate sequence, and a tracr sequence; and (b) allowing a CRISPR complex to bind to a target polynucleotide to effect cleavage of the target polynucleotide within said disease gene, wherein the CRISPR complex comprises the CRISPR enzyme complexed with (1) the guide sequence that is hybridized to the target sequence within the target polynucleotide, and (2) the tracr mate sequence that is hybridized to the tracr sequence, thereby generating a model eukaryotic cell comprising a mutated disease gene. In some embodiments, said cleavage comprises cleaving one or two strands at the location of the target sequence by said CRISPR enzyme. In some embodiments, said cleavage results in decreased transcription of a target gene. In some embodiments, the method further comprises repairing said cleaved target polynucleotide by homologous recombination with an exogenous template polynucleotide, wherein said

repair results in a mutation comprising an insertion, deletion, or substitution of one or more nucleotides of said target polynucleotide. In some embodiments, said mutation results in one or more amino acid changes in a protein expression from a gene comprising the target sequence.

[0023] In one aspect, the invention provides a method for developing a biologically active agent that modulates a cell signaling event associated with a disease gene. In some embodiments, a disease gene is any gene associated with an increase in the risk of having or developing a disease. In some embodiments, the method comprises (a) contacting a test compound with a model cell of any one of the described embodiments; and (b) detecting a change in a readout that is indicative of a reduction or an augmentation of a cell signaling event associated with said mutation in said disease gene, thereby developing said biologically active agent that modulates said cell signaling event associated with said disease gene.

[0024] In one aspect, the invention provides a recombinant polynucleotide comprising a guide sequence upstream of a tracr mate sequence, wherein the guide sequence when expressed directs sequence-specific binding of a CRISPR complex to a corresponding target sequence present in a eukaryotic cell. In some embodiments, the target sequence is a viral sequence present in a eukaryotic cell. In some embodiments, the target sequence is a proto-oncogene or an oncogene.

[0025] In one aspect the invention provides for a method of selecting one or more prokaryotic cell(s) by introducing one or more mutations in a gene in the one or more prokaryotic cell (s), the method comprising: introducing one or more vectors into the prokaryotic cell (s), wherein the one or more vectors drive expression of one or more of: a CRISPR enzyme, a guide sequence linked to a tracr mate sequence, a tracr sequence, and a editing template; wherein the editing template comprises the one or more mutations that abolish CRISPR enzyme cleavage; allowing homologous recombination of the editing template with the target polynucleotide in the cell(s) to be selected; allowing a CRISPR complex to bind to a target polynucleotide to effect cleavage of the target polynucleotide within said gene, wherein the CRISPR complex comprises the CRISPR enzyme complexed with (1) the guide sequence that is hybridized to the target sequence within the target polynucleotide, and (2) the tracr mate sequence that is hybridized to the tracr sequence, wherein binding of the CRISPR complex to the target polynucleotide induces cell death, thereby allowing one or more prokaryotic cell(s) in which one or more mutations have been introduced to be selected. In a preferred embodiment, the CRISPR enzyme is Cas9. In another aspect of the invention the cell to be selected may be a eukaryotic cell. Aspects of the invention allow for selection of specific cells without requiring a selection marker or a two-step process that may include a counter-selection system.

[0026] In some aspects the invention provides a non-naturally occurring or engineered composition comprising a CRISPR-Cas system chimeric RNA (chiRNA) polynucleotide sequence, wherein the polynucleotide sequence comprises (a) a guide sequence capable of hybridizing to a target sequence in a eukaryotic cell, (b) a tracr mate sequence, and (c) a tracr sequence wherein (a), (b) and (c) are arranged in a 5' to 3' orientation, wherein when transcribed, the tracr mate sequence hybridizes to the tracr sequence and the guide sequence directs sequence-specific binding of a CRISPR complex to the target sequence, wherein the CRISPR complex comprises a CRISPR enzyme complexed with (1) the guide sequence that is hybridized to the target sequence, and (2) the tracr mate sequence that is hybridized to the tracr sequence,

[0027] or

[0028] a CRISPR enzyme system, wherein the system is encoded by a vector system comprising one or more vectors comprising I. a first regulatory element operably linked to a CRISPR-Cas system chimeric RNA (chiRNA) polynucleotide sequence, wherein the polynucleotide sequence comprises (a) one or more guide sequences capable of hybridizing to one or more target sequences in a eukaryotic cell, (b) a tracr mate sequence, and (c) one or more tracr sequences, and II. a second regulatory element operably linked to an enzyme-coding sequence encoding a CRISPR enzyme comprising at least one or more nuclear localization sequences, wherein (a), (b) and (c) are arranged in a 5' to 3'orientation, wherein components I and II are located on the same or different vectors of the system, wherein when transcribed, the tracr mate sequence hybridizes to the tracr sequence and the guide sequence directs sequence-specific binding of a CRISPR complex to the target sequence, wherein the CRISPR complex comprises the CRISPR enzyme complexed with (1) the guide sequence that is hybridized to the target sequence, and (2) the tracr mate sequence that is hybridized to the tracr sequence, or a multiplexed CRISPR enzyme system, wherein the system is encoded by a vector system comprising one or more vectors comprising I. a first regulatory element operably linked to (a) one or more guide sequences capable of hybridizing to a target sequence in a cell, and (b) at least one or more tracr mate sequences, II. a second regulatory element operably linked to an enzyme-coding sequence encoding a CRISPR enzyme, and III. a third regulatory element operably linked to a tracr sequence, wherein components I, II and III are located on the same or different vectors of the system, wherein when transcribed, the tracr mate sequence hybridizes to the tracr sequence and the guide sequence directs sequence-specific binding of a CRISPR complex to the target sequence, wherein the CRISPR complex comprises the CRISPR enzyme complexed with (1) the guide sequence that is hybridized to the target sequence, and (2) the tracr mate sequence that is hybridized to the tracr sequence, and wherein in the multiplexed system multiple guide sequences and a single tracr sequence is used; and wherein one or more of the guide, tracr and tracr mate sequences are modified to improve stability.

[0029] In aspects of the invention, the modification comprises an engineered secondary structure. For example, the modification can comprise a reduction in a region of hybridization between the tracr mate sequence and the tracr sequence. For example, the modification also may comprise fusing the tracr mate sequence and the tracr sequence through an artificial loop. The modification may comprise the tracr sequence having a length between 40 and 120 bp. In embodiments of the invention, the tracr sequence is between 40 bp and full length of the tracr. In certain embodiments, the length of tracRNA includes at least nucleotides 1-67 and in some embodiments at least nucleotides 1-85 of the wild type tracRNA. In some embodiments, at least nucleotides corresponding to nucleotides 1-67 or 1-85 of wild type *S. pyogenes* Cas9 tracRNA may be used. Where the CRISPR system uses enzymes other than Cas9,

or other than SpCas9, then corresponding nucleotides in the relevant wild type tracRNA may be present. In some embodiments, the length of tracRNA includes no more than nucleotides 1-67 or 1-85 of the wild type tracRNA. The modification may comprise sequence optimization. In certain aspects, sequence optimization may comprise reducing the incidence of polyT sequences in the tracr and/or tracr mate sequence. Sequence optimization may be combined with reduction in the region of hybridization between the tracr mate sequence and the tracr sequence; for example, a reduced length tracr sequence.

[0030] In an aspect the invention provides the CRISPR-Cas system or CRISPR enzyme system wherein the modification comprises reduction in polyT sequences in the tracr and/or tracr mate sequence. In some aspects of the invention, one or more Ts present in a poly-T sequence of the relevant wild type sequence (that is, a stretch of more than 3, 4, 5, 6, or more contiguous T bases; in some embodiments, a stretch of no more than 10, 9, 8, 7, 6 contiguous T bases) may be substituted with a non-T nucleotide, e.g., an A, so that the string is broken down into smaller stretches of Ts with each stretch having 4, or fewer than 4 (for example, 3 or 2) contiguous Ts. Bases other than A may be used for substitution, for example C or G, or non-naturally occurring nucleotides or modified nucleotides. If the string of Ts is involved in the formation of a hairpin (or stem loop), then it is advantageous that the complementary base for the non-T base be changed to complement the non-T nucleotide. For example, if the non-T base is an A, then its complement may be changed to a T, e.g., to preserve or assist in the preservation of secondary structure. For instance, 5'-TTTTT can be altered to become 5'-TTTAT and the complementary 5'-AAAAA can be changed into 5'-ATAAA.

[0031] In an aspect the invention provides the CRISPR-Cas system or CRISPR enzyme system wherein the modification comprises adding a polyT terminator sequence. In an aspect the invention provides the CRISPR-Cas system or CRISPR enzyme system wherein the modification comprises adding a polyT terminator sequence in tracr and/or tracr mate sequences. In an aspect the invention provides the CRISPR-Cas system or CRISPR enzyme system wherein the modification comprises adding a polyT terminator sequence in the guide sequence. The polyT terminator sequence may comprise 5 contiguous T bases, or more than 5.

[0032] In an aspect the invention provides the CRISPR-Cas system or CRISPR enzyme system wherein the modification comprises altering loops and/or hairpins. In an aspect the invention provides the CRISPR-Cas system or CRISPR enzyme system wherein the modification comprises providing a minimum of two hairpins in the guide sequence. In an aspect the invention provides the CRISPR-Cas system or CRISPR enzyme system wherein the modification comprises providing a hairpin formed by complementation between the tracr and tracr mate (direct repeat) sequence. In an aspect the invention provides the CRISPR-Cas system or CRISPR enzyme system wherein the modification comprises providing one or more further hairpin(s) at or towards the 3' end of the tracrRNA sequence. For example, a hairpin may be formed by providing self complementary sequences within the tracrRNA sequence joined by a loop such that a hairpin is formed on self folding. In an aspect the invention provides the CRISPR-Cas system or CRISPR enzyme system wherein the modification comprises providing additional hairpins added to the 3' of the guide sequence. In an aspect the invention provides the CRISPR-Cas system or CRISPR enzyme system wherein the modification comprises extending the 5' end of the guide sequence. In an aspect the invention provides the CRISPR-Cas system or CRISPR enzyme system wherein the modification comprises providing one or more hairpins in the 5' end of the guide sequence. In an aspect the invention provides the CRISPR-Cas system or CRISPR enzyme system wherein the modification comprises appending the sequence (5'-AGGACGAAGTCCTAA) (SEQ ID NO: 1) to the 5' end of the guide sequence. Other sequences suitable for forming hairpins will be known to the skilled person, and may be used in certain aspects of the invention. In some aspects of the invention, at least 2, 3, 4, 5, or more additional hairpins are provided. In some aspects of the invention, no more than 10, 9, 8, 7, 6 additional hairpins are provided. In an aspect the invention provides the CRISPR-Cas system or CRISPR enzyme system wherein the modification comprises two hairpins. In an aspect the invention provides the CRISPR-Cas system or CRISPR enzyme system wherein the modification comprises three hairpins. In an aspect the invention provides the CRISPR-Cas system or CRISPR enzyme system wherein the modification comprises at most five hairpins.

[0033] In an aspect the invention provides the CRISPR-Cas system or CRISPR enzyme system wherein the modification comprises providing cross linking, or providing one or more modified nucleotides in the polynucleotide sequence. Modified nucleotides and/or cross linking may be provided in any or all of the tracr, tracr mate, and/or guide sequences, and/or in the enzyme coding sequence, and/or in vector sequences. Modifications may include inclusion of at least one non naturally occurring nucleotide, or a modified nucleotide, or analogs thereof. Modified nucleotides may be modified at the ribose, phosphate, and/or base moiety. Modified nucleotides may include 2'-O-methyl analogs, 2'-deoxy analogs, or 2'-fluoro analogs. The nucleic acid backbone may be modified, for example, a phosphorothioate backbone may be used. The use of locked nucleic acids (LNA) or bridged nucleic acids (BNA) may also be possible. Further examples of modified bases include, but are not limited to, 2-aminopurine, 5-bromo-uridine, pseudouridine, inosine, 7-methylguanosine.

[0034] It will be understood that any or all of the above modifications may be provided in isolation or in combination in a given CRISPR-Cas system or CRISPR enzyme system. Such a system may include one, two, three, four, five, or more of said modifications.

[0035] In an aspect the invention provides the CRISPR-Cas system or CRISPR enzyme system wherein the CRISPR enzyme is a type II CRISPR system enzyme, e.g., a Cas9 enzyme. In an aspect the invention provides the CRISPR-Cas system or CRISPR enzyme system wherein the CRISPR enzyme is comprised of less than one thousand amino acids, or less than four thousand amino acids. In an aspect the invention provides the CRISPR-Cas system or CRISPR enzyme system wherein the Cas9 enzyme is StCas9 or St1Cas9, or the Cas9 enzyme is a Cas9 enzyme from an organism selected from the group consisting of genus *Streptococcus, Campylobacter, Nitratifractor, Staphylococcus, Parvibaculum, Roseburia, Neisseria, Gluconacetobacter*, Azospirillum, Sphaerochaeta, *Lactobacillus, Eubacterium* or *Corynebacter*. In an aspect the invention provides the

CRISPR-Cas system or CRISPR enzyme system wherein the CRISPR enzyme is a nuclease directing cleavage of both strands at the location of the target sequence.

[0036] In an aspect the invention provides the CRISPR-Cas system or CRISPR enzyme system wherein the first regulatory element is a polymerase III promoter. In an aspect the invention provides the CRISPR-Cas system or CRISPR enzyme system wherein the second regulatory element is a polymerase II promoter.

[0037] In an aspect the invention provides the CRISPR-Cas system or CRISPR enzyme system wherein the guide sequence comprises at least fifteen nucleotides.

[0038] In an aspect the invention provides the CRISPR-Cas system or CRISPR enzyme system wherein the modification comprises optimized tracr sequence and/or optimized guide sequence RNA and/or co-fold structure of tracr sequence and/or tracr mate sequence(s) and/or stabilizing secondary structures of tracr sequence and/or tracr sequence with a reduced region of base-pairing and/or tracr sequence fused RNA elements; and/or, in the multiplexed system there are two RNAs comprising a tracer and comprising a plurality of guides or one RNA comprising a plurality of chimerics.

[0039] In aspects of the invention the chimeric RNA architecture is further optimized according to the results of mutagenesis studies. In chimeric RNA with two or more hairpins, mutations in the proximal direct repeat to stabilize the hairpin may result in ablation of CRISPR complex activity. Mutations in the distal direct repeat to shorten or stabilize the hairpin may have no effect on CRISPR complex activity. Sequence randomization in the bulge region between the proximal and distal repeats may significantly reduce CRISPR complex activity. Single base pair changes or sequence randomization in the linker region between hairpins may result in complete loss of CRISPR complex activity. Hairpin stabilization of the distal hairpins that follow the first hairpin after the guide sequence may result in maintenance or improvement of CRISPR complex activity. Accordingly, in preferred embodiments of the invention, the chimeric RNA architecture may be further optimized by generating a smaller chimeric RNA which may be beneficial for therapeutic delivery options and other uses and this may be achieved by altering the distal direct repeat so as to shorten or stabilize the hairpin. In further preferred embodiments of the invention, the chimeric RNA architecture may be further optimized by stabilizing one or more of the distal hairpins. Stabilization of hairpins may include modifying sequences suitable for forming hairpins. In some aspects of the invention, at least 2, 3, 4, 5, or more additional hairpins are provided. In some aspects of the invention, no more than 10, 9, 8, 7, 6 additional hairpins are provided. In some aspects of the invention stabilization may be cross linking and other modifications. Modifications may include inclusion of at least one non naturally occurring nucleotide, or a modified nucleotide, or analogs thereof. Modified nucleotides may be modified at the ribose, phosphate, and/or base moiety. Modified nucleotides may include 2'-O-methyl analogs, 2'-deoxy analogs, or 2'-fluoro analogs. The nucleic acid backbone may be modified, for example, a phosphorothioate backbone may be used. The use of locked nucleic acids (LNA) or bridged nucleic acids (BNA) may also be possible. Further examples of modified bases include, but are not limited to, 2-aminopurine, 5-bromo-uridine, pseudouridine, inosine, 7-methylguanosine.

[0040] In an aspect the invention provides the CRISPR-Cas system or CRISPR enzyme system wherein the CRISPR enzyme is codon-optimized for expression in a eukaryotic cell.

[0041] Accordingly, in some aspects of the invention, the length of tracRNA required in a construct of the invention, e.g., a chimeric construct, need not necessarily be fixed, and in some aspects of the invention it can be between 40 and 120 bp, and in some aspects of the invention up to the full length of the tracr, e.g., in some aspects of the invention, until the 3' end of tracr as punctuated by the transcription termination signal in the bacterial genome. In certain embodiments, the length of tracRNA includes at least nucleotides 1-67 and in some embodiments at least nucleotides 1-85 of the wild type tracRNA. In some embodiments, at least nucleotides corresponding to nucleotides 1-67 or 1-85 of wild type *S. pyogenes* Cas9 tracRNA may be used. Where the CRISPR system uses enzymes other than Cas9, or other than SpCas9, then corresponding nucleotides in the relevant wild type tracRNA may be present. In some embodiments, the length of tracRNA includes no more than nucleotides 1-67 or 1-85 of the wild type tracRNA With respect to sequence optimization (e.g., reduction in polyT sequences), e.g., as to strings of Ts internal to the tracr mate (direct repeat) or tracrRNA, in some aspects of the invention, one or more Ts present in a poly-T sequence of the relevant wild type sequence (that is, a stretch of more than 3, 4, 5, 6, or more contiguous T bases; in some embodiments, a stretch of no more than 10, 9, 8, 7, 6 contiguous T bases) may be substituted with a non-T nucleotide, e.g., an A, so that the string is broken down into smaller stretches of Ts with each stretch having 4, or fewer than 4 (for example, 3 or 2) contiguous Ts. If the string of Ts is involved in the formation of a hairpin (or stem loop), then it is advantageous that the complementary base for the non-T base be changed to complement the non-T nucleotide. For example, if the non-T base is an A, then its complement may be changed to a T, e.g., to preserve or assist in the preservation of secondary structure. For instance, 5'-TTTTT can be altered to become 5'-TTTAT and the complementary 5'-AAAAA can be changed into 5'-ATAAA. As to the presence of polyT terminator sequences in tracr+tracr mate transcript, e.g., a polyT terminator (TTTTT or more), in some aspects of the invention it is advantageous that such be added to end of the transcript, whether it is in two RNA (tracr and tracr mate) or single guide RNA form. Concerning loops and hairpins in tracr and tracr mate transcripts, in some aspects of the invention it is advantageous that a minimum of two hairpins be present in the chimeric guide RNA. A first hairpin can be the hairpin formed by complementation between the tracr and tracr mate (direct repeat) sequence. A second hairpin can be at the 3' end of the tracrRNA sequence, and this can provide secondary structure for interaction with Cas9. Additional hairpins may be added to the 3' of the guide RNA, e.g., in some aspects of the invention to increase the stability of the guide RNA. Additionally, the 5' end of the guide RNA, in some aspects of the invention, may be extended. In some aspects of the invention, one may consider 20 bp in the 5' end as a guide sequence. The 5' portion may be extended. One or more hairpins can be provided in the 5' portion, e.g., in some aspects of the invention, this may also improve the stability of the guide RNA. In some aspects of the invention, the specific hairpin can be provided by appending the sequence (5'-AGGACGAAGTCCTAA) (SEQ ID NO: 1) to

the 5' end of the guide sequence, and, in some aspects of the invention, this may help improve stability. Other sequences suitable for forming hairpins will be known to the skilled person, and may be used in certain aspects of the invention. In some aspects of the invention, at least 2, 3, 4, 5, or more additional hairpins are provided. In some aspects of the invention, no more than 10, 9, 8, 7, 6 additional hairpins are provided. The foregoing also provides aspects of the invention involving secondary structure in guide sequences. In some aspects of the invention there may be cross linking and other modifications, e.g., to improve stability. Modifications may include inclusion of at least one non naturally occurring nucleotide, or a modified nucleotide, or analogs thereof. Modified nucleotides may be modified at the ribose, phosphate, and/or base moiety. Modified nucleotides may include 2'-O-methyl analogs, 2'-deoxy analogs, or 2'-fluoro analogs. The nucleic acid backbone may be modified, for example, a phosphorothioate backbone may be used. The use of locked nucleic acids (LNA) or bridged nucleic acids (BNA) may also be possible. Further examples of modified bases include, but are not limited to, 2-aminopurine, 5-bromouridine, pseudouridine, inosine, 7-methylguanosine. Such modifications or cross linking may be present in the guide sequence or other sequences adjacent the guide sequence.

[0042] Accordingly, it is an object of the invention not to encompass within the invention any previously known product, process of making the product, or method of using the product such that Applicants reserve the right and hereby disclose a disclaimer of any previously known product, process, or method. It is further noted that the invention does not intend to encompass within the scope of the invention any product, process, or making of the product or method of using the product, which does not meet the written description and enablement requirements of the USPTO (35 U.S.C. § 112, first paragraph) or the EPO (Article 83 of the EPC), such that Applicants reserve the right and hereby disclose a disclaimer of any previously described product, process of making the product, or method of using the product.

[0043] It is noted that in this disclosure and particularly in the claims and/or paragraphs, terms such as "comprises", "comprised", "comprising" and the like can have the meaning attributed to it in U.S. Patent law; e.g., they can mean "includes", "included", "including", and the like; and that terms such as "consisting essentially of" and "consists essentially of" have the meaning ascribed to them in U.S. Patent law, e.g., they allow for elements not explicitly recited, but exclude elements that are found in the prior art or that affect a basic or novel characteristic of the invention. These and other embodiments are disclosed or are obvious from and encompassed by, the following Detailed Description.

BRIEF DESCRIPTION OF THE DRAWINGS

[0044] The novel features of the invention are set forth with particularity in the appended claims. A better understanding of the features and advantages of the present invention will be obtained by reference to the following detailed description that sets forth illustrative embodiments, in which the principles of the invention are utilized, and the accompanying drawings of which:

[0045] FIG. **1** shows a schematic model of the CRISPR system. The Cas9 nuclease from *Streptococcus pyogenes* (yellow) is targeted to genomic DNA by a synthetic guide RNA (sgRNA) consisting of a 20-nt guide sequence (blue) and a scaffold (red). The guide sequence base-pairs with the DNA target (blue), directly upstream of a requisite 5'-NGG protospacer adjacent motif (PAM; magenta), and Cas9 mediates a double-stranded break (DSB) ~3 bp upstream of the PAM (red triangle).

[0046] FIG. **2**A-F illustrates an exemplary CRISPR system, a possible mechanism of action, an example adaptation for expression in eukaryotic cells, and results of tests assessing nuclear localization and CRISPR activity. FIG. **2**C discloses SEQ ID NOS 103-104, respectively, in order of appearance. FIG. **2**E discloses SEQ ID NOS 105-107, respectively, in order of appearance. FIG. **2**F discloses SEQ ID NOS 108-112, respectively, in order of appearance.

[0047] FIG. **3**A-C illustrates an exemplary expression cassette for expression of CRISPR system elements in eukaryotic cells, predicted structures of example guide sequences, and CRISPR system activity as measured in eukaryotic and prokaryotic cells. FIG. **3**A discloses SEQ ID NO: 113. FIG. **3**B discloses SEQ ID NOS 114-122, respectively, in order of appearance.

[0048] FIG. **4**A-D illustrates results of an evaluation of SpCas9 specificity for an example target. FIG. **4**A discloses SEQ ID NOS 123, 106 and 124-134, respectively, in order of appearance. FIG. **4**C discloses SEQ ID NO: 123.

[0049] FIG. **5**A-G illustrates an exemplary vector system and results for its use in directing homologous recombination in eukaryotic cells. FIG. **5**E discloses SEQ ID NO: 135. FIG. **5**F discloses SEQ ID NOS 136 and 137, respectively, in order of appearance. FIG. **5**G discloses SEQ ID NOS 138-142, respectively, in order of appearance.

[0050] FIG. **6**A-C illustrates a comparison of different tracrRNA transcripts for Cas9-mediated gene targeting. FIG. **6**A discloses SEQ ID NOS 143 and 144, respectively, in order of appearance.

[0051] FIG. **7**A-D illustrates an exemplary CRISPR system, an example adaptation for expression in eukaryotic cells, and results of tests assessing CRISPR activity. FIG. **7**B discloses SEQ ID NOS 145 and 146, respectively, in order of appearance. FIG. **7**C discloses SEQ ID NO: 147.

[0052] FIG. **8**A-C illustrates exemplary manipulations of a CRISPR system for targeting of genomic loci in mammalian cells. FIG. **8**A discloses SEQ ID NO: 148. FIG. **8**B discloses SEQ ID NOS 149-151, respectively, in order of appearance.

[0053] FIG. **9**A-B illustrates the results of a Northern blot analysis of crRNA processing in mammalian cells. FIG. **9**A discloses SEQ ID NO: 152.

[0054] FIG. **10**A-C illustrates a schematic representation of chimeric RNAs and results of SURVEYOR assays for CRISPR system activity in eukaryotic cells. FIG. **10**A discloses SEQ ID NO: 153

[0055] FIG. **11**A-B illustrates a graphical representation of the results of SURVEYOR assays for CRISPR system activity in eukaryotic cells.

[0056] FIG. **12** illustrates predicted secondary structures for exemplary chimeric RNAs comprising a guide sequence, tracr mate sequence, and tracr sequence. FIG. **12** discloses SEQ ID NOS 83-102, respectively, in order of appearance.

[0057] FIG. **13**A-D is a phylogenetic tree of Cas genes

[0058] FIG. **14**A-F shows the phylogenetic analysis revealing five families of Cas9s, including three groups of large Cas9s (~1400 amino acids) and two of small Cas9s (~1100 amino acids).

[0059] FIG. **15** shows a graph depicting the function of different optimized guide RNAs.

[0060] FIG. **16** shows the sequence and structure of different guide chimeric RNAs. FIG. **16** discloses SEQ ID NOS 154-165, respectively, in order of appearance.

[0061] FIG. **17** shows the co-fold structure of the tracrRNA and direct repeat. FIG. **17** discloses SEQ ID NO: 166.

[0062] FIGS. **18** A and B shows data from the St1Cas9 chimeric guide RNA optimization in vitro. FIG. **18**A discloses SEQ ID NOS 167-172, respectively, in order of appearance. FIG. **18**B discloses SEQ ID NOS 173-178, respectively, in order of appearance.

[0063] FIG. **19**A-B shows cleavage of either unmethylated or methylated targets by SpCas9 cell lysate. FIG. **19**A discloses SEQ ID NOS 179, 179, 180 and 180, respectively, in order of appearance.

[0064] FIG. **20**A-G shows the optimization of guide RNA architecture for SpCas9-mediated mammalian genome editing. (a) Schematic of bicistronic expression vector (PX330) for U6 promoter-driven single guide RNA (sgRNA) and CBh promoter-driven human codon-optimized *Streptococcus pyogenes* Cas9 (hSpCas9) used for all subsequent experiments. The sgRNA consists of a 20-nt guide sequence (blue) and scaffold (red), truncated at various positions as indicated. (b) SURVEYOR assay for SpCas9-mediated indels at the human EMX1 and PVALB loci. Arrows indicate the expected SURVEYOR fragments (n=3). (c) Northern blot analysis for the four sgRNA truncation architectures, with U1 as loading control. (d) Both wildtype (wt) or nickase mutant (D10A) of SpCas9 promoted insertion of a HindIII site into the human EMX1 gene. Single stranded oligonucleotides (ssODNs), oriented in either the sense or antisense direction relative to genome sequence, were used as homologous recombination templates. (e) Schematic of the human SERPINB5 locus. sgRNAs and PAMs are indicated by colored bars above sequence; methylcytosine (Me) are highlighted (pink) and numbered relative to the transcriptional start site (TSS, +1). (f) Methylation status of SERPINB5 assayed by bisulfite sequencing of 16 clones. Filled circles, methylated CpG; open circles, unmethylated CpG. (g) Modification efficiency by three sgRNAs targeting the methylated region of SERPINB5, assayed by deep sequencing (n=2). Error bars indicate Wilson intervals (Online Methods). FIG. **20**A discloses SEQ ID NO: 153. FIG. **20**E discloses SEQ ID NO: 181.

[0065] FIG. **21**A-B shows the further optimization of CRISPR-Cas sgRNA architecture. (a) Schematic of four additional sgRNA architectures, I-IV. Each consists of a 20-nt guide sequence (blue) joined to the direct repeat (DR, grey), which hybridizes to the tracrRNA (red). The DR-tracrRNA hybrid is truncated at +12 or +22, as indicated, with an artificial GAAA stem loop. tracrRNA truncation positions are numbered according to the previously reported transcription start site for tracrRNA. sgRNA architectures II and IV carry mutations within their poly-U tracts, which could serve as premature transcriptional terminators. (b) SURVEYOR assay for SpCas9-mediated indels at the human EMX1 locus for target sites 1-3. Arrows indicate the expected SURVEYOR fragments (n=3). FIG. **21**A discloses SEQ ID NOS 153 and 182-184, respectively, in order of appearance.

[0066] FIG. **22** illustrates visualization of some target sites in the human genome. FIG. **22** discloses SEQ ID NOS 185-263, respectively, in order of appearance.

[0067] FIG. **23**A-B shows (A) a schematic of the sgRNA and (B) the SURVEYOR analysis of five sgRNA variants for SaCas9 for an optimal truncated architecture with highest cleavage efficiency. FIG. **23**A discloses SEQ ID NO: 264.

[0068] The figures herein are for illustrative purposes only and are not necessarily drawn to scale.

DETAILED DESCRIPTION OF THE INVENTION

[0069] The terms "polynucleotide", "nucleotide", "nucleotide sequence", "nucleic acid" and "oligonucleotide" are used interchangeably. They refer to a polymeric form of nucleotides of any length, either deoxyribonucleotides or ribonucleotides, or analogs thereof. Polynucleotides may have any three dimensional structure, and may perform any function, known or unknown. The following are non limiting examples of polynucleotides: coding or non-coding regions of a gene or gene fragment, loci (locus) defined from linkage analysis, exons, introns, messenger RNA (mRNA), transfer RNA, ribosomal RNA, short interfering RNA (siRNA), short-hairpin RNA (shRNA), micro-RNA (miRNA), ribozymes, cDNA, recombinant polynucleotides, branched polynucleotides, plasmids, vectors, isolated DNA of any sequence, isolated RNA of any sequence, nucleic acid probes, and primers. A polynucleotide may comprise one or more modified nucleotides, such as methylated nucleotides and nucleotide analogs. If present, modifications to the nucleotide structure may be imparted before or after assembly of the polymer. The sequence of nucleotides may be interrupted by non nucleotide components. A polynucleotide may be further modified after polymerization, such as by conjugation with a labeling component.

[0070] In aspects of the invention the terms "chimeric RNA", "chimeric guide RNA", "guide RNA", "single guide RNA" and "synthetic guide RNA" are used interchangeably and refer to the polynucleotide sequence comprising the guide sequence, the tracr sequence and the tracr mate sequence. The term "guide sequence" refers to the about 20 bp sequence within the guide RNA that specifies the target site and may be used interchangeably with the terms "guide" or "spacer". The term "tracr mate sequence" may also be used interchangeably with the term "direct repeat(s)".

[0071] As used herein the term "wild type" is a term of the art understood by skilled persons and means the typical form of an organism, strain, gene or characteristic as it occurs in nature as distinguished from mutant or variant forms.

[0072] As used herein the term "variant" should be taken to mean the exhibition of qualities that have a pattern that deviates from what occurs in nature.

[0073] The terms "non-naturally occurring" or "engineered" are used interchangeably and indicate the involvement of the hand of man. The terms, when referring to nucleic acid molecules or polypeptides mean that the nucleic acid molecule or the polypeptide is at least substantially free from at least one other component with which they are naturally associated in nature and as found in nature.

[0074] "Complementarity" refers to the ability of a nucleic acid to form hydrogen bond(s) with another nucleic acid sequence by either traditional Watson-Crick base-pairing or other non-traditional types. A percent complementarity indicates the percentage of residues in a nucleic acid molecule

which can form hydrogen bonds (e.g., Watson-Crick base pairing) with a second nucleic acid sequence (e.g., 5, 6, 7, 8, 9, 10 out of 10 being 50%, 60%, 70%, 80%, 90%, and 100% complementary). "Perfectly complementary" means that all the contiguous residues of a nucleic acid sequence will hydrogen bond with the same number of contiguous residues in a second nucleic acid sequence. "Substantially complementary" as used herein refers to a degree of complementarity that is at least 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, 97%, 98%, 99%, or 100% over a region of 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 30, 35, 40, 45, 50, or more nucleotides, or refers to two nucleic acids that hybridize under stringent conditions.

[0075] As used herein, "stringent conditions" for hybridization refer to conditions under which a nucleic acid having complementarity to a target sequence predominantly hybridizes with the target sequence, and substantially does not hybridize to non-target sequences. Stringent conditions are generally sequence-dependent, and vary depending on a number of factors. In general, the longer the sequence, the higher the temperature at which the sequence specifically hybridizes to its target sequence. Non-limiting examples of stringent conditions are described in detail in Tijssen (1993), Laboratory Techniques In Biochemistry And Molecular Biology-Hybridization With Nucleic Acid Probes Part I, Second Chapter "Overview of principles of hybridization and the strategy of nucleic acid probe assay", Elsevier, N.Y.

[0076] "Hybridization" refers to a reaction in which one or more polynucleotides react to form a complex that is stabilized via hydrogen bonding between the bases of the nucleotide residues. The hydrogen bonding may occur by Watson Crick base pairing, Hoogstein binding, or in any other sequence specific manner. The complex may comprise two strands forming a duplex structure, three or more strands forming a multi stranded complex, a single self hybridizing strand, or any combination of these. A hybridization reaction may constitute a step in a more extensive process, such as the initiation of PCR, or the cleavage of a polynucleotide by an enzyme. A sequence capable of hybridizing with a given sequence is referred to as the "complement" of the given sequence.

[0077] As used herein, "stabilization" or "increasing stability" with respect to components of the CRISPR system relate to securing or steadying the structure of the molecule. This may be accomplished by introduction of one or mutations, including single or multiple base pair changes, increasing the number of hair pins, cross linking, breaking up particular stretches of nucleotides and other modifications. Modifications may include inclusion of at least one non naturally occurring nucleotide, or a modified nucleotide, or analogs thereof. Modified nucleotides may be modified at the ribose, phosphate, and/or base moiety. Modified nucleotides may include 2'-O-methyl analogs, 2'-deoxy analogs, or 2'-fluoro analogs. The nucleic acid backbone may be modified, for example, a phosphorothioate backbone may be used. The use of locked nucleic acids (LNA) or bridged nucleic acids (BNA) may also be possible. Further examples of modified bases include, but are not limited to, 2-aminopurine, 5-bromo-uridine, pseudouridine, inosine, 7-methylguanosine. These modifications may apply to any component of the CRISPR system. In a preferred embodiment these modifications are made to the RNA components, e.g. the guide RNA or chimeric polynucleotide sequence.

[0078] As used herein, "expression" refers to the process by which a polynucleotide is transcribed from a DNA template (such as into and mRNA or other RNA transcript) and/or the process by which a transcribed mRNA is subsequently translated into peptides, polypeptides, or proteins. Transcripts and encoded polypeptides may be collectively referred to as "gene product." If the polynucleotide is derived from genomic DNA, expression may include splicing of the mRNA in a eukaryotic cell.

[0079] The terms "polypeptide", "peptide" and "protein" are used interchangeably herein to refer to polymers of amino acids of any length. The polymer may be linear or branched, it may comprise modified amino acids, and it may be interrupted by non amino acids. The terms also encompass an amino acid polymer that has been modified; for example, disulfide bond formation, glycosylation, lipidation, acetylation, phosphorylation, or any other manipulation, such as conjugation with a labeling component. As used herein the term "amino acid" includes natural and/or unnatural or synthetic amino acids, including glycine and both the D or L optical isomers, and amino acid analogs and peptidomimetics.

[0080] The terms "subject," "individual," and "patient" are used interchangeably herein to refer to a vertebrate, preferably a mammal, more preferably a human. Mammals include, but are not limited to, murines, simians, humans, farm animals, sport animals, and pets. Tissues, cells and their progeny of a biological entity obtained in vivo or cultured in vitro are also encompassed. In some embodiments, a subject may be an invertebrate animal, for example, an insect or a nematode; while in others, a subject may be a plant or a fungus.

[0081] The terms "therapeutic agent", "therapeutic capable agent" or "treatment agent" are used interchangeably and refer to a molecule or compound that confers some beneficial effect upon administration to a subject. The beneficial effect includes enablement of diagnostic determinations; amelioration of a disease, symptom, disorder, or pathological condition; reducing or preventing the onset of a disease, symptom, disorder or condition; and generally counteracting a disease, symptom, disorder or pathological condition.

[0082] As used herein, "treatment" or "treating," or "palliating" or "ameliorating" are used interchangeably. These terms refer to an approach for obtaining beneficial or desired results including but not limited to a therapeutic benefit and/or a prophylactic benefit. By therapeutic benefit is meant any therapeutically relevant improvement in or effect on one or more diseases, conditions, or symptoms under treatment. For prophylactic benefit, the compositions may be administered to a subject at risk of developing a particular disease, condition, or symptom, or to a subject reporting one or more of the physiological symptoms of a disease, even though the disease, condition, or symptom may not have yet been manifested.

[0083] The term "effective amount" or "therapeutically effective amount" refers to the amount of an agent that is sufficient to effect beneficial or desired results. The therapeutically effective amount may vary depending upon one or more of: the subject and disease condition being treated, the weight and age of the subject, the severity of the disease condition, the manner of administration and the like, which can readily be determined by one of ordinary skill in the art. The term also applies to a dose that will provide an image

for detection by any one of the imaging methods described herein. The specific dose may vary depending on one or more of: the particular agent chosen, the dosing regimen to be followed, whether it is administered in combination with other compounds, timing of administration, the tissue to be imaged, and the physical delivery system in which it is carried.

[0084] The practice of the present invention employs, unless otherwise indicated, conventional techniques of immunology, biochemistry, chemistry, molecular biology, microbiology, cell biology, genomics and recombinant DNA, which are within the skill of the art. See Sambrook, Fritsch and Maniatis, MOLECULAR CLONING: A LABORATORY MANUAL, 2nd edition (1989); CURRENT PROTOCOLS IN MOLECULAR BIOLOGY (F. M. Ausubel, et al. eds., (1987)); the series METHODS IN ENZYMOLOGY (Academic Press, Inc.): PCR 2: A PRACTICAL APPROACH (M. J. MacPherson, B. D. Hames and G. R. Taylor eds. (1995)), Harlow and Lane, eds. (1988) ANTIBODIES, A LABORATORY MANUAL, and ANIMAL CELL CULTURE (R. I. Freshney, ed. (1987)).

[0085] Several aspects of the invention relate to vector systems comprising one or more vectors, or vectors as such. Vectors can be designed for expression of CRISPR transcripts (e.g. nucleic acid transcripts, proteins, or enzymes) in prokaryotic or eukaryotic cells. For example, CRISPR transcripts can be expressed in bacterial cells such as *Escherichia coli*, insect cells (using baculovirus expression vectors), yeast cells, or mammalian cells. Suitable host cells are discussed further in Goeddel, GENE EXPRESSION TECHNOLOGY: METHODS IN ENZYMOLOGY 185, Academic Press, San Diego, Calif. (1990). Alternatively, the recombinant expression vector can be transcribed and translated in vitro, for example using T7 promoter regulatory sequences and T7 polymerase.

[0086] Vectors may be introduced and propagated in a prokaryote. In some embodiments, a prokaryote is used to amplify copies of a vector to be introduced into a eukaryotic cell or as an intermediate vector in the production of a vector to be introduced into a eukaryotic cell (e.g. amplifying a plasmid as part of a viral vector packaging system). In some embodiments, a prokaryote is used to amplify copies of a vector and express one or more nucleic acids, such as to provide a source of one or more proteins for delivery to a host cell or host organism. Expression of proteins in prokaryotes is most often carried out in *Escherichia coli* with vectors containing constitutive or inducible promoters directing the expression of either fusion or non-fusion proteins. Fusion vectors add a number of amino acids to a protein encoded therein, such as to the amino terminus of the recombinant protein. Such fusion vectors may serve one or more purposes, such as: (i) to increase expression of recombinant protein; (ii) to increase the solubility of the recombinant protein; and (iii) to aid in the purification of the recombinant protein by acting as a ligand in affinity purification. Often, in fusion expression vectors, a proteolytic cleavage site is introduced at the junction of the fusion moiety and the recombinant protein to enable separation of the recombinant protein from the fusion moiety subsequent to purification of the fusion protein. Such enzymes, and their cognate recognition sequences, include Factor Xa, thrombin and enterokinase. Example fusion expression vectors include pGEX (Pharmacia Biotech Inc; Smith and Johnson, 1988. *Gene* 67: 31-40), pMAL (New England Biolabs, Beverly, Mass.) and pRIT5 (Pharmacia, Piscataway, N.J.) that fuse glutathione S-transferase (GST), maltose E binding protein, or protein A, respectively, to the target recombinant protein.

[0087] Examples of suitable inducible non-fusion *E. coli* expression vectors include pTrc (Amrann et al., (1988) *Gene* 69:301-315) and pET lid (Studier et al., GENE EXPRESSION TECHNOLOGY: METHODS IN ENZYMOLOGY 185, Academic Press, San Diego, Calif. (1990) 60-89).

[0088] In some embodiments, a vector is a yeast expression vector. Examples of vectors for expression in yeast *Saccharomyces cerivisae* include pYepSec1 (Baldari, et al., 1987. *EMBO J.* 6: 229-234), pMFa (Kuijan and Herskowitz, 1982. Cell 30: 933-943), pJRY88 (Schultz et al., 1987. Gene 54: 113-123), pYES2 (Invitrogen Corporation, San Diego, Calif.), and picZ (InVitrogen Corp, San Diego, Calif.).

[0089] In some embodiments, a vector drives protein expression in insect cells using baculovirus expression vectors. Baculovirus vectors available for expression of proteins in cultured insect cells (e.g., SF9 cells) include the pAc series (Smith, et al., 1983. *Mol. Cell. Biol.* 3: 2156-2165) and the pVL series (Lucklow and Summers, 1989. *Virology* 170: 31-39).

[0090] In some embodiments, a vector is capable of driving expression of one or more sequences in mammalian cells using a mammalian expression vector. Examples of mammalian expression vectors include pCDM8 (Seed, 1987. *Nature* 329: 840) and pMT2PC (Kaufman, et al., 1987. *EMBO J.* 6: 187-195). When used in mammalian cells, the expression vector's control functions are typically provided by one or more regulatory elements. For example, commonly used promoters are derived from polyoma, adenovirus 2, cytomegalovirus, simian virus 40, and others disclosed herein and known in the art. For other suitable expression systems for both prokaryotic and eukaryotic cells see, e.g., Chapters 16 and 17 of Sambrook, et al., MOLECULAR CLONING: A LABORATORY MANUAL. 2nd ed., Cold Spring Harbor Laboratory, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y., 1989.

[0091] In some embodiments, the recombinant mammalian expression vector is capable of directing expression of the nucleic acid preferentially in a particular cell type (e.g., tissue-specific regulatory elements are used to express the nucleic acid). Tissue-specific regulatory elements are known in the art. Non-limiting examples of suitable tissue-specific promoters include the albumin promoter (liver-specific; Pinkert, et al., 1987. *Genes Dev.* 1: 268-277), lymphoid-specific promoters (Calame and Eaton, 1988. *Adv. Immunol.* 43: 235-275), in particular promoters of T cell receptors (Winoto and Baltimore, 1989. EMBO J. 8: 729-733) and immunoglobulins (Baneiji, et al., 1983. *Cell* 33: 729-740; Queen and Baltimore, 1983. Cell 33: 741-748), neuron-specific promoters (e.g., the neurofilament promoter; Byrne and Ruddle, 1989. *Proc. Natl. Acad. Sci. USA* 86: 5473-5477), pancreas-specific promoters (Edlund, et al., 1985. *Science* 230: 912-916), and mammary gland-specific promoters (e.g., milk whey promoter; U.S. Pat. No. 4,873,316 and European Application Publication No. 264,166). Developmentally-regulated promoters are also encompassed, e.g., the murine hox promoters (Kessel and Gruss, 1990. *Science* 249: 374-379) and the α-fetoprotein promoter (Campes and Tilghman, 1989. *Genes Dev.* 3: 537-546).

[0092] In some embodiments, a regulatory element is operably linked to one or more elements of a CRISPR

system so as to drive expression of the one or more elements of the CRISPR system. In general, CRISPRs (Clustered Regularly Interspaced Short Palindromic Repeats), also known as SPIDRs (SPacer Interspersed Direct Repeats), constitute a family of DNA loci that are usually specific to a particular bacterial species. The CRISPR locus comprises a distinct class of interspersed short sequence repeats (SSRs) that were recognized in *E. coli* (Ishino et al., J. Bacteriol., 169:5429-5433 [1987]; and Nakata et al., J. Bacteriol., 171:3553-3556 [1989]), and associated genes. Similar interspersed SSRs have been identified in *Haloferax mediterranei, Streptococcus pyogenes, Anabaena*, and *Mycobacterium tuberculosis* (See, Groenen et al., Mol. Microbiol., 10:1057-1065 [1993]; Hoe et al., Emerg. Infect. Dis., 5:254-263 [1999]; Masepohl et al., Biochim. Biophys. Acta 1307: 26-30 [1996]; and Mojica et al., Mol. Microbiol., 17:85-93 [1995]). The CRISPR loci typically differ from other SSRs by the structure of the repeats, which have been termed short regularly spaced repeats (SRSRs) (Janssen et al., OMICS J. Integ. Biol., 6:23-33 [2002]; and Mojica et al., Mol. Microbiol., 36:244-246 [2000]). In general, the repeats are short elements that occur in clusters that are regularly spaced by unique intervening sequences with a substantially constant length (Mojica et al., [2000], supra). Although the repeat sequences are highly conserved between strains, the number of interspersed repeats and the sequences of the spacer regions typically differ from strain to strain (van Embden et al., J. Bacteriol., 182:2393-2401 [2000]). CRISPR loci have been identified in more than 40 prokaryotes (See e.g., Jansen et al., Mol. Microbiol., 43:1565-1575 [2002]; and Mojica et al., [2005]) including, but not limited to Aeropyrum, *Pyrobaculum, Sulfolobus, Archaeoglobus, Halocarcula, Methanobacterium, Methanococcus, Methanosarcina, Methanopyrus, Pyrococcus, Picrophilus, Thermoplasma, Corynebacterium, Mycobacterium, Streptomyces, Aquifex, Porphyromonas, Chlorobium, Thermus, Bacillus, Listeria, Staphylococcus, Clostridium, Thermoanaerobacter, Mycoplasma, Fusobacterium, Azarcus, Chromobacterium, Neisseria, Nitrosomonas, Desulfovibrio, Geobacter, Myxococcus, Campylobacter, Wolinella, Acinetobacter, Erwinia, Escherichia, Legionella, Methylococcus, Pasteurella, Photobacterium, Salmonella, Xanthomonas, Yersinia, Treponema*, and *Thermotoga.*

[0093] In general, "CRISPR system" refers collectively to transcripts and other elements involved in the expression of or directing the activity of CRISPR-associated ("Cas") genes, including sequences encoding a Cas gene, a tracr (trans-activating CRISPR) sequence (e.g. tracrRNA or an active partial tracrRNA), a tracr-mate sequence (encompassing a "direct repeat" and a tracrRNA-processed partial direct repeat in the context of an endogenous CRISPR system), a guide sequence (also referred to as a "spacer" in the context of an endogenous CRISPR system), or other sequences and transcripts from a CRISPR locus. In some embodiments, one or more elements of a CRISPR system is derived from a type I, type II, or type III CRISPR system. In some embodiments, one or more elements of a CRISPR system is derived from a particular organism comprising an endogenous CRISPR system, such as *Streptococcus pyogenes*. In general, a CRISPR system is characterized by elements that promote the formation of a CRISPR complex at the site of a target sequence (also referred to as a protospacer in the context of an endogenous CRISPR system). In the context of formation of a CRISPR complex, "target sequence" refers to a sequence to which a guide sequence is designed to have complementarity, where hybridization between a target sequence and a guide sequence promotes the formation of a CRISPR complex. Full complementarity is not necessarily required, provided there is sufficient complementarity to cause hybridisation and promote formation of a CRISPR complex. A target sequence may comprise any polynucleotide, such as DNA or RNA polynucleotides. In some embodiments, a target sequence is located in the nucleus or cytoplasm of a cell. In some embodiments, the target sequence may be within an organelle of a eukaryotic cell, for example, mitochondrion or chloroplast. A sequence or template that may be used for recombination into the the targeted locus comprising the target sequences is referred to as an "editing template" or "editing polynucleotide" or "editing sequence". In aspects of the invention, an exogenous template polynucleotide may be referred to as an editing template. In an aspect of the invention the recombination is homologous recombination.

[0094] Typically, in the context of an endogenous CRISPR system, formation of a CRISPR complex (comprising a guide sequence hybridized to a target sequence and complexed with one or more Cas proteins) results in cleavage of one or both strands in or near (e.g. within 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 50, or more base pairs from) the target sequence. Without wishing to be bound by theory, the tracr sequence, which may comprise or consist of all or a portion of a wild-type tracr sequence (e.g. about or more than about 20, 26, 32, 45, 48, 54, 63, 67, 85, or more nucleotides of a wild-type tracr sequence), may also form part of a CRISPR complex, such as by hybridization along at least a portion of the tracr sequence to all or a portion of a tracr mate sequence that is operably linked to the guide sequence. In some embodiments, the tracr sequence has sufficient complementarity to a tracr mate sequence to hybridise and participate in formation of a CRISPR complex. As with the target sequence, it is believed that complete complementarity is not needed, provided there is sufficient to be functional. In some embodiments, the tracr sequence has at least 50%, 60%, 70%, 80%, 90%, 95% or 99% of sequence complementarity along the length of the tracr mate sequence when optimally aligned. In some embodiments, one or more vectors driving expression of one or more elements of a CRISPR system are introduced into a host cell such that expression of the elements of the CRISPR system direct formation of a CRISPR complex at one or more target sites. For example, a Cas enzyme, a guide sequence linked to a tracr-mate sequence, and a tracr sequence could each be operably linked to separate regulatory elements on separate vectors. Alternatively, two or more of the elements expressed from the same or different regulatory elements, may be combined in a single vector, with one or more additional vectors providing any components of the CRISPR system not included in the first vector. CRISPR system elements that are combined in a single vector may be arranged in any suitable orientation, such as one element located 5' with respect to ("upstream" of) or 3' with respect to ("downstream" of) a second element. The coding sequence of one element may be located on the same or opposite strand of the coding sequence of a second element, and oriented in the same or opposite direction. In some embodiments, a single promoter drives expression of a transcript encoding a CRISPR enzyme and one or more of the guide sequence, tracr mate sequence (optionally oper-

ably linked to the guide sequence), and a tracr sequence embedded within one or more intron sequences (e.g. each in a different intron, two or more in at least one intron, or all in a single intron). In some embodiments, the CRISPR enzyme, guide sequence, tracr mate sequence, and tracr sequence are operably linked to and expressed from the same promoter.

[0095] In some embodiments, a vector comprises one or more insertion sites, such as a restriction endonuclease recognition sequence (also referred to as a "cloning site"). In some embodiments, one or more insertion sites (e.g. about or more than about 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, or more insertion sites) are located upstream and/or downstream of one or more sequence elements of one or more vectors. In some embodiments, a vector comprises an insertion site upstream of a tracr mate sequence, and optionally downstream of a regulatory element operably linked to the tracr mate sequence, such that following insertion of a guide sequence into the insertion site and upon expression the guide sequence directs sequence-specific binding of a CRISPR complex to a target sequence in a eukaryotic cell. In some embodiments, a vector comprises two or more insertion sites, each insertion site being located between two tracr mate sequences so as to allow insertion of a guide sequence at each site. In such an arrangement, the two or more guide sequences may comprise two or more copies of a single guide sequence, two or more different guide sequences, or combinations of these. When multiple different guide sequences are used, a single expression construct may be used to target CRISPR activity to multiple different, corresponding target sequences within a cell. For example, a single vector may comprise about or more than about 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, or more guide sequences. In some embodiments, about or more than about 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, or more such guide-sequence-containing vectors may be provided, and optionally delivered to a cell.

[0096] In some embodiments, a vector comprises a regulatory element operably linked to an enzyme-coding sequence encoding a CRISPR enzyme, such as a Cas protein. Non-limiting examples of Cas proteins include Cas1, Cas1B, Cas2, Cas3, Cas4, Cas5, Cas6, Cas7, Cas8, Cas9 (also known as Csn1 and Csx12), Cas10, Csy1, Csy2, Csy3, Cse1, Cse2, Csc1, Csc2, Csa5, Csn2, Csm2, Csm3, Csm4, Csm5, Csm6, Cmr1, Cmr3, Cmr4, Cmr5, Cmr6, Csb1, Csb2, Csb3, Csx17, Csx14, Csx10, Csx16, CsaX, Csx3, Csx1, Csx15, Csf1, Csf2, Csf3, Csf4, homologs thereof, or modified versions thereof. These enzymes are known; for example, the amino acid sequence of *S. pyogenes* Cas9 protein may be found in the SwissProt database under accession number Q99ZW2. In some embodiments, the unmodified CRISPR enzyme has DNA cleavage activity, such as Cas9. In some embodiments the CRISPR enzyme is Cas9, and may be Cas9 from *S. pyogenes* or *S. pneumoniae*. In some embodiments, the CRISPR enzyme directs cleavage of one or both strands at the location of a target sequence, such as within the target sequence and/or within the complement of the target sequence. In some embodiments, the CRISPR enzyme directs cleavage of one or both strands within about 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 50, 100, 200, 500, or more base pairs from the first or last nucleotide of a target sequence. In some embodiments, a vector encodes a CRISPR enzyme that is mutated to with respect to a corresponding wild-type enzyme such that the mutated CRISPR enzyme lacks the ability to cleave one or both strands of a target polynucleotide containing a target sequence. For example, an aspartate-to-alanine substitution (D10A) in the RuvC I catalytic domain of Cas9 from *S. pyogenes* converts Cas9 from a nuclease that cleaves both strands to a nickase (cleaves a single strand). Other examples of mutations that render Cas9 a nickase include, without limitation, H840A, N854A, and N863A. In some embodiments, a Cas9 nickase may be used in combination with guide sequenc(es), e.g., two guide sequences, which target respectively sense and antisense strands of the DNA target. This combination allows both strands to be nicked and used to induce NHEJ. Applicants have demonstrated (data not shown) the efficacy of two nickase targets (i.e., sgRNAs targeted at the same location but to different strands of DNA) in inducing mutagenic NHEJ. A single nickase (Cas9-D10A with a single sgRNA) is unable to induce NHEJ and create indels but Applicants have shown that double nickase (Cas9-D10A and two sgRNAs targeted to different strands at the same location) can do so in human embryonic stem cells (hESCs). The efficiency is about 50% of nuclease (i.e., regular Cas9 without D10 mutation) in hESCs.

[0097] As a further example, two or more catalytic domains of Cas9 (RuvC I, RuvC II, and RuvC III) may be mutated to produce a mutated Cas9 substantially lacking all DNA cleavage activity. In some embodiments, a D10A mutation is combined with one or more of H840A, N854A, or N863A mutations to produce a Cas9 enzyme substantially lacking all DNA cleavage activity. In some embodiments, a CRISPR enzyme is considered to substantially lack all DNA cleavage activity when the DNA cleavage activity of the mutated enzyme is less than about 25%, 10%, 5%, 1%, 0.1%, 0.01%, or lower with respect to its non-mutated form. Other mutations may be useful; where the Cas9 or other CRISPR enzyme is from a species other than *S. pyogenes*, mutations in corresponding amino acids may be made to achieve similar effects.

[0098] In some embodiments, an enzyme coding sequence encoding a CRISPR enzyme is codon optimized for expression in particular cells, such as eukaryotic cells. The eukaryotic cells may be those of or derived from a particular organism, such as a mammal, including but not limited to human, mouse, rat, rabbit, dog, or non-human primate. In general, codon optimization refers to a process of modifying a nucleic acid sequence for enhanced expression in the host cells of interest by replacing at least one codon (e.g. about or more than about 1, 2, 3, 4, 5, 10, 15, 20, 25, 50, or more codons) of the native sequence with codons that are more frequently or most frequently used in the genes of that host cell while maintaining the native amino acid sequence. Various species exhibit particular bias for certain codons of a particular amino acid. Codon bias (differences in codon usage between organisms) often correlates with the efficiency of translation of messenger RNA (mRNA), which is in turn believed to be dependent on, among other things, the properties of the codons being translated and the availability of particular transfer RNA (tRNA) molecules. The predominance of selected tRNAs in a cell is generally a reflection of the codons used most frequently in peptide synthesis. Accordingly, genes can be tailored for optimal gene expression in a given organism based on codon optimization. Codon usage tables are readily available, for example, at the "Codon Usage Database", and these tables can be adapted in a number of ways. See Nakamura, Y., et al. "Codon usage

tabulated from the international DNA sequence databases: status for the year 2000" Nucl. Acids Res. 28:292 (2000). Computer algorithms for codon optimizing a particular sequence for expression in a particular host cell are also available, such as Gene Forge (Aptagen; Jacobus, PA), are also available. In some embodiments, one or more codons (e.g. 1, 2, 3, 4, 5, 10, 15, 20, 25, 50, or more, or all codons) in a sequence encoding a CRISPR enzyme correspond to the most frequently used codon for a particular amino acid.

[0099] In some embodiments, a vector encodes a CRISPR enzyme comprising one or more nuclear localization sequences (NLSs), such as about or more than about 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, or more NLSs. In some embodiments, the CRISPR enzyme comprises about or more than about 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, or more NLSs at or near the amino-terminus, about or more than about 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, or more NLSs at or near the carboxy-terminus, or a combination of these (e.g. one or more NLS at the amino-terminus and one or more NLS at the carboxy terminus). When more than one NLS is present, each may be selected independently of the others, such that a single NLS may be present in more than one copy and/or in combination with one or more other NLSs present in one or more copies. In a preferred embodiment of the invention, the CRISPR enzyme comprises at most 6 NLSs. In some embodiments, an NLS is considered near the N- or C-terminus when the nearest amino acid of the NLS is within about 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 40, 50, or more amino acids along the polypeptide chain from the N- or C-terminus. Typically, an NLS consists of one or more short sequences of positively charged lysines or arginines exposed on the protein surface, but other types of NLS are known. Non-limiting examples of NLSs include an NLS sequence derived from: the NLS of the SV40 virus large T-antigen, having the amino acid sequence PKKKRKV (SEQ ID NO: 2); the NLS from nucleoplasmin (e.g. the nucleoplasmin bipartite NLS with the sequence KRPAATKKAGQAKKKK (SEQ ID NO: 3)); the c-myc NLS having the amino acid sequence PAAKRVKLD (SEQ ID NO: 4) or RQRRNELKRSP (SEQ ID NO: 5); the hRNPA1 M9 NLS having the sequence NQSSNFGPMKGGNFGGRSSGPYGGGGQYFAK-PRNQGGY (SEQ ID NO: 6); the sequence RMRIZFKNKGKDTAELRRRRVEVSVELRKAKKD-EQILKRRNV (SEQ ID NO: 7) of the IBB domain from importin-alpha; the sequences VSRKRPRP (SEQ ID NO: 8) and PPKKARED (SEQ ID NO: 9) of the myoma T protein; the sequence PQPKKKPL (SEQ ID NO: 10) of human p53; the sequence SALIKKKKKMAP (SEQ ID NO: 11) of mouse c-abl IV; the sequences DRLRR (SEQ ID NO: 12) and PKQKKRK (SEQ ID NO: 13) of the influenza virus NS1; the sequence RKLKKKIKKL (SEQ ID NO: 14) of the Hepatitis virus delta antigen; the sequence REKKKFLKRR (SEQ ID NO: 15) of the mouse Mx1 protein; the sequence KRKGDEVDGVDEVAKKKSKK (SEQ ID NO: 16) of the human poly(ADP-ribose) polymerase; and the sequence RKCLQAGMNLEARKTKK (SEQ ID NO: 17) of the steroid hormone receptors (human) glucocorticoid.

[0100] In general, the one or more NLSs are of sufficient strength to drive accumulation of the CRISPR enzyme in a detectable amount in the nucleus of a eukaryotic cell. In general, strength of nuclear localization activity may derive from the number of NLSs in the CRISPR enzyme, the particular NLS(s) used, or a combination of these factors. Detection of accumulation in the nucleus may be performed by any suitable technique. For example, a detectable marker may be fused to the CRISPR enzyme, such that location within a cell may be visualized, such as in combination with a means for detecting the location of the nucleus (e.g. a stain specific for the nucleus such as DAPI). Examples of detectable markers include fluorescent proteins (such as Green fluorescent proteins, or GFP; RFP; CFP), and epitope tags (HA tag, flag tag, SNAP tag). Cell nuclei may also be isolated from cells, the contents of which may then be analyzed by any suitable process for detecting protein, such as immunohistochemistry, Western blot, or enzyme activity assay. Accumulation in the nucleus may also be determined indirectly, such as by an assay for the effect of CRISPR complex formation (e.g. assay for DNA cleavage or mutation at the target sequence, or assay for altered gene expression activity affected by CRISPR complex formation and/or CRISPR enzyme activity), as compared to a control no exposed to the CRISPR enzyme or complex, or exposed to a CRISPR enzyme lacking the one or more NLSs.

[0101] In general, a guide sequence is any polynucleotide sequence having sufficient complementarity with a target polynucleotide sequence to hybridize with the target sequence and direct sequence-specific binding of a CRISPR complex to the target sequence. In some embodiments, the degree of complementarity between a guide sequence and its corresponding target sequence, when optimally aligned using a suitable alignment algorithm, is about or more than about 50%, 60%, 75%, 80%, 85%, 90%, 95%, 97.5%, 99%, or more. Optimal alignment may be determined with the use of any suitable algorithm for aligning sequences, non-limiting example of which include the Smith-Waterman algorithm, the Needleman-Wunsch algorithm, algorithms based on the Burrows-Wheeler Transform (e.g. the Burrows Wheeler Aligner), ClustalW, Clustal X, BLAT, Novoalign (Novocraft Technologies, ELAND (Illumina, San Diego, CA), SOAP (available at soap.genomics.org.cn), and Maq (available at maq.sourceforge.net). In some embodiments, a guide sequence is about or more than about 5, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 35, 40, 45, 50, 75, or more nucleotides in length. In some embodiments, a guide sequence is less than about 75, 50, 45, 40, 35, 30, 25, 20, 15, 12, or fewer nucleotides in length. The ability of a guide sequence to direct sequence-specific binding of a CRISPR complex to a target sequence may be assessed by any suitable assay. For example, the components of a CRISPR system sufficient to form a CRISPR complex, including the guide sequence to be tested, may be provided to a host cell having the corresponding target sequence, such as by transfection with vectors encoding the components of the CRISPR sequence, followed by an assessment of preferential cleavage within the target sequence, such as by Surveyor assay as described herein. Similarly, cleavage of a target polynucleotide sequence may be evaluated in a test tube by providing the target sequence, components of a CRISPR complex, including the guide sequence to be tested and a control guide sequence different from the test guide sequence, and comparing binding or rate of cleavage at the target sequence between the test and control guide sequence reactions. Other assays are possible, and will occur to those skilled in the art.

[0102] A guide sequence may be selected to target any target sequence. In some embodiments, the target sequence is a sequence within a genome of a cell. Exemplary target sequences include those that are unique in the target

genome. For example, for the *S. pyogenes* Cas9, a unique target sequence in a genome may include a Cas9 target site of the form nnnnnnnnNNNNNNNNNNNNXGG (SEQ ID NO: 265) where NNNNNNNNNNNNXGG (SEQ ID NO: 266) (N is A, G, T, or C; and X can be anything) has a single occurrence in the genome. A unique target sequence in a genome may include an *S. pyogenes* Cas9 target site of the form nnnnnnnnnNNNNNNNNNNNXGG (SEQ ID NO: 267) where NNNNNNNNNNNXGG (SEQ ID NO: 268) (N is A, G, T, or C; and X can be anything) has a single occurrence in the genome. For the *S. thermophilus* CRISPR1 Cas9, a unique target sequence in a genome may include a Cas9 target site of the form nnnnnnnnNNNNNNNNNNNNXXAGAAW (SEQ ID NO: 18) where NNNNNNNNNNNNXXAGAAW (SEQ ID NO: 19) (N is A, G, T, or C; X can be anything; and W is A or T) has a single occurrence in the genome. A unique target sequence in a genome may include an *S. thermophilus* CRISPR1 Cas9 target site of the form nnnnnnnnnNNNNNNNNNNNXXAGAAW (SEQ ID NO: 20) where NNNNNNNNNNNXXAGAAW (SEQ ID NO: 21) (N is A, G, T, or C; X can be anything; and W is A or T) has a single occurrence in the genome. For the *S. pyogenes* Cas9, a unique target sequence in a genome may include a Cas9 target site of the form nnnnnnnnNNNNNNNNNNNNXGGXG (SEQ ID NO: 269) where NNNNNNNNNNNNXGGXG (SEQ ID NO: 270) (N is A, G, T, or C; and X can be anything) has a single occurrence in the genome. A unique target sequence in a genome may include an *S. pyogenes* Cas9 target site of the form nnnnnnnnnNNNNNNNNNNNXGGXG (SEQ ID NO: 271) where NNNNNNNNNNNXGGXG (SEQ ID NO: 272) (N is A, G, T, or C; and X can be anything) has a single occurrence in the genome. In each of these sequences "n" may be A, G, T, or C, and need not be considered in identifying a sequence as unique.

[0103] In some embodiments, a guide sequence is selected to reduce the degree of secondary structure within the guide sequence. Secondary structure may be determined by any suitable polynucleotide folding algorithm. Some programs are based on calculating the minimal Gibbs free energy. An example of one such algorithm is mFold, as described by Zuker and Stiegler (Nucleic Acids Res. 9 (1981), 133-148). Another example folding algorithm is the online webserver RNAfold, developed at Institute for Theoretical Chemistry at the University of Vienna, using the centroid structure prediction algorithm (see e.g. A. R. Gruber et al., 2008, *Cell* 106(1): 23-24; and PA Carr and GM Church, 2009, *Nature Biotechnology* 27(12): 1151-62). Further algorithms may be found in U.S. application Ser. No. _____ (Broad Reference BI-2012/084 44790.11.2022); incorporated herein by reference.

[0104] In general, a tracr mate sequence includes any sequence that has sufficient complementarity with a tracr sequence to promote one or more of: (1) excision of a guide sequence flanked by tracr mate sequences in a cell containing the corresponding tracr sequence; and (2) formation of a CRISPR complex at a target sequence, wherein the CRISPR complex comprises the tracr mate sequence hybridized to the tracr sequence. In general, degree of complementarity is with reference to the optimal alignment of the tracr mate sequence and tracr sequence, along the length of the shorter of the two sequences. Optimal alignment may be determined by any suitable alignment algorithm, and may further account for secondary structures, such as self-complementarity within either the tracr sequence or tracr mate sequence. In some embodiments, the degree of complementarity between the tracr sequence and tracr mate sequence along the length of the shorter of the two when optimally aligned is about or more than about 25%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95%, 97.5%, 99%, or higher. Example illustrations of optimal alignment between a tracr sequence and a tracr mate sequence are provided in FIGS. **12**B and **13**B. In some embodiments, the tracr sequence is about or more than about 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 25, 30, 40, 50, or more nucleotides in length. In some embodiments, the tracr sequence and tracr mate sequence are contained within a single transcript, such that hybridization between the two produces a transcript having a secondary structure, such as a hairpin. Preferred loop forming sequences for use in hairpin structures are four nucleotides in length, and most preferably have the sequence GAAA. However, longer or shorter loop sequences may be used, as may alternative sequences. The sequences preferably include a nucleotide triplet (for example, AAA), and an additional nucleotide (for example C or G). Examples of loop forming sequences include CAAA and AAAG. In an embodiment of the invention, the transcript or transcribed polynucleotide sequence has at least two or more hairpins. In preferred embodiments, the transcript has two, three, four or five hairpins. In a further embodiment of the invention, the transcript has at most five hairpins. In some embodiments, the single transcript further includes a transcription termination sequence; preferably this is a polyT sequence, for example six T nucleotides. An example illustration of such a hairpin structure is provided in the lower portion of FIG. **13**B, where the portion of the sequence 5' of the final "N" and upstream of the loop corresponds to the tracr mate sequence, and the portion of the sequence 3' of the loop corresponds to the tracr sequence. Further non-limiting examples of single polynucleotides comprising a guide sequence, a tracr mate sequence, and a tracr sequence are as follows (listed 5' to 3'), where "N" represents a base of a guide sequence, the first block of lower case letters represent the tracr mate sequence, and the second block of lower case letters represent the tracr sequence, and the final poly-T sequence represents the transcription terminator: (1) NNNNNNNNNNNNNNNNNNNgtttttgtactctcaagatttaGAAAtaaatcttgcagaagctacaaagataaggctt catgccgaaatcaacaccctgtcattttatggcagggtgttttcgttatttaaTTTTTT (SEQ ID NO: 22); (2) NNNNNNNNNNNNNNNNNNNgtttttgtactctcaGAAAtgcagaagctacaaagataaggcttcatgccgaaatca acaccctgtcattttatggcagggtgttttcgttattaaTTTTTT (SEQ ID NO: 23); (3) NNNNNNNNNNNNNNNNNNNgtttttgtactctcaGAAAtgcagaagctacaaagataaggcttcatgccgaaatca acaccctgtcattttatggcagggtgtTTTTTT (SEQ ID NO: 24); (4) NNNNNNNNNNNNNNNNNNNgtttttagagctaGAAAtagcaagttaaaataaggctagtccgttatcaacttgaaaaagtggcaccgagtcggtgcTTTTTT (SEQ ID NO: 25); (5) NNNNNNNNNNNNNNNNNNNgtttttagagctaGAAATAGcaagttaaaataaggctagtccgttatcaacttgaaaaagtgTTTTTTT (SEQ ID NO: 26); and (6) NNNNNNNNNNNNNNNNNNNgtttttagagctagAAATAGcaagttaaaataaggctagtccgttatcaTTTTTTTT (SEQ ID NO: 27). In some embodiments, sequences (1) to (3) are used in combination with Cas9 from *S. thermophilus* CRISPR1. In some embodiments, sequences (4) to (6)

are used in combination with Cas9 from *S. pyogenes*. In some embodiments, the tracr sequence is a separate transcript from a transcript comprising the tracr mate sequence (such as illustrated in the top portion of FIG. **13**B).

[0105] In some embodiments, a recombination template is also provided. A recombination template may be a component of another vector as described herein, contained in a separate vector, or provided as a separate polynucleotide. In some embodiments, a recombination template is designed to serve as a template in homologous recombination, such as within or near a target sequence nicked or cleaved by a CRISPR enzyme as a part of a CRISPR complex. A template polynucleotide may be of any suitable length, such as about or more than about 10, 15, 20, 25, 50, 75, 100, 150, 200, 500, 1000, or more nucleotides in length. In some embodiments, the template polynucleotide is complementary to a portion of a polynucleotide comprising the target sequence. When optimally aligned, a template polynucleotide might overlap with one or more nucleotides of a target sequences (e.g. about or more than about 1, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100 or more nucleotides). In some embodiments, when a template sequence and a polynucleotide comprising a target sequence are optimally aligned, the nearest nucleotide of the template polynucleotide is within about 1, 5, 10, 15, 20, 25, 50, 75, 100, 200, 300, 400, 500, 1000, 5000, 10000, or more nucleotides from the target sequence.

[0106] In some embodiments, the CRISPR enzyme is part of a fusion protein comprising one or more heterologous protein domains (e.g. about or more than about 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, or more domains in addition to the CRISPR enzyme). A CRISPR enzyme fusion protein may comprise any additional protein sequence, and optionally a linker sequence between any two domains. Examples of protein domains that may be fused to a CRISPR enzyme include, without limitation, epitope tags, reporter gene sequences, and protein domains having one or more of the following activities: methylase activity, demethylase activity, transcription activation activity, transcription repression activity, transcription release factor activity, histone modification activity, RNA cleavage activity and nucleic acid binding activity. Non-limiting examples of epitope tags include histidine (His) tags, V5 tags, FLAG tags, influenza hemagglutinin (HA) tags, Myc tags, VSV-G tags, and thioredoxin (Trx) tags. Examples of reporter genes include, but are not limited to, glutathione-S-transferase (GST), horseradish peroxidase (HRP), chloramphenicol acetyltransferase (CAT) beta-galactosidase, beta-glucuronidase, luciferase, green fluorescent protein (GFP), HcRed, DsRed, cyan fluorescent protein (CFP), yellow fluorescent protein (YFP), and autofluorescent proteins including blue fluorescent protein (BFP). A CRISPR enzyme may be fused to a gene sequence encoding a protein or a fragment of a protein that bind DNA molecules or bind other cellular molecules, including but not limited to maltose binding protein (MBP), S-tag, Lex A DNA binding domain (DBD) fusions, GAL4 DNA binding domain fusions, and herpes simplex virus (HSV) BP16 protein fusions. Additional domains that may form part of a fusion protein comprising a CRISPR enzyme are described in US20110059502, incorporated herein by reference. In some embodiments, a tagged CRISPR enzyme is used to identify the location of a target sequence.

[0107] In some aspects, the invention provides methods comprising delivering one or more polynucleotides, such as or one or more vectors as described herein, one or more transcripts thereof, and/or one or proteins transcribed therefrom, to a host cell. In some aspects, the invention further provides cells produced by such methods, and organisms (such as animals, plants, or fungi) comprising or produced from such cells. In some embodiments, a CRISPR enzyme in combination with (and optionally complexed with) a guide sequence is delivered to a cell. Conventional viral and non-viral based gene transfer methods can be used to introduce nucleic acids in mammalian cells or target tissues. Such methods can be used to administer nucleic acids encoding components of a CRISPR system to cells in culture, or in a host organism. Non-viral vector delivery systems include DNA plasmids, RNA (e.g. a transcript of a vector described herein), naked nucleic acid, and nucleic acid complexed with a delivery vehicle, such as a liposome. Viral vector delivery systems include DNA and RNA viruses, which have either episomal or integrated genomes after delivery to the cell. For a review of gene therapy procedures, see Anderson, Science 256:808-813 (1992); Nabel & Felgner, TIBTECH 11:211-217 (1993); Mitani & Caskey, TIBTECH 11:162-166 (1993); Dillon, TIBTECH 11:167-175 (1993); Miller, Nature 357:455-460 (1992); Van Brunt, Biotechnology 6(10):1149-1154 (1988); Vigne, Restorative Neurology and Neuroscience 8:35-36 (1995); Kremer & Perricaudet, British Medical Bulletin 51(1):31-44 (1995); Haddada et al., in Current Topics in Microbiology and Immunology Doerfler and Bohm (eds) (1995); and Yu et al., Gene Therapy 1:13-26 (1994).

[0108] Methods of non-viral delivery of nucleic acids include lipofection, nucleofection, microinjection, biolistics, virosomes, liposomes, immunoliposomes, polycation or lipid:nucleic acid conjugates, naked DNA, artificial virions, and agent-enhanced uptake of DNA. Lipofection is described in e.g., U.S. Pat. Nos. 5,049,386, 4,946,787; and 4,897,355) and lipofection reagents are sold commercially (e.g., Transfectam™ and Lipofectin™). Cationic and neutral lipids that are suitable for efficient receptor-recognition lipofection of polynucleotides include those of Felgner, WO 91/17424; WO 91/16024. Delivery can be to cells (e.g. in vitro or ex vivo administration) or target tissues (e.g. in vivo administration).

[0109] The preparation of lipid:nucleic acid complexes, including targeted liposomes such as immunolipid complexes, is well known to one of skill in the art (see, e.g., Crystal, Science 270:404-410 (1995); Blaese et al., Cancer Gene Ther. 2:291-297 (1995); Behr et al., Bioconjugate Chem. 5:382-389 (1994); Remy et al., Bioconjugate Chem. 5:647-654 (1994); Gao et al., Gene Therapy 2:710-722 (1995); Ahmad et al., Cancer Res. 52:4817-4820 (1992); U.S. Pat. Nos. 4,186,183, 4,217,344, 4,235,871, 4,261,975, 4,485,054, 4,501,728, 4,774,085, 4,837,028, and 4,946,787).

[0110] The use of RNA or DNA viral based systems for the delivery of nucleic acids takes advantage of highly evolved processes for targeting a virus to specific cells in the body and trafficking the viral payload to the nucleus. Viral vectors can be administered directly to patients (in vivo) or they can be used to treat cells in vitro, and the modified cells may optionally be administered to patients (ex vivo). Conventional viral based systems could include retroviral, lentivirus, adenoviral, adeno-associated and herpes simplex virus vectors for gene transfer. Integration in the host genome is possible with the retrovirus, lentivirus, and adeno-associated

virus gene transfer methods, often resulting in long term expression of the inserted transgene. Additionally, high transduction efficiencies have been observed in many different cell types and target tissues.

[0111] The tropism of a retrovirus can be altered by incorporating foreign envelope proteins, expanding the potential target population of target cells. Lentiviral vectors are retroviral vectors that are able to transduce or infect non-dividing cells and typically produce high viral titers. Selection of a retroviral gene transfer system would therefore depend on the target tissue. Retroviral vectors are comprised of cis-acting long terminal repeats with packaging capacity for up to 6-10 kb of foreign sequence. The minimum cis-acting LTRs are sufficient for replication and packaging of the vectors, which are then used to integrate the therapeutic gene into the target cell to provide permanent transgene expression. Widely used retroviral vectors include those based upon murine leukemia virus (MuLV), gibbon ape leukemia virus (GaLV), Simian Immuno deficiency virus (SIV), human immuno deficiency virus (HIV), and combinations thereof (see, e.g., Buchscher et al., J. Virol. 66:2731-2739 (1992); Johann et al., J. Virol. 66:1635-1640 (1992); Sommnerfelt et al., Virol. 176:58-59 (1990); Wilson et al., J. Virol. 63:2374-2378 (1989); Miller et al., J. Virol. 65:2220-2224 (1991); PCT/US94/05700).

[0112] In applications where transient expression is preferred, adenoviral based systems may be used. Adenoviral based vectors are capable of very high transduction efficiency in many cell types and do not require cell division. With such vectors, high titer and levels of expression have been obtained. This vector can be produced in large quantities in a relatively simple system. Adeno-associated virus ("AAV") vectors may also be used to transduce cells with target nucleic acids, e.g., in the in vitro production of nucleic acids and peptides, and for in vivo and ex vivo gene therapy procedures (see, e.g., West et al., Virology 160:38-47 (1987); U.S. Pat. No. 4,797,368; WO 93/24641; Kotin, Human Gene Therapy 5:793-801 (1994); Muzyczka, J. Clin. Invest. 94:1351 (1994). Construction of recombinant AAV vectors are described in a number of publications, including U.S. Pat. No. 5,173,414; Tratschin et al., Mol. Cell. Biol. 5:3251-3260 (1985); Tratschin, et al., Mol. Cell. Biol. 4:2072-2081 (1984); Hermonat & Muzyczka, PNAS 81:6466-6470 (1984); and Samulski et al., J. Virol. 63:03822-3828 (1989).

[0113] Packaging cells are typically used to form virus particles that are capable of infecting a host cell. Such cells include 293 cells, which package adenovirus, and W2 cells or PA317 cells, which package retrovirus. Viral vectors used in gene therapy are usually generated by producing a cell line that packages a nucleic acid vector into a viral particle. The vectors typically contain the minimal viral sequences required for packaging and subsequent integration into a host, other viral sequences being replaced by an expression cassette for the polynucleotide(s) to be expressed. The missing viral functions are typically supplied in trans by the packaging cell line. For example, AAV vectors used in gene therapy typically only possess ITR sequences from the AAV genome which are required for packaging and integration into the host genome. Viral DNA is packaged in a cell line, which contains a helper plasmid encoding the other AAV genes, namely rep and cap, but lacking ITR sequences. The cell line may also also infected with adenovirus as a helper. The helper virus promotes replication of the AAV vector and expression of AAV genes from the helper plasmid. The helper plasmid is not packaged in significant amounts due to a lack of ITR sequences. Contamination with adenovirus can be reduced by, e.g., heat treatment to which adenovirus is more sensitive than AAV. Additional methods for the delivery of nucleic acids to cells are known to those skilled in the art. See, for example, US20030087817, incorporated herein by reference.

[0114] In some embodiments, a host cell is transiently or non-transiently transfected with one or more vectors described herein. In some embodiments, a cell is transfected as it naturally occurs in a subject. In some embodiments, a cell that is transfected is taken from a subject. In some embodiments, the cell is derived from cells taken from a subject, such as a cell line. A wide variety of cell lines for tissue culture are known in the art. Examples of cell lines include, but are not limited to, C8161, CCRF-CEM, MOLT, mIMCD-3, NHDF, HeLa-S3, Huh1, Huh4, Huh7, HUVEC, HASMC, HEKn, HEKa, MiaPaCell, Panc1, PC-3, TF1, CTLL-2, C1R, Rat6, CV1, RPTE, A10, T24, J82, A375, ARH-77, Calu1, SW480, SW620, SKOV3, SK-UT, CaCo2, P388D1, SEM-K2, WEHI-231, HB56, TIB55, Jurkat, J45.01, LRMB, Bcl-1, BC-3, IC21, DLD2, Raw264.7, NRK, NRK-52E, MRC5, MEF, Hep G2, HeLa B, HeLa T4, COS, COS-1, COS-6, COS-M6A, BS-C-1 monkey kidney epithelial, BALB/3T3 mouse embryo fibroblast, 3T3 Swiss, 3T3-L1, 132-d5 human fetal fibroblasts; 10.1 mouse fibroblasts, 293-T, 3T3, 721, 9L, A2780, A2780ADR, A2780cis, A172, A20, A253, A431, A-549, ALC, B16, B35, BCP-1 cells, BEAS-2B, bEnd.3, BHK-21, BR 293, BxPC3, C3H-10T1/2, C6/36, Cal-27, CHO, CHO-7, CHO-IR, CHO-K1, CHO-K2, CHO-T, CHO Dhfr −/−, COR-L23, COR-L23/CPR, COR-L23/5010, COR-L23/R23, COS-7, COV-434, CML T1, CMT, CT26, D17, DH82, DU145, DuCaP, EL4, EM2, EM3, EMT6/AR1, EMT6/AR10.0, FM3, H1299, H69, HB54, HB55, HCA2, HEK-293, HeLa, Hepa1c1c7, HL-60, HMEC, HT-29, Jurkat, JY cells, K562 cells, Ku812, KCL22, KG1, KYO1, LNCap, Ma-Mel 1-48, MC-38, MCF-7, MCF-10A, MDA-MB-231, MDA-MB-468, MDA-MB-435, MDCK II, MDCK II, MOR/0.2R, MONO-MAC 6, MTD-1A, MyEnd, NCI-H69/CPR, NCI-H69/LX10, NCI-H69/LX20, NCI-H69/LX4, NIH-3T3, NALM-1, NW-145, OPCN/OPCT cell lines, Peer, PNT-1A/PNT 2, RenCa, RIN-5F, RMA/RMAS, Saos-2 cells, Sf-9, SkBr3, T2, T-47D, T84, THP1 cell line, U373, U87, U937, VCaP, Vero cells, WM39, WT-49, X63, YAC-1, YAR, and transgenic varieties thereof. Cell lines are available from a variety of sources known to those with skill in the art (see, e.g., the American Type Culture Collection (ATCC) (Manassas, Va.)). In some embodiments, a cell transfected with one or more vectors described herein is used to establish a new cell line comprising one or more vector-derived sequences. In some embodiments, a cell transiently transfected with the components of a CRISPR system as described herein (such as by transient transfection of one or more vectors, or transfection with RNA), and modified through the activity of a CRISPR complex, is used to establish a new cell line comprising cells containing the modification but lacking any other exogenous sequence. In some embodiments, cells transiently or non-transiently transfected with one or more vectors described herein, or cell lines derived from such cells are used in assessing one or more test compounds.

[0115] In some embodiments, one or more vectors described herein are used to produce a non-human trans-

genic animal or transgenic plant. In some embodiments, the transgenic animal is a mammal, such as a mouse, rat, or rabbit. In certain embodiments, the organism or subject is a plant. In certain embodiments, the organism or subject or plant is algae. Methods for producing transgenic plants and animals are known in the art, and generally begin with a method of cell transfection, such as described herein. Transgenic animals are also provided, as are transgenic plants, especially crops and algae. The transgenic animal or plant may be useful in applications outside of providing a disease model. These may include food or feed production through expression of, for instance, higher protein, carbohydrate, nutrient or vitamins levels than would normally be seen in the wildtype. In this regard, transgenic plants, especially pulses and tubers, and animals, especially mammals such as livestock (cows, sheep, goats and pigs), but also poultry and edible insects, are preferred.

[0116] Transgenic algae or other plants such as rape may be particularly useful in the production of vegetable oils or biofuels such as alcohols (especially methanol and ethanol), for instance. These may be engineered to express or over-express high levels of oil or alcohols for use in the oil or biofuel industries.

[0117] In one aspect, the invention provides for methods of modifying a target polynucleotide in a eukaryotic cell. In some embodiments, the method comprises allowing a CRISPR complex to bind to the target polynucleotide to effect cleavage of said target polynucleotide thereby modifying the target polynucleotide, wherein the CRISPR complex comprises a CRISPR enzyme complexed with a guide sequence hybridized to a target sequence within said target polynucleotide, wherein said guide sequence is linked to a tracr mate sequence which in turn hybridizes to a tracr sequence.

[0118] In one aspect, the invention provides a method of modifying expression of a polynucleotide in a eukaryotic cell. In some embodiments, the method comprises allowing a CRISPR complex to bind to the polynucleotide such that said binding results in increased or decreased expression of said polynucleotide; wherein the CRISPR complex comprises a CRISPR enzyme complexed with a guide sequence hybridized to a target sequence within said polynucleotide, wherein said guide sequence is linked to a tracr mate sequence which in turn hybridizes to a tracr sequence.

[0119] With recent advances in crop genomics, the ability to use CRISPR-Cas systems to perform efficient and cost effective gene editing and manipulation will allow the rapid selection and comparison of single and and multiplexed genetic manipulations to transform such genomes for improved production and enhanced traits. In this regard reference is made to US patents and publications: U.S. Pat. No. 6,603,061—*Agrobacterium*-Mediated Plant Transformation Method; U.S. Pat. No. 7,868,149—Plant Genome Sequences and Uses Thereof and US 2009/0100536—Transgenic Plants with Enhanced Agronomic Traits, all the contents and disclosure of each of which are herein incorporated by reference in their entirety. In the practice of the invention, the contents and disclosure of Morrell et al "Crop genomics: advances and applications" Nat Rev Genet. 2011 Dec. 29; 13(2):85-96 are also herein incorporated by reference in their entirety. In an advantageous embodiment of the invention, the CRISPR/Cas9 system is used to engineer microalgae (Example 14). Accordingly, reference herein to animal cells may also apply, mutatis mutandis, to plant cells unless otherwise apparent.

[0120] In one aspect, the invention provides for methods of modifying a target polynucleotide in a eukaryotic cell, which may be in vivo, ex vivo or in vitro. In some embodiments, the method comprises sampling a cell or population of cells from a human or non-human animal or plant (including micro-algae), and modifying the cell or cells. Culturing may occur at any stage ex vivo. The cell or cells may even be re-introduced into the non-human animal or plant (including micro-algae).

[0121] In one aspect, the invention provides kits containing any one or more of the elements disclosed in the above methods and compositions. In some embodiments, the kit comprises a vector system and instructions for using the kit. In some embodiments, the vector system comprises (a) a first regulatory element operably linked to a tracr mate sequence and one or more insertion sites for inserting a guide sequence upstream of the tracr mate sequence, wherein when expressed, the guide sequence directs sequence-specific binding of a CRISPR complex to a target sequence in a eukaryotic cell, wherein the CRISPR complex comprises a CRISPR enzyme complexed with (1) the guide sequence that is hybridized to the target sequence, and (2) the tracr mate sequence that is hybridized to the tracr sequence; and/or (b) a second regulatory element operably linked to an enzyme-coding sequence encoding said CRISPR enzyme comprising a nuclear localization sequence. Elements may provide individually or in combinations, and may provided in any suitable container, such as a vial, a bottle, or a tube. In some embodiments, the kit includes instructions in one or more languages, for example in more than one language.

[0122] In some embodiments, a kit comprises one or more reagents for use in a process utilizing one or more of the elements described herein. Reagents may be provided in any suitable container. For example, a kit may provide one or more reaction or storage buffers. Reagents may be provided in a form that is usable in a particular assay, or in a form that requires addition of one or more other components before use (e.g. in concentrate or lyophilized form). A buffer can be any buffer, including but not limited to a sodium carbonate buffer, a sodium bicarbonate buffer, a borate buffer, a Tris buffer, a MOPS buffer, a HEPES buffer, and combinations thereof. In some embodiments, the buffer is alkaline. In some embodiments, the buffer has a pH from about 7 to about 10. In some embodiments, the kit comprises one or more oligonucleotides corresponding to a guide sequence for insertion into a vector so as to operably link the guide sequence and a regulatory element. In some embodiments, the kit comprises a homologous recombination template polynucleotide.

[0123] In one aspect, the invention provides methods for using one or more elements of a CRISPR system. The CRISPR complex of the invention provides an effective means for modifying a target polynucleotide. The CRISPR complex of the invention has a wide variety of utility including modifying (e.g., deleting, inserting, translocating, inactivating, activating) a target polynucleotide in a multiplicity of cell types. As such the CRISPR complex of the invention has a broad spectrum of applications in, e.g., gene therapy, drug screening, disease diagnosis, and prognosis. An exemplary CRISPR complex comprises a CRISPR

enzyme complexed with a guide sequence hybridized to a target sequence within the target polynucleotide. The guide sequence is linked to a tracr mate sequence, which in turn hybridizes to a tracr sequence.

[0124] The target polynucleotide of a CRISPR complex can be any polynucleotide endogenous or exogenous to the eukaryotic cell. For example, the target polynucleotide can be a polynucleotide residing in the nucleus of the eukaryotic cell. The target polynucleotide can be a sequence coding a gene product (e.g., a protein) or a non-coding sequence (e.g., a regulatory polynucleotide or a junk DNA). Without wishing to be bound by theory, it is believed that the target sequence should be associated with a PAM (protospacer adjacent motif); that is, a short sequence recognised by the CRISPR complex. The precise sequence and length requirements for the PAM differ depending on the CRISPR enzyme used, but PAMs are typically 2-5 base pair sequences adjacent the protospacer (that is, the target sequence) Examples of PAM sequences are given in the examples section below, and the skilled person will be able to identify further PAM sequences for use with a given CRISPR enzyme.

[0125] The target polynucleotide of a CRISPR complex may include a number of disease-associated genes and polynucleotides as well as signaling biochemical pathway-associated genes and polynucleotides as listed in U.S. provisional patent applications 61/736,527 and 61/748,427 having Broad reference BI-2011/008/WSGR Docket No. 44063-701.101 and BI-2011/008/WSGR Docket No. 44063-701.102 respectively, both entitled SYSTEMS METHODS AND COMPOSITIONS FOR SEQUENCE MANIPULATION filed on Dec. 12, 2012 and Jan. 2, 2013, respectively, the contents of all of which are herein incorporated by reference in their entirety.

[0126] Examples of target polynucleotides include a sequence associated with a signaling biochemical pathway, e.g., a signaling biochemical pathway-associated gene or polynucleotide. Examples of target polynucleotides include a disease associated gene or polynucleotide. A "disease-associated" gene or polynucleotide refers to any gene or polynucleotide which is yielding transcription or translation products at an abnormal level or in an abnormal form in cells derived from a disease-affected tissues compared with tissues or cells of a non disease control. It may be a gene that becomes expressed at an abnormally high level; it may be a gene that becomes expressed at an abnormally low level, where the altered expression correlates with the occurrence and/or progression of the disease. A disease-associated gene also refers to a gene possessing mutation(s) or genetic variation that is directly responsible or is in linkage disequilibrium with a gene(s) that is responsible for the etiology of a disease. The transcribed or translated products may be known or unknown, and may be at a normal or abnormal level.

[0127] Examples of disease-associated genes and polynucleotides are available from McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, Md.) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, Md.), available on the World Wide Web.

[0128] Examples of disease-associated genes and polynucleotides are listed in Tables A and B. Disease specific information is available from McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, Md.) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, Md.), available on the World Wide Web. Examples of signaling biochemical pathway-associated genes and polynucleotides are listed in Table C.

[0129] Mutations in these genes and pathways can result in production of improper proteins or proteins in improper amounts which affect function. Further examples of genes, diseases and proteins are hereby incorporated by reference from U.S. Provisional applications 61/736,527 and 61/748,427. Such genes, proteins and pathways may be the target polynucleotide of a CRISPR complex.

TABLE A

DISEASE/DISORDERS	GENE(S)
Neoplasia	PTEN; ATM; ATR; EGFR; ERBB2; ERBB3; ERBB4; Notch1; Notch2; Notch3; Notch4; AKT; AKT2; AKT3; HIF; HIF1a; HIF3a; Met; HRG; Bcl2; PPAR alpha; PPAR gamma; WT1 (Wilms Tumor); FGF Receptor Family members (5 members: 1, 2, 3, 4, 5); CDKN2a; APC; RB retinoblastoma); MEN1; VHL; BRCA1; BRCA2; AR (Androgen Receptor); TSG101; IGF; IGF Receptor; Igf1 (4 variants); Igf2 (3 variants); Igf 1 Receptor; Igf 2 Receptor; Bax; Bcl2; caspases family (9 members: 1, 2, 3, 4, 6, 7, 8, 9, 12); Kras; Apc
Age-related Macular Degeneration	Abcr; Ccl2; Cc2; cp (ceruloplasmin); Timp3; cathepsinD; Vldlr; Ccr2
Schizophrenia	Neuregulin1 (Nrg1); Erb4 (receptor for Neuregulin); Complexin1 (Cplx1); Tph1 Tryptophan hydroxylase; Tph2 Tryptophan hydroxylase 2; Neurexin 1; GSK3; GSK3a; GSK3b
Disorders	5-HTT (Slc6a4); COMT; DRD (Drd1a); SLC6A3; DAOA; DTNBP1; Dao (Dao1)
Trinucleotide Repeat Disorders	HTT (Huntington's Dx); SBMA/SMAX1/AR (Kennedy's Dx); FXN/X25 (Friedrich's Ataxia); ATX3 (Machado-Joseph's Dx); ATXN1 and ATXN2 (spinocerebellar ataxias); DMPK (myotonic dystrophy); Atrophin-1 and Atn1 (DRPLA Dx); CBP (Creb-BP - global instability); VLDLR (Alzheimer's); Atxn7; Atxn10

TABLE A-continued

DISEASE/DISORDERS	GENE(S)
Fragile X Syndrome	FMR2; FXR1; FXR2; mGLUR5
Secretase Related Disorders	APH-1 (alpha and beta); Presenilin (Psen1); nicastrin (Ncstn); PEN-2
Others	Nos1; Parp1; Nat1; Nat2
Prion - related disorders	Prp
ALS	SOD1; ALS2; STEX; FUS; TARDBP; VEGF (VEGF-a; VEGF-b; VEGF-c)
Drug addiction	Prkce (alcohol); Drd2; Drd4; ABAT (alcohol); GRIA2; Grm5; Grin1; Htr1b; Grin2a; Drd3; Pdyn; Gria1 (alcohol)
Autism	Mecp2; BZRAP1; MDGA2; Sema5A; Neurexin 1; Fragile X (FMR2 (AFF2); FXR1; FXR2; Mglur5)
Alzheimer's Disease	E1; CHIP; UCH; UBB; Tau; LRP; PICALM; Clusterin; PS1; SORL1; CR1; Vldlr; Uba1; Uba3; CHIP28 (Aqp1, Aquaporin 1); Uchl1; Uchl3; APP
Inflammation	IL-10; IL-1 (IL-1a; IL-1b); IL-13; IL-17 (IL-17a (CTLA8); IL-17b; IL-17c; IL-17d; IL- -17f; II-23; Cx3cr1; ptpn22; TNFa; NOD2/CARD15 for IBD; IL-6; IL-12 (IL-12a; IL-12b); CTLA4; Cx3cl1
Parkinson's Disease	x-Synuclein; DJ-1; LRRK2; Parkin; PINK1

TABLE B

Blood and coagulation diseases and disorders	Anemia (CDAN1, CDA1, RPS19, DBA, PKLR, PK1, NT5C3, UMPH1, PSN1, RHAG, RH50A, NRAMP2, SPTB, ALAS2, ANH1, ASB, ABCB7, ABC7, ASAT); Bare lymphocyte syndrome (TAPBP, TpSN, TAP2, ABCB3, PSF2, RING11, MHC2TA, C2TA, RFX5, RFXAP, RFX5), Bleeding disorders (TBXA2R, P2RX1, P2X1); Factor H and factor H-like 1 (HF1, CFH, HUS); Factor V and factor VIII (MCFD2); Factor VII deficiency (F7); Factor X deficiency (F10); Factor XI deficiency (F11); Factor XII deficiency (F12, HAF); Factor XIIIA deficiency (F13A1, F13A); Factor XIIIB deficiency (F13B); Fanconi anemia (FANCA, FACA, FA1, FA, FAA, FAAP95, FAAP90, FLJ34064, FANCB, FANCC, FACC, BRCA2, FANCD1, FANCD2, FANCD, FACD, FAD, FANCE, FACE, FANCF, XRCC9, FANCG, BRIP1, BACH1, FANCJ, PHF9, FANCL, FANCM, KIAA1596); Hemophagocytic lymphohistiocytosis disorders (PRF1, HPLH2, UNC13D, MUNC13-4, HPLH3, HLH3, FHL3); Hemophilia A (F8, F8C, HEMA); Hemophilia B (F9, HEMB), Hemorrhagic disorders (PI, ATT, F5); Leukocyde deficiencies and disorders (ITGB2, CD18, LCAMB, LAD, EIF2B1, EIF2BA, EIF2B2, EIF2B3, EIF2B5, LVWM, CACH, CLE, EIF2B4); Sickle cell anemia (HBB); Thalassemia (HBA2, HBB, HBD, LCRB, HBA1)
Cell dysregulation and oncology diseases and disorders	B-cell non-Hodgkin lymphoma (BCL7A, BCL7); Leukemia TAL1, TCL5, SCL, TAL2, FLT3, NBS1, NBS, ZNFN1A1, IK1, LYF1, HOXD4, HOX4B, BCR, CML, PHL, ALL, ARNT, KRAS2, RASK2, GMPS, AF10, ARHGEF12, LARG, KIAA0382, CALM, CLTH, CEBPA, CEBP, CHIC2, BTL, FLT3, KIT, PBT, LPP, NPM1, NUP214, D9S46E, CAN, CAIN, RUNX1, CBFA2, AML1, WHSC1L1, NSD3, FLT3, AF1Q, NPM1, NUMA1, ZNF145, PLZF, PML, MYL, STAT5B, AF10, CALM, CLTH, ARL11, ARLTS1, P2RX7, P2X7, BCR, CML, PHL, ALL, GRAF, NF1, VRNF, WSS, NFNS, PTPN11, PTP2C, SHP2, NS1, BCL2, CCND1, PRAD1, BCL1, TCRA, GATA1, GF1, ERYF1, NFE1, ABL1, NQO1, DIA4, NMOR1, NUP214, D9S46E, CAN, CAIN).
Inflammation and immune related diseases and disorders	AIDS (KIR3DL1, NKAT3, NKB1, AMB11, KIR3DS1, IFNG, CXCL12, SDF1); Autoimmune lymphoproliferative syndrome (TNFRSF6, APT1, FAS, CD95, ALPS1A); Combined immunodeficiency, (IL2RG, SCIDX1, SCIDX, IMD4); HIV-1 (CCL5, SCYA5, D17S136E, TCP228), HIV susceptibility or infection (IL10, CSIF, CMKBR2, CCR2, CMKBR5, CCCKR5 (CCR5)); Immunodeficiencies (CD3E, CD3G, AICDA, AID, HIGM2, TNFRSF5, CD40, UNG, DGU, HIGM4, TNFSF5, CD40LG, HIGM1, IGM, FOXP3, IPEX, AIID, XPID, PIDX, TNFRSF14B, TACI); Inflammation (IL-10, IL-1 (IL-1a, IL-1b), IL-13, IL-17 (IL-17a (CTLA8), IL-17b, IL-17c, IL-17d, IL-17f), II-23, Cx3cr1, ptpn22, TNFa, NOD2/CARD15 for IBD, IL-6, IL-12 (IL-12a, IL-12b), CTLA4, Cx3cl1); Severe combined immunodeficiencies (SCIDs) (JAK3, JAKL, DCLRE1C, ARTEMIS, SCIDA, RAG1, RAG2, ADA, PTPRC, CD45, LCA, IL7R, CD3D, T3D, IL2RG, SCIDX1, SCIDX, IMD4).
Metabolic, liver, kidney and protein diseases and disorders	Amyloid neuropathy (TTR, PALB); Amyloidosis (APOA1, APP, AAA, CVAP, AD1, GSN, FGA, LYZ, TTR, PALB); Cirrhosis (KRT18, KRT8, CIRH1A, NAIC, TEX292, KIAA1988); Cystic fibrosis (CFTR, ABCC7, CF, MRP7); Glycogen storage diseases (SLC2A2, GLUT2, G6PC, G6PT, G6PT1, GAA, LAMP2, LAMPB, AGL, GDE, GBE1, GYS2,

TABLE B-continued

	PYGL, PFKM); Hepatic adenoma, 142330 (TCF1, HNF1A, MODY3), Hepatic failure; early onset, and neurologic disorder (SCOD1, SCO1), Hepatic lipase deficiency (LIPC), Hepatoblastoma, cancer and carcinomas (CTNNB1, PDGFRL, PDGRL, PRLTS, AXIN1, AXIN, CTNNB1, TP53, P53, LFS1, IGF2R, MPRI, MET, CASP8, MCH5; Medullary cystic kidney disease (UMOD, HNFJ, FJHN, MCKD2, ADMCKD2); Phenylketonuria (PAH, PKU1, QDPR, DHPR, PTS); Polycystic kidney and hepatic disease (FCYT, PKHD1, ARPKD, PKD1, PKD2, PKD4, PKDTS, PRKCSH, G19P1, PCLD, SEC63).
Muscular/Skeletal diseases and disorders	Becker muscular dystrophy (DMD, BMD, MYF6) Duchenne Muscular Dystrophy (DMD, BMD); Emery-Dreifuss muscular dystrophy (LMNA, LMN1, EMD2, FPLD, CMD1A, HGPS, LGMD1B, LMNA, LMN1, EMD2, FPLD, CMD1A); Facioscapulohumeral muscular dystrophy (FSHMD1A, FSHD1A); Muscular dystrophy (FKRP, MDC1C, LGMD2I, LAMA2, LAMM, LARGE, KIAA0609, MDC1D, FCMD, TTID, MYOT, CAPN3, CANP3, DYSF, LGMD2B, SGCG, LGMD2C, DMDA1, SCG3, SGCA, ADL, DAG2, LGMD2D, DMDA2, SGCB, LGMD2E, SGCD, SGD, LGMD2F, CMD1L, TCAP, LGMD2G, CMD1N, TRIM32, HT2A, LGMD2H, FKRP, MDC1C, LGMD2I, TTN, CMD1G, TMD, LGMD2J, POMT1, CAV3, LGMD1C, SEPN1, SELN, RSMD1, PLEC1, PLTN, EBS1); Osteopetrosis (LRP5, BMND1, LRP7, LR3, OPPG, VBCH2, CLCN7, CLC7, OPTA2, OSTM1, GL, TCIRG1, TIRC7, OC116, OPTB1); Muscular atrophy (VAPB, VAPC, ALS8, SMN1, SMA1, SMA2, SMA3, SMA4, BSCL2, SPG17, GARS, SMAD1, CMT2D, HEXB, IGHMBP2, SMUBP2, CATF1, SMARD1).
Neurological and neuronal diseases and disorders	ALS (SOD1, ALS2, STEX, FUS, TARDBP, VEGF (VEGF-a, VEGF-b, VEGF-c); Alzheimer disease (APP, AAA, CVAP, AD1, APOE, AD2, PSEN2, AD4, STM2, APBB2, FE65L1, NOS3, PLAU, URK, ACE, DCP1, ACE1, MPO, PACIP1, PAXIP1L, PTIP, A2M, BLMH, BMH, PSEN1, AD3); Autism (Mecp2, BZRAP1, MDGA2, Sema5A, Neurexin 1, GLO1, MECP2, RTT, PPMX, MRX16, MRX79, NLGN3, NLGN4, KIAA1260, AUTSX2); Fragile X Syndrome (FMR2, FXR1, FXR2, mGLUR5); Huntington's disease and disease like disorders (HD, IT15, PRNP, PRIP, JPH3, JP3, HDL2, TBP, SCA17); Parkinson disease (NR4A2, NURR1, NOT, TINUR, SNCAIP, TRP, SCA17, SNCA, NACP, PARK1, PARK4, DJ1, PARK7, LRRK2, PARK8, PINK1, PARK6, UCHL1, PARK5, SNCA, NACP, PARK1, PARK4, PRKN, PARK2, PDJ, DBH, NDUFV2); Rett syndrome (MECP2, RTT, PPMX, MRX16, MRX79, CDKL5, STK9, MECP2, RTT, PPMX, MRX16, MRX79, x-Synuclein, DJ-1); Schizophrenia (Neuregulin1 (Nrg1), Erb4 (receptor for Neuregulin), Complexin1 (CpLx1), Tph1 Tryptophan hydroxylase, Tph2, Tryptophan hydroxylase 2, Neurexin 1, GSK3, GSK3a, GSK3b, 5-HTT (Slc6a4), COMT, DRD (Drd1a), SLC6A3, DAOA, DTNBP1, Dao (Dao1)); Secretase Related Disorders (APH-1 (alpha and beta); Presenilin (Psen1), nicastrin, (Ncstn), PEN-2, Nos1, Parp1, Nat1, NaT2); Trinucleotide Repeat Disorders (HTT (Huntington's Dx), SBMA/SMAX1/AR (Kennedy's Dx), FXN/X25 (Friedrich's Ataxia); ATX3 (Machado-Joseph's Dx), ATXN1 and ATXN2 spinocerebellar ataxias), DMPK (myotonic dystrophy), Atrophin1 and Atn1 (DRPLA Dx), CBP (Creb-BP - global instability), VLDLR (Alzheimer's), Atxn7, Atxn10).
Occular diseases and disorders	Age-related macular degeneration (Abcr, Ccl2, Cc2, cp (ceruloplasmin), Timp3, cathepsinD, Vldlr, Ccr2); Cataract (CRYAA, CRYA1, CRYBB2, CRYB2, PITX3, BFSP2, CP49, CP47, CRYAA, CRYA1, PAX6, AN2, MGDA, CRYBA1, CRYB1, CRYGC, CRYG3, CCL, LIM2, MP19, CRYGD, CRYG4, BFSP2, CP49, CP47, HSF4, CTM, HSF4, CTM, MIP, AQP0, CRYAB, CRYA2, CTPP2, CRYBB1, CRYGD, CRYG4, CRYBB2, CRYB2, CRYGC, CRYG3, CCL, CRYAA, CRYA1, GJA8, CX50, CAE1, GJA3, CX46, CZP3, CAE3, CCM1, CAM, KRIT1); Corneal clouding and dystrophy (APOA1, TGFBI, CSD2 CDGG1, CSD, BIGH3, CDG2, TACSTD2 TROP2, M1S1, VSX1, RINX, PPCD, PPD, KTCN, COL8A2, FECD, PPCD2, PIP5K3, CFD); Cornea plana congenital (KERA, CNA2); Glaucoma (MYOC, TIGR, GLC1A, JOAG, GPOA, OPTN, GLC1E, FIP2, HYPL, NRP, CYP1B1, GLC3A, OPA1, NTG, NPG, CYP1B1, GLC3A); Leber congenital amaurosis (CRB1, RP12, CRX, CORD2, CRD, RPGRIP1, LCA6, CORD9, RPE65, RP20, AIPL1, LCA4, GUCY2D, GUC2D, LCA1, CORD6, RDH12, LCA3); Macular dystrophy (ELOVL4, ADMD, STGD2, STGD3, RDS, RP7, PRPH2, PRPH, AVMD, AOFMD, VMD2).

TABLE C

CELLULAR FUNCTION	GENES
PI3K/AKT Signaling	PRKCE; ITGAM; ITGA5; IRAK1; PRKAA2; EIF2AK2; PTEN; EIF4E; PRKCZ; GRK6; MAPK1; TSC1; PLK1; AKT2; IKBKB; PIK3CA; CDK8; CDKN1B; NFKB2; BCL2; PIK3CB; PPP2R1A; MAPK8; BCL2L1; MAPK3; TSC2; ITGA1; KRAS; EIF4EBP1; RELA; PRKCD; NOS3; PRKAA1; MAPK9; CDK2; PPP2CA; PIM1; ITGB7; YWHAZ; ILK; TP53; RAF1; IKBKG; RELB; DYRK1A; CDKN1A; ITGB1; MAP2K2; JAK1; AKT1; JAK2; PIK3R1; CHUK; PDPK1; PPP2R5C; CTNNB1; MAP2K1; NFKB1; PAK3; ITGB3; CCND1; GSK3A; FRAP1; SFN; ITGA2; TTK; CSNK1A1; BRAF; GSK3B; AKT3; FOXO1; SGK; HSP90AA1; RPS6KB1
ERK/MAPK Signaling	PRKCE; ITGAM; ITGA5; HSPB1; IRAK1; PRKAA2; EIF2AK2; RAC1; RAP1A; TLN1; EIF4E; ELK1; GRK6; MAPK1; RAC2; PLK1; AKT2; PIK3CA; CDK8; CREB1; PRKCI; PTK2; FOS; RPS6KA4; PIK3CB; PPP2R1A; PIK3C3; MAPK8; MAPK3; ITGA1; ETS1; KRAS; MYCN; EIF4EBP1; PPARG; PRKCD; PRKAA1; MAPK9; SRC; CDK2; PPP2CA; PIM1; PIK3C2A; ITGB7; YWHAZ; PPP1CC; KSR1; PXN; RAF1; FYN; DYRK1A; ITGB1; MAP2K2; PAK4; PIK3R1; STAT3; PPP2R5C; MAP2K1; PAK3; ITGB3; ESR1; ITGA2; MYC; TTK; CSNK1A1; CRKL; BRAF; ATF4; PRKCA; SRF; STAT1; SGK
Glucocorticoid Receptor Signaling	RAC1; TAF4B; EP300; SMAD2; TRAF6; PCAF; ELK1; MAPK1; SMAD3; AKT2; IKBKB; NCOR2; UBE2I; PIK3CA; CREB1; FOS; HSPA5; NFKB2; BCL2; MAP3K14; STAT5B; PIK3CB; PIK3C3; MAPK8; BCL2L1; MAPK3; TSC22D3; MAPK10; NRIP1; KRAS; MAPK13; RELA; STAT5A; MAPK9; NOS2A; PBX1; NR3C1; PIK3C2A; CDKN1C; TRAF2; SERPINE1; NCOA3; MAPK14; TNF; RAF1; IKBKG; MAP3K7; CREBBP; CDKN1A; MAP2K2; JAK1; IL8; NCOA2; AKT1; JAK2; PIK3R1; CHUK; STAT3; MAP2K1; NFKB1; TGFBR1; ESR1; SMAD4; CEBPB; JUN; AR; AKT3; CCL2; MMP1; STAT1; IL6; HSP90AA1
Axonal Guidance Signaling	PRKCE; ITGAM; ROCK1; ITGA5; CXCR4; ADAM12; IGF1; RAC1; RAP1A; EIF4E; PRKCZ; NRP1; NTRK2; ARHGEF7; SMO; ROCK2; MAPK1; PGF; RAC2; PTPN11; GNAS; AKT2; PIK3CA; ERBB2; PRKCI; PTK2; CFL1; GNAQ; PIK3CB; CXCL12; PIK3C3; WNT11; PRKD1; GNB2L1; ABL1; MAPK3; ITGA1; KRAS; RHOA; PRKCD; PIK3C2A; ITGB7; GLI2; PXN; VASP; RAF1; FYN; ITGB1; MAP2K2; PAK4; ADAM17; AKT1; PIK3R1; GLI1; WNT5A; ADAM10; MAP2K1; PAK3; ITGB3; CDC42; VEGFA; ITGA2; EPHA8; CRKL; RND1; GSK3B; AKT3; PRKCA
Ephrin Receptor Signaling	PRKCE; ITGAM; ROCK1; ITGA5; CXCR4; IRAK1; PRKAA2; EIF2AK2; RAC1; RAP1A; GRK6; ROCK2; MAPK1; PGF; RAC2; PTPN11; GNAS; PLK1; AKT2; DOK1; CDK8; CREB1; PTK2; CFL1; GNAQ; MAP3K14; CXCL12; MAPK8; GNB2L1; ABL1; MAPK3; ITGA1; KRAS; RHOA; PRKCD; PRKAA1; MAPK9; SRC; CDK2; PIM1; ITGB7; PXN; RAF1; FYN; DYRK1A; ITGB1; MAP2K2; PAK4; AKT1; JAK2; STAT3; ADAM10; MAP2K1; PAK3; ITGB3; CDC42; VEGFA; ITGA2; EPHA8; TTK; CSNK1A1; CRKL; BRAF; PTPN13; ATF4; AKT3; SGK
Actin Cytoskeleton Signaling	ACTN4; PRKCE; ITGAM; ROCK1; ITGA5; IRAK1; PRKAA2; EIF2AK2; RAC1; INS; ARHGEF7; GRK6; ROCK2; MAPK1; RAC2; PLK1; AKT2; PIK3CA; CDK8; PTK2; CFL1; PIK3CB; MYH9; DIAPH1; PIK3C3; MAPK8; F2R; MAPK3; SLC9A1; ITGA1; KRAS; RHOA; PRKCD; PRKAA1; MAPK9; CDK2; PIM1; PIK3C2A; ITGB7; PPP1CC; PXN; VIL2; RAF1; GSN; DYRK1A; ITGB1; MAP2K2; PAK4; PIP5K1A; PIK3R1; MAP2K1; PAK3; ITGB3; CDC42; APC; ITGA2; TTK; CSNK1A1; CRKL; BRAF; VAV3; SGK
Huntington's Disease Signaling	PRKCE; IGF1; EP300; RCOR1; PRKCZ; HDAC4; TGM2; MAPK1; CAPNS1; AKT2; EGFR; NCOR2; SP1; CAPN2; PIK3CA; HDAC5; CREB1; PRKCI; HSPA5; REST; GNAQ; PIK3CB; PIK3C3; MAPK8; IGF1R; PRKD1; GNB2L1; BCL2L1; CAPN1; MAPK3; CASP8; HDAC2;

TABLE C-continued

CELLULAR FUNCTION	GENES
	HDAC7A; PRKCD; HDAC11; MAPK9; HDAC9; PIK3C2A; HDAC3; TP53; CASP9; CREBBP; AKT1; PIK3R1; PDPK1; CASP1; APAF1; FRAP1; CASP2; JUN; BAX; ATF4; AKT3; PRKCA; CLTC; SGK; HDAC6; CASP3
Apoptosis Signaling	PRKCE; ROCK1; BID; IRAK1; PRKAA2; EIF2AK2; BAK1; BIRC4; GRK6; MAPK1; CARNS1; PLK1; AKT2; IKBKB; CAPN2; CDK8; FAS; NFKB2; BCL2; MAP3K14; MAPK8; BCL2L1; CAPN1; MAPK3; CASP8; KRAS; RELA; PRKCD; PRKAA1; MAPK9; CDK2; PIM1; TP53; TNF; RAF1; IKBKG; RELB; CASP9; DYRK1A; MAP2K2; CHUK; APAF1; MAP2K1; NFKB1; PAK3; LMNA; CASP2; BIRC2; TTK; CSNK1A1; BRAF; BAX; PRKCA; SGK; CASP3; BIRC3; PARP1
B Cell Receptor Signaling	RAC1; PTEN; LYN; ELK1; MAPK1; RAC2; PTPN11; AKT2; IKBKB; PIK3CA; CREB1; SYK; NFKB2; CAMK2A; MAP3K14; PIK3CB; PIK3C3; MAPK8; BCL2L1; ABL1; MAPK3; ETS1; KRAS; MAPK13; RELA; PTPN6; MAPK9; EGR1; PIK3C2A; BTK; MAPK14; RAF1; IKBKG; RELB; MAP3K7; MAP2K2; AKT1; PIK3R1; CHUK; MAP2K1; NFKB1; CDC42; GSK3A; FRAP1; BCL6; BCL10; JUN; GSK3B; ATF4; AKT3; VAV3; RPS6KB1
Leukocyte Extravasation Signaling	ACTN4; CD44; PRKCE; ITGAM; ROCK1; CXCR4; CYBA; RAC1; RAP1A; PRKCZ; ROCK2; RAC2; PTPN11; MMP14; PIK3CA; PRKCI; PIK2; PIK3CB; CXCL12; PIK3C3; MAPK8; PRKD1; ABL1; MAPK10; CYBB; MAPK13; RHOA; PRKCD; MAPK9; SRC; PIK3C2A; BTK; MAPK14; NOX1; PXN; VIL2; VASP; ITGB1; MAP2K2; CTNND1; PIK3R1; CTNNB1; CLDN1; CDC42; F11R; ITK; CRKL; VAV3; CTTN; PRKCA; MMP1; MMP9
Integrin Signaling	ACTN4; ITGAM; ROCK1; ITGA5; RAC1; PTEN; RAP1A; TLN1; ARHGEF7; MAPK1; RAC2; CAPNS1; AKT2; CAPN2; PIK3CA; PTK2; PIK3CB; PIK3C3; MAPK8; CAV1; CAPN1; ABL1; MAPK3; ITGA1; KRAS; RHOA; SRC; PIK3C2A; ITGB7; PPP1CC; ILK; PXN; VASP; RAF1; FYN; ITGB1; MAP2K2; PAK4; AKT1; PIK3R1; TNK2; MAP2K1; PAK3; ITGB3; CDC42; RND3; ITGA2; CRKL; BRAF; GSK3B; AKT3
Acute Phase Response Signaling	IRAK1; SOD2; MYD88; TRAF8; ELK1; MAPK1; PTPN11; AKT2; IKBKB; PIK3CA; FOS; NFKB2; MAP3K14; PIK3CB; MAPK8; RIPK1; MAPK3; IL6ST; KRAS; MAPK13; IL6R; RELA; SOCS1; MAPK9; FTL; NR3C1; TRAF2; SERPINE1; MAPK14; TNF; RAF1; PDK1; IKBKG; RELB; MAP3K7; MAP2K2; AKT1; JAK2; PIK3R1; CHUK; STAT3; MAP2K1; NFKB1; FRAP1; CEBPB; JUN; AKT3; IL1R1; IL6
PTEN Signaling	ITGAM; ITGA5; RAC1; PTEN; PRKCZ; BCL2L11; MAPK1; RAC2; AKT2; EGFR; IKBKB; CBL; PIK3CA; CDKN1B; PTK2; NFKB2; BCL2; PIK3CB; BCL2L1; MAPK3; ITGA1; KRAS; ITGB7; ILK; PDGFRB; INSR; RAF1; IKBKG; CASP9; CDKN1A; ITGB1; MAP2K2; AKT1; PIK3R1; CHUK; PDGFRA; PDPK1; MAP2K1; NFKB1; ITGB3; CDC42; CCND1; GSK3A; ITGA2; GSK3B; AKT3; FOXO1; CASP3; RPS6KB1
p53 Signaling	PTEN; EP300; BBC3; PCAF; FASN; BRCA1; GADD45A; BIRC5; AKT2; PIK3CA; CHEK1; TP53INP1; BCL2; PIK3CB; PIK3C3; MAPK8; THBS1; ATR; BCL2L1; E2F1; PMAIP1; CHEK2; TNFRSF10B; TP73; RB1; HDAC9; CDK2; PIK3C2A; MAPK14; TP53; LRDD; CDKN1A; HIPK2; AKT1; PIK3R1; RRM2B; APAF1; CTNNB1; SIRT1; CCND1; PRKDC; ATM; SFN; CDKN2A; JUN; SNAI2; GSK3B; BAX; AKT3
Aryl Hydrocarbon Receptor Signaling	HSPB1; EP300; FASN; TGM2; RXRA; MAPK1; NQO1; NCOR2; SP1; ARNT; CDKN1B; FOS; CHEK1; SMARCA4; NFKB2; MAPK8; ALDH1A1; ATR; E2F1; MAPK3; NRIP1; CHEK2; RELA; TP73; GSTP1; RB1; SRC; CDK2; AHR; NFE2L2; NCOA3; TP53; TNF; CDKN1A; NCOA2; APAF1; NFKB1; CCND1; ATM; ESR1; CDKN2A; MYC; JUN; ESR2; BAX; IL6; CYP1B1; HSP90AA1
Xenobiotic Metabolism Signaling	PRKCE; EP300; PRKCZ; RXRA; MAPK1; NQO1; NCOR2; PIK3CA; ARNT; PRKCI; NFKB2; CAMK2A; PIK3CB; PPP2R1A; PIK3C3; MAPK8; PRKD1; ALDH1A1; MAPK3; NRIP1; KRAS; MAPK13; PRKCD; GSTP1; MAPK9; NOS2A; ABCB1; AHR; PPP2CA; FTL;

TABLE C-continued

CELLULAR FUNCTION	GENES
	NFE2L2; PIK3C2A; PPARGC1A; MAPK14; TNT; RAF1; CREBBP; MAP2K2; PIK3R1; PPP2R5C; MAP2K1; NFTKB1; KEAP1; PRKCA; EIF2AK3; IL6; CYP1B1; HSP90AA1
SAPK/JNK Signaling	PRKCE; IRAK1; PRKAA2; EIF2AK2; RAC1; ELK1; GRK6; MAPK1; GADD45A; RAC2; PLK1; AKT2; PIK3CA; FADD; CDK8; PIK3CB; PIK3C3; MAPK8; RIPK1; GNB2L1; IRS1; MAPK3; MAPK10; DAXX; KRAS; PRKCD; PRKAA1; MAPK9; CDK2; PIM1; PIK3C2A; TRAF2; TP53; LCK; MAP3K7; DYRK1A; MAP2K2; PIK3R1; MAP2K1; PAK3; CDC42; JUN; TTK; CSNK1A1; CRKL; BRAF; SGK
PPAr/RXR Signaling	PRKAA2; EP300; INS; SMAD2; TRAF6; PPARA; FASN; RXRA; MAPK1; SMAD3; GNAS; IKBKB; NCOR2; ABCA1; GNAQ; NFKB2; MAP3K14; STAT5B; MAPK8; IRS1; MAPK3; KRAS; RELA; PRKAA1; PPARGC1A; NCOA3; MAPK14; INSR; RAF1; IKBKG; RELB; MAP3K7; CREBBP; MAP2K2; JAK2; CHUK; MAP2K1; NFKB1; TGFBR1; SMAD4; JUN; IL1R1; PRKCA; IL6; HSP90AA1; ADIPOQ
NF-KB Signaling	IRAK1; EIF2AK2; EP300; INS; MYD88; PRKCZ; TRAF6; TBK1; AKT2; EGFR; IKBKB; PIK3CA; BTRC; NFKB2; MAP3K14; PIK3CB; PIK3C3; MAPK8; RIPK1; HDAC2; KRAS; RELA; PIK3C2A; TRAF2; TLR4; PDGFRB; TNF; INSR; LCK; IKBKG; RELB; MAP3K7; CREBBP; AKT1; PIK3R1; CHUK; PDGFRA; NFKB1; TLR2; BCL10; GSK3B; AKT3; TNFAIP3; IL1R1
Neuregulin Signaling	ERBB4; PRKCE; ITGAM; ITGA5; PTEN; PRKCZ; ELK1; MAPK1; PTPN11; AKT2; EGFR; ERBB2; PRKCI; CDKN1B; STAT5B; PRKD1; MAPK3; ITGA1; KRAS; PRKCD; STAT5A; SRC; ITGB7; RAF1; ITGB1; MAP2K2; ADAM17; AKT1; PIK3R1; PDPKI; MAP2K1; ITGB3; EREG; FRAP1; PSEN1; ITGA2; MYC; NRG1; CRKL; AKT3; PRKCA; HSP90AA1; RPS6KB1
Wnt & Beta catenin Signaling	CD44; EP300; LRP6; DVL3; CSNK1E; GJA1; SMO; AKT2; PIN1; CDH1; BTRC; GNAQ; MARK2; PPP2R1A; WNT11; SRC; DKK1; PPP2CA; SOX6; SFRP2; ILK; LEF1; SOX9; TP53; MAP3K7; CREBBP; TCF7L2; AKT1; PPP2R5C; WNT5A; LRP5; CTNNB1; TGFBR1; CCND1; GSK3A; DVL1; APC; CDKN2A; MYC; CSNK1A1; GSK3B; AKT3; SOX2
Insulin Receptor Signaling	PTEN; INS; EIF4E; PTPN1; PRKCZ; MAPK1; TSC1; PTPN11; AKT2; CBL; PIK3CA; PRKCI; PIK3CB; P1K3C3; MAPK8; IRS1; MAPK3; TSC2; KRAS; EIF4EBP1; SLC2A4; PIK3C2A; PPP1CC; INSR; RAF1; FYN; MAP2K2; JAK1; AKT1; JAK2; PIK3R1; PDPK1; MAP2K1; GSK3A; FRAP1; CRKL; GSK3B; AKT3; FOXO1; SGK; RPS6KB1
IL-6 Signaling	HSPB1; TRAF6; MAPKAPK2; ELK1; MAPK1; PTPN11; IKBKB; FOS; NFKB2; MAP3K14; MAPK8; MAPK3; MAPK10; IL6ST; KRAS; MAPK13; IL6R; RELA; SOCS1; MAPK9; ABCB1; TRAF2; MAPK14; TNF; RAF1; IKBKG; RELB; MAP3K7; MAP2K2; IL8; JAK2; CHUK; STAT3; MAP2K1; NFKB1; CEBPB; JUN; IL1R1; SRF; IL6
Hepatic Cholestasis	PRKCE; IRAK1; INS; MYD88; PRKCZ; TRAF6; PPARA; RXRA; IKBKB; PRKCI; NFKB2; MAP3K14; MAPK8; PRKD1; MAPK10; RELA; PRKCD; MAPK9; ABCB1; TRAF2; TLR4; TNF; INSR; IKBKG; RELB; MAP3K7; IL8; CHUK; NR1H2; TJP2; NFKB1; ESR1; SREBF1; FGFR4; JUN; IL1R1; PRKCA; IL6
IGF- I Signaling	IGF1; PRKCZ; ELK1; MAPK1; PTPN11; NEDD4; AKT2; PIK3CA; PRKCI; PTK2; FOS; PIK3CB; PIK3C3; MAPK8; IGF1R; IRS1; MAPK3; IGFBP7; KRAS; PIK3C2A; YWHAZ; PXN; RAF1; CASP9; MAP2K2; AKT1; PIK3R1; PDPK1; MAP2K1; IGFBP2; SFN; JUN; CYR61; AKT3; FOXO1; SRF; CTGF; RPS6KB1
NRF2-mediated Oxidative Stress Response	PRKCE; EP300; SOD2; PRKCZ; MAPK1; SQSTM1; NQO1; PIK3CA; PRKCI; FOS; PIK3CB; PIK3C3; MAPK8; PRKD1; MAPK3; KRAS; PRKCD; GSTP1; MAPK9; FTL; NFE2L2; PIK3C2A; MAPK14; RAF1; MAP3K7; CREBBP; MAP2K2; AKT1; PIK3R1; MAP2K1; PPIB; JUN; KEAP1; GSK3B; ATF4; PRKCA; EIF2AK3; HSP90AA1

TABLE C-continued

CELLULAR FUNCTION	GENES
Hepatic Fibrosis/Hepatic Stellate Cell Activation	EDN1; IGF1; KDR; FLT1; SMAD2; FGFR1; MET; PGF; SMAD3; EGFR; FAS; CSF1; NFKB2; BCL2; MYH9; IGF1R; IL6R; RELA; TLR4; PDGFRB; TNF; RELB; IL8; PDGFRA; NFKB1; TGFBR1; SMAD4; VEGFA; BAX; IL1R1; CCL2; HGF; MMP1; STAT1; IL6; CTGF; MMP9
PPAR Signaling	EP300; INS; TRAF6; PPARA; RXRA; MAPK1; IKBKB; NCOR2; FOS; NFKB2; MAP3K14; STAT5B; MAPK3; NRIP1; KRAS; PPARG; RELA; STAT5A; TRAF2; PPARGC1A; PDGFRB; TNF; INSR; RAF1; IKBKG; RELB; MAP3K7; CREBBP; MAP2K2; CHUK; PDGFRA; MAP2K1; NFKB1; JUN; IL1R1; HSP90AA1
Fc Epsilon RI Signaling	PRKCE; RAC1; PRKCZ; LYN; MAPK1; RAC2; PTPN11; AKT2; PIK3CA; SYK; PRKCI; PIK3CB; PIK3C3; MAPK8; PRKD1; MAPK3; MAPK10; KRAS; MAPK13; PRKCD; MAPK9; PIK3C2A; BTK; MAPK14; TNF; RAF1; FYN; MAP2K2; AKT1; PIK3R1; PDPK1; MAP2K1; AKT3; VAV3; PRKCA
G-Protein Coupled Receptor Signaling	PRKCE; RAP1A; RGS16; MAPK1; GNAS; AKT2; IKBKB; PIK3CA; CREB1; GNAQ; NFKB2; CAMK2A; PIK3CB; PIK3C3; MAPK3; KRAS; RELA; SRC; PIK3C2A; RAF1; IKBKG; RELB; FYN; MAP2K2; AKT1; PIK3R1; CHUK; PDPK1; STAT3; MAP2K1; NFKB1; BRAF; ATF4; AKT3; PRKCA
Inositol Phosphate Metabolism	PRKCE; IRAK1; PRKAA2; EIF2AK2; PTEN; GRK6; MAPK1; PLK1; AKT2; PIK3CA; CDK8; PIK3CB; PIK3C3; MAPK8; MAPK3; PRKCD; PRKAA1; MAPK9; CDK2; PIM1; PIK3C2A; DYRK1A; MAP2K2; PIP5K1A; PIK3R1; MAP2K1; PAK3; ATM; TTK; CSNK1A1; BRAF; SGK
PDGF Signaling	EIF2AK2; ELK1; ABL2; MAPK1; PIK3CA; FOS; PIK3CB; PIK3C3; MAPK8; CAV1; ABL1; MAPK3; KRAS; SRC; PIK3C2A; PDGFRB; RAF1; MAP2K2; JAK1; JAK2; PIK3R1; PDGFRA; STAT3; SPHK1; MAP2K1; MYC; JUN; CRKL; PRKCA; SRF; STAT1; SPHK2
VEGF Signaling	ACTN4; ROCK1; KDR; FLT1; ROCK2; MAPK1; PGF; AKT2; PIK3CA; ARNT; PTK2; BCL2; PIK3CB; PIK3C3; BCL2L1; MAPK3; KRAS; HIF1A; NOS3; PIK3C2A; PXN; RAF1; MAP2K2; ELAVL1; AKT1; PIK3R1; MAP2K1; SFN; VEGFA; AKT3; FOXO1; PRKCA
Natural Killer Cell Signaling	PRKCE; RAC1; PRKCZ; MAPK1; RAC2; PTPN11; KIR2DL3; AKT2; PIK3CA; SYK; PRKCI; PIK3CB; PIK3C3; PRKD1; MAPK3; KRAS; PRKCD; PTPN6; PIK3C2A; LCK; RAF1; FYN; MAP2K2; PAK4; AKT1; PIK3R1; MAP2K1; PAK3; AKT3; VAV3; PRKCA
Cell Cycle: G1/S Checkpoint Regulation	HDAC4; SMAD3; SUV39H1; HDAC5; CDKN1B; BTRC; ATR; ABL1; E2F1; HDAC2; HDAC7A; RB1; HDAC11; HDAC9; CDK2; E2F2; HDAC3; TP53; CDKN1A; CCND1; E2F4; ATM; RBL2; SMAD4; CDKN2A; MYC; NRG1; GSK3B; RBL1; HDAC6
T Cell Receptor Signaling	RAC1; ELK1; MAPK1; IKBKB; CBL; PIK3CA; FOS; NFKB2; PIK3CB; PIK3C3; MAPK8; MAPK3; KRAS; RELA; PIK3C2A; BTK; LCK; RAF1; IKBKG; RELB; FYN; MAP2K2; PIK3R1; CHUK; MAP2K1; NFKB1; ITK; BCL10; JUN; VAV3
Death Receptor Signaling	CRADD; HSPB1; BID; BIRC4; TBK1; IKBKB; FADD; FAS; NFKB2; BCL2; MAP3K14; MAPK8; RIPK1; CASP8; DAXX; TNFRSF10B; RELA; TRAF2; TNF; IKBKG; RELB; CASP9; CHUK; APAF1; NFKB1; CASP2; BIRC2; CASP3; BIRC3
FGF Signaling	RAC1; FGFR1; MET; MAPKAPK2; MAPK1; PTPN11; AKT2; PIK3CA; CREB1; PIK3CB; PIK3C3; MAPK8; MAPK3; MAPK13; PTPN6; PIK3C2A; MAPK14; RAF1; AKT1; PIK3R1; STAT3; MAP2K1; FGFR4; CRKL; ATF4; AKT3; PRKCA; HGF
GM-CSF Signaling	LYN; ELK1; MAPK1; PTPN11; AKT2; PIK3CA; CAMK2A; STAT5B; PIK3CB; PIK3C3; GNB2L1; BCL2L1; MAPK3; ETS1; KRAS; RUNX1; PIM1; PIK3C2A; RAF1; MAP2K2; AKT1; JAK2; PIK3R1; STAT3; MAP2K1; CCND1; AKT3; STAT1
Amyotrophic Lateral Sclerosis Signaling	BID; IGF1; RAC1; BIRC4; PGF; CAPNS1; CAPN2; PIK3CA; BCL2; PIK3CB; PIK3C3; BCL2L1; CAPN1; PIK3C2A; TP53; CASP9; PIK3R1; RAB5A; CASP1; APAF1; VEGFA; BIRC2; BAX; AKT3; CASP3; BIRC3

TABLE C-continued

CELLULAR FUNCTION	GENES
JAK/Stat Signaling	PTPN1; MAPK1; PTPN11; AKT2; PIK3CA; STAT5B; PIK3CB; PIK3C3; MAPK3; KRAS; SOCS1; STAT5A; PTPN6; PIK3C2A; RAF1; CDKN1A; MAP2K2; JAK1; AKT1; JAK2; PIK3R1; STAT3; MAP2K1; FRAP1; AKT3; STAT1
Nicotinate and Nicotinamide Metabolism	PRKCE; IRAK1; PRKAA2; EIF2AK2; GRK6; MAPK1; PLK1; AKT2; CDK8; MAPK8; MAPK3; PRKCD; PRKAA1; PBEF1; MAPK9; CDK2; PIM1; DYRK1A; MAP2K2; MAP2K1; PAK3; NT5E; TTK; CSNK1A1; BRAF; SGK
Chemokine Signaling	CXCR4; ROCK2; MAPK1; PTK2; FOS; CFL1; GNAQ; CAMK2A; CXCL12; MAPK8; MAPK3; KRAS; MAPK13; RHOA; CCR3; SRC; PPP1CC; MAPK14; NOX1; RAF1; MAP2K2; MAP2K1; JUN; CCL2; PRKCA
IL-2 Signaling	ELK1; MAPK1; PTPN11; AKT2; PIK3CA; SYK; FOS; STAT5B; PIK3CB; PIK3C3; MAPK8; MAPK3; KRAS; SOCS1; STAT5A; PIK3C2A; LCK; RAF1; MAP2K2; JAK1; AKT1; PIK3R1; MAP2K1; JUN; AKT3
Synaptic Long Term Depression	PRKCE; IGF1; PRKCZ; PRDX6; LYN; MAPK1; GNAS; PRKCI; GNAQ; PPP2R1A; IGF1R; PRKD1; MAPK3; KRAS; GRN; PRKCD; NOS3; NOS2A; PPP2CA; YWHAZ; RAF1; MAP2K2; PPP2R5C; MAP2K1; PRKCA
Estrogen Receptor Signaling	TAF4B; EP300; CARM1; PCAF; MAPK1; NCOR2; SMARCA4; MAPK3; NRIP1; KRAS; SRC; NR3C1; HDAC3; PPARGC1A; RBM9; NCOA3; RAF1; CREBBP; MAP2K2; NCOA2; MAP2K1; PRKDC; ESR1; ESR2
Protein Ubiquitination Pathway	TRAF6; SMURF1; BIRC4; BRCA1; UCHL1; NEDD4; CBL; UBE2I; BTRC; HSPA5; USP7; USP10; FBXW7; USP9X; STUB1; USP22; B2M; BIRC2; PARK2; USP8; USP1; VHL; HSP90AA1; BIRC3
IL-10 Signaling	TRAF6; CCR1; ELK1; IKBKB; SP1; FOS; NFKB2; MAP3K14; MAPK8; MAPK13; RELA; MAPK14; TNF; IKBKG; RELB; MAP3K7; JAK1; CHUK; STAT3; NFKB1; JUN; IL1R1; IL6
VDR/RXR Activation	PRKCE; EP300; PRKCZ; RXRA; GADD45A; HES1; NCOR2; SP1; PRKCI; CDKN1B; PRKD1; PRKCD; RUNX2; KLF4; YY1; NCOA3; CDKN1A; NCOA2; SPP1; LRP5; CEBPB; FOXO1; PRKCA
TGF-beta Signaling	EP300; SMAD2; SMURF1; MAPK1; SMAD3; SMAD1; FOS; MAPK8; MAPK3; KRAS; MAPK9; RUNX2; SERPINE1; RAF1; MAP3K7; CREBBP; MAPP2K2; MAP2K1; TGFBR1; SMAD4; JUN; SMAD5
Toll-like Receptor Signaling	IRAK1; EIF2AK2; MYD88; TRAF6; PPARA; ELK1; IKBKB; FOS; NFKB2; MAP3K14; MAPK8; MAPK13; RELA; TLR4; MAPK14; IKBKG; RELB; MAP3K7; CHUK; NFKB1; TLR2; JUN
p38 MAPK Signaling	HSPB1; IRAK1; TRAF6; MAPKAPK2; ELK1; FADD; FAS; CREB1; DDIT3; RPS6KA4; DAXX; MAPK13; TRAF2; MAPK14; TNF; MAP3K7; TGFBR1; MYC; ATF4; IL1R1; SRF; STAT1
Neurotrophin/TRK Signaling	NTRK2; MAPK1; PTPN11; PIK3CA; CREB1; FOS; PIK3CB; PIK3C3; MAPK8; MAPK3; KRAS; PIK3C2A; RAF1; MAP2K2; AKT1; PIK3R1; PDPK1; MAP2K1; CDC42; JUN; ATF4
FXR/RXR Activation	INS; PPARA; FASN; RXRA; AKT2; SDC1; MAPK8; APOB; MAPK10; PPARG; MTTP; MAPK9; PPARGC1A; TNF; CREBBP; AKT1; SREBF1; FGFR4; AKT3; FOXO1
Synaptic Long Term Potentiation	PRKCE; RAP1A; EP300; PRKCZ; MAPK1; CREB1; PRKCI; GNAQ; CAMK2A; PRKD1; MAPK3; KRAS; PRKCD; PPP1CC; RAF1; CREBBP; MAP2K2; MAP2K1; ATF4; PRKCA
Calcium Signaling	RAP1A; EP300; HDAC4; MAPK1; HDAC5; CREB1; CAMK2A; MYH9; MAPK3; HDAC2; HDAC7A; HDAC11; HDAC9; HDAC3; CREBBP; CALR; CAMKK2; ATF4; HDAC6
EGF Signaling	ELK1; MAPK1; EGFR; PIK3CA; FOS; PIK3CB; PIK3C3; MAPK8; MAPK3; PIK3C2A; RAF1; JAK1; PIK3R1; STAT3; MAP2K1; JUN; PRKCA; SRF; STAT1
Hypoxia Signaling in the Cardiovascular System	EDN1; PTEN; EP300; NQO1; UBE2I; CREB1; ARNT; HIF1A; SLC2A4; NOS3; TP53; LDHA; AKT1; ATM; VEGFA; JUN; ATF4; VHL; HSP90AA1

TABLE C-continued

CELLULAR FUNCTION	GENES
LPS/IL-1 Mediated Inhibition of RXR Function	IRAK1; MYD88; TRAF6; PPARA; RXRA; ABCA1; MAPK8; ALDH1A1; GSTP1; MAPK9; ABCB1; TRAF2; TLR4; TNF; MAP3K7; NR1H2; SREBF1; JUN; IL1R1
LXR/RXR Activation	FASN; RXRA; NCOR2; ABCA1; NFKB2; IRF3; RELA; NOS2A; TLR4; TNF; RELB; LDLR; NR1H2; NFKB1; SREBF1; IL1R1; CCL2; IL6; MMP9
Amyloid Processing	PRKCE; CSNK1E; MAPK1; CAPNS1; AKT2; CAPN2; CAPN1; MAPK3; MAPK13; MAPT; MAPK14; AKT1; PSEN1; CSNK1A1; GSK3B; AKT3; APP
IL-4 Signaling	AKT2; PIK3CA; PIK3CB; PIK3C3; IRS1; KRAS; SOCS1; PTPN6; NR3C1; PIK3C2A; JAK1; AKT1; JAK2; PIK3R1; FRAP1; AKT3; RPS6KB1
Cell Cycle: G2/M DNA Damage Checkpoint Regulation	EP300; PCAF; BRCA1; GADD45A; PLK1; BTRC; CHEK1; ATR; CHEK2; YWHAZ; TP53; CDKN1A; PRKDC; ATM; SFN; CDKN2A
Nitric Oxide Signaling in the Cardiovascular System	KDR; FLT1; PGF; AKT2; PIK3CA; PIK3CB; PIK3C3; CAV1; PRKCD; NOS3; PIK3C2A; AKT1; PIK3R1; VEGFA; AKT3; HSP90AA1
Purine Metabolism	NME2; SMARCA4; MYH9; RRM2; ADAR; EIF2AK4; PKM2; ENTPD1; RAD51; RRM2B; TJP2; RAD51C; NT5E; POLD1; NME1
cAMP-mediated Signaling	RAP1A; MAPK1; GNAS; CREB1; CAMK2A; MAPK3; SRC; RAF1; MAP2K2; STAT3; MAP2K1; BRAF; ATF4
Mitochondrial Dysfunction	SOD2; MAPK8; CASP8; MAPK10; MAPK9; CASP9; PARK7; PSEN1; PARK2; APP; CASP3
Notch Signaling	HES1; JAG1; NUMB; NOTCH4; ADAM17; NOTCH2; PSEN1; NOTCH3; NOTCH1; DLL4
Endoplasmic Reticulum Stress Pathway	HSPA5; MAPK8; XBP1; TRAF2; ATF6; CASP9; ATF4; EIF2AK3; CASP3
Pyrimidine Metabolism	NME2; AICDA; RRM2; EIF2AK4; ENTPD1; RRM2B; NT5E; POLD1; NME1
Parkinson's Signaling	UCHL1; MAPK8; MAPK13; MAPK14; CASP9; PARK7; PARK2; CASP3
Cardiac & Beta Adrenergic Signaling	GNAS; GNAQ; PPP2R1A; GNB2L1; PPP2CA; PPP1CC; PPP2R5C
Glycolysis/ Gluconeogenesis	HK2; GCK; GPI; ALDH1A1; PKM2; LDHA; HK1
Interferon Signaling	IRF1; SOCS1; JAK1; JAK2; IFITM1; STAT1; IFIT3
Sonic Hedgehog Signaling	ARRB2; SMO; GLI2; DYRK1A; GLI1; GSK3B; DYRK1B
Glycerophospholipid Metabolism	PLD1; GRN; GPAM; YWHAZ; SPHK1; SPHK2
Phospholipid Degradation	PRDX6; PLD1; GRN; YWHAZ; SPHK1; SPHK2
Tryptophan Metabolism	SIAH2; PRMT5; NEDD4; ALDH1A1; CYP1B1; SIAH1
Lysine Degradation	SUV39H1; EHMT2; NSD1; SETD7; PPP2R5C
Nucleotide Excision Repair Pathway	ERCC5; ERCC4; XPA; XPC; ERCC1
Starch and Sucrose Metabolism	UCHL1; HK2; GCK; GPI; HK1
Aminosugars Metabolism	NQO1; HK2; GCK; HK1
Arachidonic Acid Metabolism	PRDX6; GRN; YWHAZ; GYP1B1
Circadian Rhythm Signaling	CSNK1E; CREB1; ATF4; NR1D1
Coagulation System	BDKRB1; F2R; SERPINE1; F3
Dopamine Receptor Signaling	PPP2R1A; PPP2CA; PPP1CC; PPP2R5C
Glutathione Metabolism	IDH2; GSTP1; ANPEP; IDH1
Glycerolipid Metabolism	ALDH1A1; GPAM; SPHK1; SPHK2
Linoleic Acid Metabolism	PRDX6; GRN; YWHAZ; CYP1B1
Methionine Metabolism	DNMT1; DNMT3B; AHCY; DNMT3A
Pyruvate Metabolism	GLO1; ALDH1A1; PKM2; LDHA
Arginine and Proline Metabolism	ALDH1A1; NOS3; NOS2A
Eicosanoid Signaling	PRDX6; GRN; YWHAZ
Fructose and Mannose Metabolism	HK2; GCK; HK1
Galactose Metabolism	HK2; GCK; HK1
Stilbene, Coumarine and Lignin Biosynthesis	PRDX6; PRDX1; TYR

TABLE C-continued

CELLULAR FUNCTION	GENES
Antigen Presentation Pathway	CALR; B2M
Biosynthesis of Steroids	NQO1; DHCR7
Butanoate Metabolism	ALDH1A1; NLGN1
Citrate Cycle	IDH2; IDH1
Fatty Acid Metabolism	ALDH1A1; CYP1B1
Glycerophospholipid Metabolism	PRDX6; CHKA
Histidine Metabolism	PRMT5; ALDH1A1
Inositol Metabolism	ERO1L; APEX1
Metabolism of Xenobiotics by Cytochrome p450	GSTP1; CYP1B1
Methane Metabolism	PRDX6; PRDX1
Phenylalanine Metabolism	PRDX6; PRDX1
Propanoate Metabolism	ALDH1A1; LDHA
Selenoamino Acid Metabolism	PRMT5; AHCY
Sphingolipid Metabolism	SPHK1; SPHK2
Aminophosphonate Metabolism	PRMT5
Androgen and Estrogen Metabolism	PRMT5
Ascorbate and Aldarate Metabolism	ALDH1A1
Bile Acid Biosynthesis	ALDH1A1
Cysteine Metabolism	LDHA
Fatty Acid Biosynthesis	FASN
Glutamate Receptor Signaling	GNB2L1
NRF2-mediated Oxidative Stress Response	PRDX1
Pentose Phosphate Pathway	GPI
Pentose and Glucuronate Interconversions	UCHL1
Retinol Metabolism	ALDH1A1
Riboflavin Metabolism	TYR
Tyrosine Metabolism	PRMT5, TYR
Ubiquinone Biosynthesis	PRMT5
Valine, Leucine and Isoleucine Degradation	ALDH1A1
Glycine, Serine and Threonine Metabolism	CHKA
Lysine Degradation	ALDH1A1
Pain/Taste	TRPM5; TRPA1
Pain	TRPM7; TRPC5; TRPC6; TRPC1; Cnr1; cnr2; Grk2; Trpa1; Pomc; Cgrp; Crf; Pka; Era; Nr2b; TRPM5; Prkaca; Prkacb; Prkar1a; Prkar2a
Mitochondrial Function	AIF; CytC; SMAC (Diablo); Aifm-1; Aifm-2
Developmental Neurology	BMP-4; Chordin (Chrd); Noggin (Nog); WNT (Wnt2; Wnt2b; Wnt3a; Wnt4; Wnt5a; Wnt6; Wnt7b; Wnt8b; Wnt9a; Wnt9b; Wnt10a; Wnt10b; Wnt16); beta-catenin; Dkk-1; Frizzled related proteins; Otx2; Gbx2; FGF-8; Reelin; Dab1; unc-86 (Pou4f1 or Brn3a); Numb; Reln

[0130] Embodiments of the invention also relate to methods and compositions related to knocking out genes, amplifying genes and repairing particular mutations associated with DNA repeat instability and neurological disorders (Robert D. Wells, Tetsuo Ashizawa, Genetic Instabilities and Neurological Diseases, Second Edition, Academic Press, Oct. 13, 2011—Medical). Specific aspects of tandem repeat sequences have been found to be responsible for more than twenty human diseases (New insights into repeat instability: role of RNA•DNA hybrids. McIvor E I, Polak U, Napierala M. RNA Biol. 2010 September-October; 7(5):551-8). The CRISPR-Cas system may be harnessed to correct these defects of genomic instability.

[0131] A further aspect of the invention relates to utilizing the CRISPR-Cas system for correcting defects in the EMP2A and EMP2B genes that have been identified to be associated with Lafora disease. Lafora disease is an autosomal recessive condition which is characterized by progressive myoclonus epilepsy which may start as epileptic seizures in adolescence. A few cases of the disease may be caused by mutations in genes yet to be identified. The disease causes seizures, muscle spasms, difficulty walking, dementia, and eventually death. There is currently no therapy that has proven effective against disease progression. Other genetic abnormalities associated with epilepsy

may also be targeted by the CRISPR-Cas system and the underlying genetics is further described in Genetics of Epilepsy and Genetic Epilepsies, edited by Giuliano Avanzini, Jeffrey L. Noebels, Mariani Foundation Paediatric Neurology:20; 2009).

[0132] In yet another aspect of the invention, the CRISPR-Cas system may be Cl used to correct ocular defects that arise from several genetic mutations further described in Genetic Diseases of the Eye, Second Edition, edited by Elias I. Traboulsi, Oxford University Press, 2012.

[0133] Several further aspects of the invention relate to correcting defects associated with a wide range of genetic diseases which are further described on the website of the National Institutes of Health under the topic subsection Genetic Disorders. The genetic brain diseases may include but are not limited to Adrenoleukodystrophy, Agenesis of the Corpus Callosum, Aicardi Syndrome, Alpers' Disease, Alzheimer's Disease, Barth Syndrome, Batten Disease, CADASIL, Cerebellar Degeneration, Fabry's Disease, Gerstmann-Straussler-Scheinker Disease, Huntington's Disease and other Triplet Repeat Disorders, Leigh's Disease, Lesch-Nyhan Syndrome, Menkes Disease, Mitochondrial Myopathies and NINDS Colpocephaly. These diseases are further described on the website of the National Institutes of Health under the subsection Genetic Brain Disorders.

[0134] In some embodiments, the condition may be neoplasia. In some embodiments, where the condition is neoplasia, the genes to be targeted are any of those listed in Table A (in this case PTEN asn so forth). In some embodiments, the condition may be Age-related Macular Degeneration. In some embodiments, the condition may be a Schizophrenic Disorder. In some embodiments, the condition may be a Trinucleotide Repeat Disorder. In some embodiments, the condition may be Fragile X Syndrome. In some embodiments, the condition may be a Secretase Related Disorder. In some embodiments, the condition may be a Prion-related disorder. In some embodiments, the condition may be ALS. In some embodiments, the condition may be a drug addiction. In some embodiments, the condition may be Autism. In some embodiments, the condition may be Alzheimer's Disease. In some embodiments, the condition may be inflammation. In some embodiments, the condition may be Parkinson's Disease.

[0135] Examples of proteins associated with Parkinson's disease include but are not limited to α-synuclein, DJ-1, LRRK2, PINK1, Parkin, UCHL1, Synphilin-1, and NURR1.

[0136] Examples of addiction-related proteins may include ABAT for example.

[0137] Examples of inflammation-related proteins may include the monocyte chemoattractant protein-1 (MCP1) encoded by the Ccr2 gene, the C-C chemokine receptor type 5 (CCR5) encoded by the Ccr5 gene, the IgG receptor IIB (FCGR2b, also termed CD32) encoded by the Fcgr2b gene, or the Fc epsilon R1g (FCER1g) protein encoded by the Fcer1g gene, for example.

[0138] Examples of cardiovascular diseases associated proteins may include IL1B (interleukin 1, beta), XDH (xanthine dehydrogenase), TP53 (tumor protein p53), PTGIS (prostaglandin 12 (prostacyclin) synthase), MB (myoglobin), IL4 (interleukin 4), ANGPT1 (angiopoietin 1), ABCG8 (ATP-binding cassette, sub-family G (WHITE), member 8), or CTSK (cathepsin K), for example.

[0139] Examples of Alzheimer's disease associated proteins may include the very low density lipoprotein receptor protein (VLDLR) encoded by the VLDLR gene, the ubiquitin-like modifier activating enzyme 1 (UBA1) encoded by the UBA1 gene, or the NEDD8-activating enzyme E1 catalytic subunit protein (UBEIC) encoded by the UBA3 gene, for example.

[0140] Examples of proteins associated with Autism Spectrum Disorder may include the benzodiazapine receptor (peripheral) associated protein 1 (BZRAP1) encoded by the BZRAP1 gene, the AF4/FMR2 family member 2 protein (AFF2) encoded by the AFF2 gene (also termed MFR2), the fragile X mental retardation autosomal homolog 1 protein (FXR1) encoded by the FXR1 gene, or the fragile X mental retardation autosomal homolog 2 protein (FXR2) encoded by the FXR2 gene, for example.

[0141] Examples of proteins associated with Macular Degeneration may include the ATP-binding cassette, sub-family A (ABC1) member 4 protein (ABCA4) encoded by the ABCR gene, the apolipoprotein E protein (APOE) encoded by the APOE gene, or the chemokine (C-C motif) Ligand 2 protein (CCL2) encoded by the CCL2 gene, for example.

[0142] Examples of proteins associated with Schizophrenia may include NRG1, ErbB4, CPLX1, TPH1, TPH2, NRXN1, GSK3A, BDNF, DISC1, GSK3B, and combinations thereof.

[0143] Examples of proteins involved in tumor suppression may include ATM (ataxia telangiectasia mutated), ATR (ataxia telangiectasia and Rad3 related), EGFR (epidermal growth factor receptor), ERBB2 (v-erb-b2 erythroblastic leukemia viral oncogene homolog 2), ERBB3 (v-erb-b2 erythroblastic leukemia viral oncogene homolog 3), ERBB4 (v-erb-b2 erythroblastic leukemia viral oncogene homolog 4), Notch 1, Notch2, Notch 3, or Notch 4, for example.

[0144] Examples of proteins associated with a secretase disorder may include PSENEN (presenilin enhancer 2 homolog (*C. elegans*)), CTSB (cathepsin B), PSEN1 (presenilin 1), APP (amyloid beta (A4) precursor protein), APH1B (anterior pharynx defective 1 homolog B (*C. elegans*)), PSEN2 (presenilin 2 (Alzheimer disease 4)), or BACE1 (beta-site APP-cleaving enzyme 1), for example.

[0145] Examples of proteins associated with Amyotrophic Lateral Sclerosis may include SOD1 (superoxide dismutase 1), ALS2 (amyotrophic lateral sclerosis 2), FUS (fused in sarcoma), TARDBP (TAR DNA binding protein), VAGFA (vascular endothelial growth factor A), VAGFB (vascular endothelial growth factor B), and VAGFC (vascular endothelial growth factor C), and any combination thereof.

[0146] Examples of proteins associated with prion diseases may include SOD1 (superoxide dismutase 1), ALS2 (amyotrophic lateral sclerosis 2), FUS (fused in sarcoma), TARDBP (TAR DNA binding protein), VAGFA (vascular endothelial growth factor A), VAGFB (vascular endothelial growth factor B), and VAGFC (vascular endothelial growth factor C), and any combination thereof.

[0147] Examples of proteins related to neurodegenerative conditions in prion disorders may include A2M (Alpha-2-Macroglobulin), AATF (Apoptosis antagonizing transcription factor), ACPP (Acid phosphatase prostate), ACTA2 (Actin alpha 2 smooth muscle aorta), ADAM22 (ADAM metallopeptidase domain), ADORA3 (Adenosine A3 receptor), or ADRA1D (Alpha-1D adrenergic receptor for Alpha-1D adrenoreceptor), for example.

[0148] Examples of proteins associated with Immunodeficiency may include A2M [alpha-2-macroglobulin];

AANAT [arylalkylamine N-acetyltransferase]; ABCA1 [ATP-binding cassette, sub-family A (ABC1), member 1]; ABCA2 [ATP-binding cassette, sub-family A (ABC1), member 2]; or ABCA3 [ATP-binding cassette, sub-family A (ABC1), member 3]; for example.

[0149] Examples of proteins associated with Trinucleotide Repeat Disorders include AR (androgen receptor), FMR1 (fragile X mental retardation 1), HTT (huntingtin), or DMPK (dystrophia myotonica-protein kinase), FXN (frataxin), ATXN2 (ataxin 2), for example.

[0150] Examples of proteins associated with Neurotransmission Disorders include SST (somatostatin), NOS1 (nitric oxide synthase 1 (neuronal)), ADRA2A (adrenergic, alpha-2A-, receptor), ADRA2C (adrenergic, alpha-2C-, receptor), TACR1 (tachykinin receptor 1), or HTR2c (5-hydroxytryptamine (serotonin) receptor 2C), for example.

[0151] Examples of neurodevelopmental-associated sequences include A2BP1 [ataxin 2-binding protein 1], AADAT [aminoadipate aminotransferase], AANAT [arylalkylamine N-acetyltransferase], ABAT [4-aminobutyrate aminotransferase], ABCA1 [ATP-binding cassette, sub-family A (ABC1), member 1], or ABCA13 [ATP-binding cassette, sub-family A (ABC1), member 13], for example.

[0152] Further examples of preferred conditions treatable with the present system include may be selected from: Aicardi-Goutieres Syndrome; Alexander Disease; Allan-Herndon-Dudley Syndrome; POLG-Related Disorders; Alpha-Mannosidosis (Type II and III); Alstrom Syndrome; Angelman; Syndrome; Ataxia-Telangiectasia; Neuronal Ceroid-Lipofuscinoses; Beta-Thalassemia; Bilateral Optic Atrophy and (Infantile) Optic Atrophy Type 1; Retinoblastoma (bilateral); Canavan Disease; Cerebrooculofacioskeletal Syndrome 1 [COFS1]; Cerebrotendinous Xanthomatosis; Cornelia de Lange Syndrome; MAPT-Related Disorders; Genetic Prion Diseases; Dravet Syndrome; Early-Onset Familial Alzheimer Disease; Friedreich Ataxia [FRDA]; Fryns Syndrome; Fucosidosis; Fukuyama Congenital Muscular Dystrophy; Galactosialidosis; Gaucher Disease; Organic Acidemias; Hemophagocytic Lymphohistiocytosis; Hutchinson-Gilford Progeria Syndrome; Mucolipidosis II; Infantile Free Sialic Acid Storage Disease; PLA2G6-Associated Neurodegeneration; Jervell and Lange-Nielsen Syndrome; Junctional Epidermolysis Bullosa; Huntington Disease; Krabbe Disease (Infantile); Mitochondrial DNA-Associated Leigh Syndrome and NARP; Lesch-Nyhan Syndrome; LIS1-Associated Lissencephaly; Lowe Syndrome; Maple Syrup Urine Disease; MECP2 Duplication Syndrome; ATP7A-Related Copper Transport Disorders; LAMA2-Related Muscular Dystrophy; Arylsulfatase A Deficiency; Mucopolysaccharidosis Types I, II or III; Peroxisome Biogenesis Disorders, Zellweger Syndrome Spectrum; Neurodegeneration with Brain Iron Accumulation Disorders; Acid Sphingomyelinase Deficiency; Niemann-Pick Disease Type C; Glycine Encephalopathy; ARX-Related Disorders; Urea Cycle Disorders; COL1A1/2-Related Osteogenesis Imperfecta; Mitochondrial DNA Deletion Syndromes; PLP1-Related Disorders; Perry Syndrome; Phelan-McDermid Syndrome; Glycogen Storage Disease Type II (Pompe Disease) (Infantile); MAPT-Related Disorders; MECP2-Related Disorders; Rhizomelic Chondrodysplasia Punctata Type 1; Roberts Syndrome; Sandhoff Disease; Schindler Disease-Type 1; Adenosine Deaminase Deficiency; Smith-Lemli-Opitz Syndrome; Spinal Muscular Atrophy; Infantile-Onset Spinocerebellar Ataxia; Hexosaminidase A Deficiency; Thanatophoric Dysplasia Type 1; Collagen Type VI-Related Disorders; Usher Syndrome Type I; Congenital Muscular Dystrophy; Wolf-Hirschhorn Syndrome; Lysosomal Acid Lipase Deficiency; and Xeroderma Pigmentosum.

[0153] Chronic administration of protein therapeutics may elicit unacceptable immune responses to the specific protein. The immunogenicity of protein drugs can be ascribed to a few immunodominant helper T lymphocyte (HTL) epitopes. Reducing the MHC binding affinity of these HTL epitopes contained within these proteins can generate drugs with lower immunogenicity (Tangri S, et al. ("Rationally engineered therapeutic proteins with reduced immunogenicity" J Immunol. 2005 Mar. 15; 174(6):3187-96.) In the present invention, the immunogenicity of the CRISPR enzyme in particular may be reduced following the approach first set out in Tangri et al with respect to erythropoietin and subsequently developed. Accordingly, directed evolution or rational design may be used to reduce the immunogenicity of the CRISPR enzyme (for instance a Cas9) in the host species (human or other species).

[0154] In plants, pathogens are often host-specific. For example, *Fusarium oxysporum* f. sp. *lycopersici* causes tomato wilt but attacks only tomato, and *F. oxysporum f. dianthii Puccinia graminis* f. sp. *tritici* attacks only wheat. Plants have existing and induced defenses to resist most pathogens. Mutations and recombination events across plant generations lead to genetic variability that gives rise to susceptibility, especially as pathogens reproduce with more frequency than plants. In plants there can be non-host resistance, e.g., the host and pathogen are incompatible. There can also be Horizontal Resistance, e.g., partial resistance against all races of a pathogen, typically controlled by many genes and Vertical Resistance, e.g., complete resistance to some races of a pathogen but not to other races, typically controlled by a few genes. In a Gene-for-Gene level, plants and pathogens evolve together, and the genetic changes in one balance changes in other. Accordingly, using Natural Variability, breeders combine most useful genes for Yield, Quality, Uniformity, Hardiness, Resistance. The sources of resistance genes include native or foreign Varieties, Heirloom Varieties, Wild Plant Relatives, and Induced Mutations, e.g., treating plant material with mutagenic agents. Using the present invention, plant breeders are provided with a new tool to induce mutations. Accordingly, one skilled in the art can analyze the genome of sources of resistance genes, and in Varieties having desired characteristics or traits employ the present invention to induce the rise of resistance genes, with more precision than previous mutagenic agents and hence accelerate and improve plant breeding programs.

[0155] As will be apparent, it is envisaged that the present system can be used to target any polynucleotide sequence of interest. Some examples of conditions or diseases that might be usefully treated using the present system are included in the Tables above and examples of genes currently associated with those conditions are also provided there. However, the genes exemplified are not exhaustive.

EXAMPLES

[0156] The following examples are given for the purpose of illustrating various embodiments of the invention and are not meant to limit the present invention in any fashion. The present examples, along with the methods described herein

are presently representative of preferred embodiments, are exemplary, and are not intended as limitations on the scope of the invention. Changes therein and other uses which are encompassed within the spirit of the invention as defined by the scope of the claims will occur to those skilled in the art.

Example 1: CRISPR Complex Activity in the Nucleus of a Eukaryotic Cell

[0157] An example type II CRISPR system is the type II CRISPR locus from *Streptococcus pyogenes* SF370, which contains a cluster of four genes Cas9, Cas1, Cas2, and Csn1, as well as two non-coding RNA elements, tracrRNA and a characteristic array of repetitive sequences (direct repeats) interspaced by short stretches of non-repetitive sequences (spacers, about 30 bp each). In this system, targeted DNA double-strand break (DSB) is generated in four sequential steps (FIG. **2**A). First, two non-coding RNAs, the pre-crRNA array and tracrRNA, are transcribed from the CRISPR locus. Second, tracrRNA hybridizes to the direct repeats of pre-crRNA, which is then processed into mature crRNAs containing individual spacer sequences. Third, the mature crRNA:tracrRNA complex directs Cas9 to the DNA target consisting of the protospacer and the corresponding PAM via heteroduplex formation between the spacer region of the crRNA and the protospacer DNA. Finally, Cas9 mediates cleavage of target DNA upstream of PAM to create a DSB within the protospacer (FIG. **2**A). This example describes an example process for adapting this RNA-programmable nuclease system to direct CRISPR complex activity in the nuclei of eukaryotic cells.

[0158] To improve expression of CRISPR components in mammalian cells, two genes from the SF370 locus 1 of *Streptococcus pyogenes* (*S. pyogenes*) were codon-optimized, Cas9 (SpCas9) and RNase III (SpRNase III). To facilitate nuclear localization, a nuclear localization signal (NLS) was included at the amino (N)- or carboxyl (C)-termini of both SpCas9 and SpRNase III (FIG. **2**B). To facilitate visualization of protein expression, a fluorescent protein marker was also included at the N- or C-termini of both proteins (FIG. **2**B). A version of SpCas9 with an NLS attached to both N- and C-termini (2xNLS-SpCas9) was also generated. Constructs containing NLS-fused SpCas9 and SpRNase III were transfected into 293FT human embryonic kidney (HEK) cells, and the relative positioning of the NLS to SpCas9 and SpRNase III was found to affect their nuclear localization efficiency. Whereas the C-terminal NLS was sufficient to target SpRNase III to the nucleus, attachment of a single copy of these particular NLS's to either the N- or C-terminus of SpCas9 was unable to achieve adequate nuclear localization in this system. In this example, the C-terminal NLS was that of nucleoplasmin (KRPAATKK-AGQAKKKK) (SEQ ID NO: 3), and the C-terminal NLS was that of the SV40 large T-antigen (PKKKRKV) (SEQ ID NO: 2). Of the versions of SpCas9 tested, only 2xNLS-SpCas9 exhibited nuclear localization (FIG. **2**B).

[0159] The tracrRNA from the CRISPR locus of *S. pyogenes* SF370 has two transcriptional start sites, giving rise to two transcripts of 89-nucleotides (nt) and 171 nt that are subsequently processed into identical 75 nt mature tracrRNAs. The shorter 89 nt tracrRNA was selected for expression in mammalian cells (expression constructs illustrated in FIG. **6**, with functionality as determined by results of Surveryor assay shown in FIG. **6**B). Transcription start sites are marked as +1, and transcription terminator and the sequence probed by northern blot are also indicated. Expression of processed tracrRNA was also confirmed by Northern blot. FIG. **7**C shows results of a Northern blot analysis of total RNA extracted from 293FT cells transfected with U6 expression constructs carrying long or short tracrRNA, as well as SpCas9 and DR-EMX1(1)-DR. Left and right panels are from 293FT cells transfected without or with SpRNase III, respectively. U6 indicate loading control blotted with a probe targeting human U6 snRNA. Transfection of the short tracrRNA expression construct led to abundant levels of the processed form of tracrRNA (~75 bp). Very low amounts of long tracrRNA are detected on the Northern blot.

[0160] To promote precise transcriptional initiation, the RNA polymerase III-based U6 promoter was selected to drive the expression of tracrRNA (FIG. **2**C). Similarly, a U6 promoter-based construct was developed to express a pre-crRNA array consisting of a single spacer flanked by two direct repeats (DRs, also encompassed by the term "tracrmate sequences"; FIG. **2**C). The initial spacer was designed to target a 33-base-pair (bp) target site (30-bp protospacer plus a 3-bp CRISPR motif (PAM) sequence satisfying the NGG recognition motif of Cas9) in the human EMX1 locus (FIG. **2**C), a key gene in the development of the cerebral cortex.

[0161] To test whether heterologous expression of the CRISPR system (SpCas9, SpRNase III, tracrRNA, and pre-crRNA) in mammalian cells can achieve targeted cleavage of mammalian chromosomes, HEK 293FT cells were transfected with combinations of CRISPR components. Since DSBs in mammalian nuclei are partially repaired by the non-homologous end joining (NHEJ) pathway, which leads to the formation of indels, the Surveyor assay was used to detect potential cleavage activity at the target EMX1 locus (see e.g. Guschin et al., 2010, Methods Mol Biol 649: 247). Co-transfection of all four CRISPR components was able to induce up to 5.0% cleavage in the protospacer (see FIG. **2**D). Co-transfection of all CRISPR components minus SpRNase III also induced up to 4.7% indel in the protospacer, suggesting that there may be endogenous mammalian RNases that are capable of assisting with crRNA maturation, such as for example the related Dicer and Drosha enzymes. Removing any of the remaining three components abolished the genome cleavage activity of the CRISPR system (FIG. **2**D). Sanger sequencing of amplicons containing the target locus verified the cleavage activity: in 43 sequenced clones, 5 mutated alleles (11.6%) were found. Similar experiments using a variety of guide sequences produced indel percentages as high as 29% (see FIGS. **4**-**8**, **10** and **11**). These results define a three-component system for efficient CRISPR-mediated genome modification in mammalian cells.

[0162] To optimize the cleavage efficiency, Applicants also tested whether different isoforms of tracrRNA affected the cleavage efficiency and found that, in this example system, only the short (89-bp) transcript form was able to mediate cleavage of the human EMX1 genomic locus. FIG. **9** provides an additional Northern blot analysis of crRNA processing in mammalian cells. FIG. **9**A illustrates a schematic showing the expression vector for a single spacer flanked by two direct repeats (DR-EMX1(1)-DR). The 30 bp spacer targeting the human EMX1 locus protospacer 1 and the direct repeat sequences are shown in the sequence beneath FIG. **9**A. The line indicates the region whose reverse-complement sequence was used to generate North-

ern blot probes for EMX1(1) crRNA detection. FIG. **9**B shows a Northern blot analysis of total RNA extracted from 293FT cells transfected with U6 expression constructs carrying DR-EMX1(1)-DR. Left and right panels are from 293FT cells transfected without or with SpRNase III respectively. DR-EMX1(1)-DR was processed into mature crRNAs only in the presence of SpCas9 and short tracrRNA and was not dependent on the presence of SpRNase III. The mature crRNA detected from transfected 293FT total RNA is ~33 bp and is shorter than the 39-42 bp mature crRNA from *S. pyogenes*. These results demonstrate that a CRISPR system can be transplanted into eukaryotic cells and reprogrammed to facilitate cleavage of endogenous mammalian target polynucleotides.

[0163] FIG. **2** illustrates the bacterial CRISPR system described in this example. FIG. **2**A illustrates a schematic showing the CRISPR locus 1 from *Streptococcus pyogenes* SF370 and a proposed mechanism of CRISPR-mediated DNA cleavage by this system. Mature crRNA processed from the direct repeat-spacer array directs Cas9 to genomic targets consisting of complimentary protospacers and a protospacer-adjacent motif (PAM). Upon target-spacer base pairing, Cas9 mediates a double-strand break in the target DNA. FIG. **2**B illustrates engineering of *S. pyogenes* Cas9 (SpCas9) and RNase III (SpRNase III) with nuclear localization signals (NLSs) to enable import into the mammalian nucleus. FIG. **2**C illustrates mammalian expression of SpCas9 and SpRNase III driven by the constitutive EF1a promoter and tracrRNA and pre-crRNA array (DR-Spacer-DR) driven by the RNA Pol3 promoter U6 to promote precise transcription initiation and termination. A protospacer from the human EMX1 locus with a satisfactory PAM sequence is used as the spacer in the pre-crRNA array. FIG. **2**D illustrates surveyor nuclease assay for SpCas9-mediated minor insertions and deletions. SpCas9 was expressed with and without SpRNase III, tracrRNA, and a pre-crRNA array carrying the EMX1-target spacer. FIG. **2**E illustrates a schematic representation of base pairing between target locus and EMX1-targeting crRNA, as well as an example chromatogram showing a micro deletion adjacent to the SpCas9 cleavage site. FIG. **2**F illustrates mutated alleles identified from sequencing analysis of 43 clonal amplicons showing a variety of micro insertions and deletions. Dashes indicate deleted bases, and non-aligned or mismatched bases indicate insertions or mutations. Scale bar=10 μm.

[0164] To further simplify the three-component system, a chimeric crRNA-tracrRNA hybrid design was adapted, where a mature crRNA (comprising a guide sequence) is fused to a partial tracrRNA via a stem-loop to mimic the natural crRNA:tracrRNA duplex (FIG. **3**A).

[0165] Guide sequences can be inserted between BbsI sites using annealed oligonucleotides. Protospacers on the sense and anti-sense strands are indicated above and below the DNA sequences, respectively. A modification rate of 6.3% and 0.75% was achieved for the human PVALB and mouse Th loci respectively, demonstrating the broad applicability of the CRISPR system in modifying different loci across multiple organisms While cleavage was only detected with one out of three spacers for each locus using the chimeric constructs, all target sequences were cleaved with efficiency of indel production reaching 27% when using the co-expressed pre-crRNA arrangement (FIGS. **4** and **5**).

[0166] FIG. **5** provides a further illustration that SpCas9 can be reprogrammed to target multiple genomic loci in mammalian cells. FIG. **5**A provides a schematic of the human EMX1 locus showing the location of five protospacers, indicated by the underlined sequences. FIG. **5**B provides a schematic of the pre-crRNA/trcrRNA complex showing hybridization between the direct repeat region of the pre-crRNA and tracrRNA (top), and a schematic of a chimeric RNA design comprising a 20 bp guide sequence, and tracr mate and tracr sequences consisting of partial direct repeat and tracrRNA sequences hybridized in a hairpin structure (bottom). Results of a Surveyor assay comparing the efficacy of Cas9-mediated cleavage at five protospacers in the human EMX1 locus is illustrated in FIG. **5**C. Each protospacer is targeted using either processed pre-crRNA/tracrRNA complex (crRNA) or chimeric RNA (chiRNA).

[0167] Since the secondary structure of RNA can be crucial for intermolecular interactions, a structure prediction algorithm based on minimum free energy and Boltzmann-weighted structure ensemble was used to compare the putative secondary structure of all guide sequences used in our genome targeting experiment (FIG. **3**B) (see e.g. Gruber et al., 2008, Nucleic Acids Research, 36: W70). Analysis revealed that in most cases, the effective guide sequences in the chimeric crRNA context were substantially free of secondary structure motifs, whereas the ineffective guide sequences were more likely to form internal secondary structures that could prevent base pairing with the target protospacer DNA. It is thus possible that variability in the spacer secondary structure might impact the efficiency of CRISPR-mediated interference when using a chimeric crRNA.

[0168] FIG. **3** illustrates example expression vectors. FIG. **3**A provides a schematic of a bi-cistronic vector for driving the expression of a synthetic crRNA-tracrRNA chimera (chimeric RNA) as well as SpCas9. The chimeric guide RNA contains a 20-bp guide sequence corresponding to the protospacer in the genomic target site. FIG. **3**B provides a schematic showing guide sequences targeting the human EMX1, PVALB, and mouse Th loci, as well as their predicted secondary structures. The modification efficiency at each target site is indicated below the RNA secondary structure drawing (EMX1, n=216 amplicon sequencing reads; PVALB, n=224 reads; Th, n=265 reads). The folding algorithm produced an output with each base colored according to its probability of assuming the predicted secondary structure, as indicated by a rainbow scale that is reproduced in FIG. **3**B in gray scale. Further vector designs for SpCas9 are shown in FIG. **3**A, including single expression vectors incorporating a U6 promoter linked to an insertion site for a guide oligo, and a Cbh promoter linked to SpCas9 coding sequence.

[0169] To test whether spacers containing secondary structures are able to function in prokaryotic cells where CRISPRs naturally operate, transformation interference of protospacer-bearing plasmids were tested in an *E. coli* strain heterologously expressing the *S. pyogenes* SF370 CRISPR locus 1 (FIG. **3**C). The CRISPR locus was cloned into a low-copy *E. coli* expression vector and the crRNA array was replaced with a single spacer flanked by a pair of DRs (pCRISPR). *E. coli* strains harboring different pCRISPR plasmids were transformed with challenge plasmids containing the corresponding protospacer and PAM sequences (FIG. **3**C). In the bacterial assay, all spacers facilitated efficient CRISPR interference (FIG. **3**C). These results sug-

gest that there may be additional factors affecting the efficiency of CRISPR activity in mammalian cells.

[0170] To investigate the specificity of CRISPR-mediated cleavage, the effect of single-nucleotide mutations in the guide sequence on protospacer cleavage in the mammalian genome was analyzed using a series of EMX1-targeting chimeric crRNAs with single point mutations (FIG. **4**A). FIG. **4**B illustrates results of a Surveyor nuclease assay comparing the cleavage efficiency of Cas9 when paired with different mutant chimeric RNAs. Single-base mismatch up to 12-bp 5' of the PAM substantially abrogated genomic cleavage by SpCas9, whereas spacers with mutations at farther upstream positions retained activity against the original protospacer target (FIG. **4**B). In addition to the PAM, SpCas9 has single-base specificity within the last 12-bp of the spacer. Furthermore, CRISPR is able to mediate genomic cleavage as efficiently as a pair of TALE nucleases (TALEN) targeting the same EMX1 protospacer. FIG. **4**C provides a schematic showing the design of TALENs targeting EMX1, and FIG. **4**D shows a Surveyor gel comparing the efficiency of TALEN and Cas9 (n=3).

[0171] Having established a set of components for achieving CRISPR-mediated gene editing in mammalian cells through the error-prone NHEJ mechanism, the ability of CRISPR to stimulate homologous recombination (HR), a high fidelity gene repair pathway for making precise edits in the genome, was tested. The wild type SpCas9 is able to mediate site-specific DSBs, which can be repaired through both NHEJ and HR. In addition, an aspartate-to-alanine substitution (D10A) in the RuvC I catalytic domain of SpCas9 was engineered to convert the nuclease into a nickase (SpCas9n; illustrated in FIG. **5**A) (see e.g. Sapranausaks et al., 2011, Nucleic Acid Research, 39: 9275; Gasiunas et al., 2012, Proc. Natl. Acad. Sci. USA, 109: E2579), such that nicked genomic DNA undergoes the high-fidelity homology-directed repair (HDR). Surveyor assay confirmed that SpCas9n does not generate indels at the EMX1 protospacer target. As illustrated in FIG. **5**B, co-expression of EMX1-targeting chimeric crRNA with SpCas9 produced indels in the target site, whereas co-expression with SpCas9n did not (n=3). Moreover, sequencing of 327 amplicons did not detect any indels induced by SpCas9n. The same locus was selected to test CRISPR-mediated HR by co-transfecting HEK 293FT cells with the chimeric RNA targeting EMX1, hSpCas9 or hSpCas9n, as well as a HR template to introduce a pair of restriction sites (HindIII and NheI) near the protospacer. FIG. **5**C provides a schematic illustration of the HR strategy, with relative locations of recombination points and primer annealing sequences (arrows). SpCas9 and SpCas9n indeed catalyzed integration of the HR template into the EMX1 locus. PCR amplification of the target region followed by restriction digest with HindIII revealed cleavage products corresponding to expected fragment sizes (arrows in restriction fragment length polymorphism gel analysis shown in FIG. **5**D), with SpCas9 and SpCas9n mediating similar levels of HR efficiencies. Applicants further verified HR using Sanger sequencing of genomic amplicons (FIG. **5**E). These results demonstrate the utility of CRISPR for facilitating targeted gene insertion in the mammalian genome. Given the 14-bp (12-bp from the spacer and 2-bp from the PAM) target specificity of the wild type SpCas9, the availability of a nickase can significantly reduce the likelihood of off-target modifications, since single strand breaks are not substrates for the error-prone NHEJ pathway.

[0172] Expression constructs mimicking the natural architecture of CRISPR loci with arrayed spacers (FIG. **2**A) were constructed to test the possibility of multiplexed sequence targeting. Using a single CRISPR array encoding a pair of EMX1- and PVALB-targeting spacers, efficient cleavage at both loci was detected (FIG. **4**F, showing both a schematic design of the crRNA array and a Surveyor blot showing efficient mediation of cleavage). Targeted deletion of larger genomic regions through concurrent DSBs using spacers against two targets within EMX1 spaced by 119 bp was also tested, and a 1.6% deletion efficacy (3 out of 182 amplicons; FIG. **5**G) was detected. This demonstrates that the CRISPR system can mediate multiplexed editing within a single genome.

Example 2: CRISPR System Modifications and Alternatives

[0173] The ability to use RNA to program sequence-specific DNA cleavage defines a new class of genome engineering tools for a variety of research and industrial applications. Several aspects of the CRISPR system can be further improved to increase the efficiency and versatility of CRISPR targeting. Optimal Cas9 activity may depend on the availability of free Mg^{2+} at levels higher than that present in the mammalian nucleus (see e.g. Jinek et al., 2012, Science, 337:816), and the preference for an NGG motif immediately downstream of the protospacer restricts the ability to target on average every 12-bp in the human genome. Some of these constraints can be overcome by exploring the diversity of CRISPR loci across the microbial metagenome (see e.g. Makarova et al., 2011, Nat Rev Microbiol, 9:467). Other CRISPR loci may be transplanted into the mammalian cellular milieu by a process similar to that described in Example 1. The modification efficiency at each target site is indicated below the RNA secondary structures. The algorithm generating the structures colors each base according to its probability of assuming the predicted secondary structure. RNA guide spacers 1 and 2 induced 14% and 6.4%, respectively. Statistical analysis of cleavage activity across biological replica at these two protospacer sites is also provided in FIG. **7**.

Example 3: Sample Target Sequence Selection Algorithm

[0174] A software program is designed to identify candidate CRISPR target sequences on both strands of an input DNA sequence based on desired guide sequence length and a CRISPR motif sequence (PAM) for a specified CRISPR enzyme. For example, target sites for Cas9 from *S. pyogenes*, with PAM sequences NGG, may be identified by searching for 5'-N_x-NGG-3' both on the input sequence and on the reverse-complement of the input. Likewise, target sites for Cas9 of *S. thermophilus* CRISPR1, with PAM sequence NNAGAAW, may be identified by searching for 5'-N_x-NNAGAAW-3' both on the input sequence and on the reverse-complement of the input. Likewise, target sites for Cas9 of *S. thermophilus* CRISPR3, with PAM sequence NGGNG, may be identified by searching for 5'-N_x-NGGNG-3' both on the input sequence and on the reverse-complement of the input. The value "x" in N_x may be fixed by the program or specified by the user, such as 20.

[0175] Since multiple occurrences in the genome of the DNA target site may lead to nonspecific genome editing, after identifying all potential sites, the program filters out sequences based on the number of times they appear in the relevant reference genome. For those CRISPR enzymes for which sequence specificity is determined by a 'seed' sequence, such as the 11-12 bp 5' from the PAM sequence, including the PAM sequence itself, the filtering step may be based on the seed sequence. Thus, to avoid editing at additional genomic loci, results are filtered based on the number of occurrences of the seed:PAM sequence in the relevant genome. The user may be allowed to choose the length of the seed sequence. The user may also be allowed to specify the number of occurrences of the seed:PAM sequence in a genome for purposes of passing the filter. The default is to screen for unique sequences. Filtration level is altered by changing both the length of the seed sequence and the number of occurrences of the sequence in the genome. The program may in addition or alternatively provide the sequence of a guide sequence complementary to the reported target sequence(s) by providing the reverse complement of the identified target sequence(s).

[0176] Further details of methods and algorithms to optimize sequence selection can be found found in U.S. application Ser. No. _____ (Broad Reference BI-2012/084 44790.11.2022); incorporated herein by reference.

Example 4: Evaluation of Multiple Chimeric crRNA-tracrRNA Hybrids

[0177] This example describes results obtained for chimeric RNAs (chiRNAs; comprising a guide sequence, a tracr mate sequence, and a tracr sequence in a single transcript) having tracr sequences that incorporate different lengths of wild-type tracrRNA sequence. FIG. **18***a* illustrates a schematic of a bicistronic expression vector for chimeric RNA and Cas9. Cas9 is driven by the CBh promoter and the chimeric RNA is driven by a U6 promoter. The chimeric guide RNA consists of a 20 bp guide sequence (Ns) joined to the tracr sequence (running from the first "U" of the lower strand to the end of the transcript), which is truncated at various positions as indicated. The guide and tracr sequences are separated by the tracr-mate sequence GUUUUAGAGCUA (SEQ ID NO: 28) followed by the loop sequence GAAA. Results of SURVEYOR assays for Cas9-mediated indels at the human EMX1 and PVALB loci are illustrated in FIGS. **18***b* and **18***c*, respectively. Arrows indicate the expected SURVEYOR fragments. ChiRNAs are indicated by their "+n" designation, and crRNA refers to a hybrid RNA where guide and tracr sequences are expressed as separate transcripts. Quantification of these results, performed in triplicate, are illustrated by histogram in FIGS. **11***a* and **11***b*, corresponding to FIGS. **10***b* and **10***c*, respectively ("N.D." indicates no indels detected). Protospacer IDs and their corresponding genomic target, protospacer sequence, PAM sequence, and strand location are provided in Table D. Guide sequences were designed to be complementary to the entire protospacer sequence in the case of separate transcripts in the hybrid system, or only to the underlined portion in the case of chimeric RNAs.

TABLE D

protospacer ID	genomic target	protospacer sequence (5' to 3')	PAM	Strand	SEQ ID NO:
1	EMX1	GGACATCGATGTCACCTCCAATGACTAGGG	TGG	+	29
2	EMX1	CATTGGAGGTGACATCGATGTCCTCCCCAT	TGG	-	30
3	EMX1	GGAAGGGCCTGAGTCCGAGCAGAAGAAGAA	GGG	+	31
4	PVALB	GGTGGCGAGAGGGGCCGAGATTGGGTGTTC	AGG	+	32
5	PVALB	ATGCAGGAGGGTGGCGAGAGGGGCCGAGAT	TGG	+	33

Cell Culture and Transfection

[0178] Human embryonic kidney (HEK) cell line 293FT (Life Technologies) was maintained in Dulbecco's modified Eagle's Medium (DMEM) supplemented with 10% fetal bovine serum (HyClone), 2 mM GlutaMAX (Life Technologies), 100U/mL penicillin, and 100 g/mL streptomycin at 37° C. with 5% C02 incubation. 293FT cells were seeded onto 24-well plates (Corning) 24 hours prior to transfection at a density of 150,000 cells per well. Cells were transfected using Lipofectamine 2000 (Life Technologies) following the manufacturer's recommended protocol. For each well of a 24-well plate, a total of 500 ng plasmid was used.

SURVEYOR Assay for Genome Modification

[0179] 293FT cells were transfected with plasmid DNA as described above. Cells were incubated at 37° C. for 72 hours post-transfection prior to genomic DNA extraction. Genomic DNA was extracted using the QuickExtract DNA Extraction Solution (Epicentre) following the manufacturer's protocol. Briefly, pelleted cells were resuspended in QuickExtract solution and incubated at 65° C. for 15 minutes and 98° C. for 10 minutes. The genomic region flanking the CRISPR target site for each gene was PCR amplified (primers listed in Table E), and products were purified using QiaQuick Spin Column (Qiagen) following the manufacturer's protocol. 400 ng total of the purified PCR products were mixed with 2 μl 10×Taq DNA Polymerase PCR buffer (Enzymatics) and ultrapure water to a final volume of 20 μl, and subjected to a re-annealing process to enable heteroduplex formation: 95° C. for 10 min, 95° C. to 85° C. ramping at −2° C./s, 85° C. to 25° C. at −0.25° C./s, and 25° C. hold for 1 minute. After re-annealing, products were treated with SURVEYOR nuclease and SURVEYOR enhancer S (Transgenomics) following the manufacturer's recommended protocol, and analyzed on 4-20% Novex TBE poly-acrylamide gels (Life Technologies). Gels were stained with SYBR Gold DNA stain (Life Technologies) for 30 minutes and imaged with a Gel Doc gel imaging system (Bio-rad). Quantification was based on relative band intensities.

TABLE E

primer name	genomic target	primer sequence (5' to 3')	SEQ ID NO
Sp-EMX1-F	EMX1	AAAACCACCCTTCTCTCTGGC	34
Sp-EMX1-R	EMX1	GGAGATTGGAGACACGGAGAG	35
Sp-PVALB-F	PVALB	CTGGAAAGCCAATGCCTGAC	36
Sp-PVALB-R	PVALB	GGCAGCAAACTCCTTGTCCT	37

Computational Identification of Unique CRISPR Target Sites

[0180] To identify unique target sites for the *S. pyogenes* SF370 Cas9 (SpCas9) enzyme in the human, mouse, rat, zebrafish, fruit fly, and *C. elegans* genome, we developed a software package to scan both strands of a DNA sequence and identify all possible SpCas9 target sites. For this example, each SpCas9 target site was operationally defined as a 20 bp sequence followed by an NGG protospacer adjacent motif (PAM) sequence, and we identified all sequences satisfying this 5'-N20-NGG-3' definition on all chromosomes. To prevent non-specific genome editing, after identifying all potential sites, all target sites were filtered based on the number of times they appear in the relevant reference genome. To take advantage of sequence specificity of Cas9 activity conferred by a 'seed' sequence, which can be, for example, approximately 11-12 bp sequence 5' from the PAM sequence, 5'-NNNNNNNNNN-NGG-3' sequences were selected to be unique in the relevant genome. All genomic sequences were downloaded from the UCSC Genome Browser (Human genome hg19, Mouse genome mm9, Rat genome rn5, Zebrafish genome danRer7, *D. melanogaster* genome dm4 and *C. elegans* genome ce10). The full search results are available to browse using UCSC Genome Browser information. An example visualization of some target sites in the human genome is provided in FIG. **22**.

[0181] Initially, three sites within the EMX1 locus in human HEK 293FT cells were targeted. Genome modification efficiency of each chiRNA was assessed using the SURVEYOR nuclease assay, which detects mutations resulting from DNA double-strand breaks (DSBs) and their subsequent repair by the non-homologous end joining (NHEJ) DNA damage repair pathway. Constructs designated chiRNA(+n) indicate that up to the +n nucleotide of wild-type tracrRNA is included in the chimeric RNA construct, with values of 48, 54, 67, and 85 used for n. Chimeric RNAs containing longer fragments of wild-type tracrRNA (chiRNA(+67) and chiRNA(+85)) mediated DNA cleavage at all three EMX1 target sites, with chiRNA(+85) in particular demonstrating significantly higher levels of DNA cleavage than the corresponding crRNA/tracrRNA hybrids that expressed guide and tracr sequences in separate transcripts (FIGS. **10***b* and **10***a*). Two sites in the PVALB locus that yielded no detectable cleavage using the hybrid system (guide sequence and tracr sequence expressed as separate transcripts) were also targeted using chiRNAs. chiRNA(+67) and chiRNA(+85) were able to mediate significant cleavage at the two PVALB protospacers (FIGS. **10***c* and **10***b*).

[0182] For all five targets in the EMX1 and PVALB loci, a consistent increase in genome modification efficiency with increasing tracr sequence length was observed. Without wishing to be bound by any theory, the secondary structure formed by the 3' end of the tracrRNA may play a role in enhancing the rate of CRISPR complex formation. An illustration of predicted secondary structures for each of the chimeric RNAs used in this example is provided in FIG. **21**. The secondary structure was predicted using RNAfold (http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi) using minimum free energy and partition function algorithm. Pseudocolor for each based (reproduced in grayscale) indicates the probability of pairing. Because chiRNAs with longer tracr sequences were able to cleave targets that were not cleaved by native CRISPR crRNA/tracrRNA hybrids, it is possible that chimeric RNA may be loaded onto Cas9 more efficiently than its native hybrid counterpart. To facilitate the application of Cas9 for site-specific genome editing in eukaryotic cells and organisms, all predicted unique target sites for the *S. pyogenes* Cas9 were computationally identified in the human, mouse, rat, zebra fish, *C. elegans*, and *D. melanogaster* genomes. Chimeric RNAs can be designed for Cas9 enzymes from other microbes to expand the target space of CRISPR RNA-programmable nucleases.

[0183] FIGS. **11** and **21** illustrate exemplary bicistronic expression vectors for expression of chimeric RNA including up to the +85 nucleotide of wild-type tracr RNA sequence, and SpCas9 with nuclear localization sequences. SpCas9 is expressed from a CBh promoter and terminated with the bGH polyA signal (bGH pA). The expanded sequence illustrated immediately below the schematic corresponds to the region surrounding the guide sequence insertion site, and includes, from 5' to 3', 3'-portion of the U6 promoter (first shaded region), BbsI cleavage sites (arrows), partial direct repeat (tracr mate sequence GTTTTAGAGCTA (SEQ ID NO: 38), underlined), loop sequence GAAA, and +85 tracr sequence (underlined sequence following loop sequence). An exemplary guide sequence insert is illustrated below the guide sequence insertion site, with nucleotides of the guide sequence for a selected target represented by an "N".

[0184] Sequences described in the above examples are as follows (polynucleotide sequences are 5' to 3'):

U6-short tracrRNA (*Streptococcus pyogenes* SF370):

(SEQ ID NO: 39)

GAGGGCCTATTTCCCATGATTCCTTCATATTTGCATATACGATACAAGGCTGTTAGAGAGATAATTGGAATTAATT
TGACTGTAAACACAAAGATATTAGTACAAAATACGTGACGTAGAAAGTAATAATTTCTTGGGTAGTTTGCAGTTTT
AAAATTATGTTTTAAAATGGACTATCATATGCTTACCGTAACTTGAAAGTATTTCGATTTCTTGGCTTTATATATC
TTGTGGAAAGGACGAAACACC**GGAACCATTCAAAACAGCATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTT**
GAAAAAGTGGCACCGAGTCGGTGCTTTTTTT

(bold = tracrRNA sequence; underline = terminator sequence)

-continued

U6-long tracrRNA (*Streptococcus pyogenes* SF370):

(SEQ ID NO: 40)

GAGGGCCTATTTCCCATGATTCCTTCATATTTGCATATACGATACAAGGCTGTTAGAGAGATAATTGGAATTAATT
TGACTGTAAACACAAAGATATTAGTACAAAATACGTGACGTAGAAAGTAATAATTTCTTGGGTAGTTTGCAGTTTT
AAAATTATGTTTTAAAATGGACTATCATATGCTTACCGTAACTTGAAAGTATTTCGATTTCTTGGCTTTATATATC
TTGTGGAAAGGACGAAACACCGGTAGTATTAAGTATTGTTTTATGGCTGATAAATTTCTTTGAATTTCTCCTTGAT
TATTTGTTATAAAAGTTATAAAATAATCTTGTTGGAACCATTCAAAACAGCATAGCAAGTTAAAATAAGGCTAGTC
CGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTTTT

U6-DR-BbsI backbone-DR (*Streptococcus pyogenes* SF370):

(SEQ ID NO: 41)

GAGGGCCTATTTCCCATGATTCCTTCATATTTGCATATACGATACAAGGCTGTTAGAGAGATAATTGGAATTAATT
TGACTGTAAACACAAAGATATTAGTACAAAATACGTGACGTAGAAAGTAATAATTTCTTGGGTAGTTTGCAGTTTT
AAAATTATGTTTTAAAATGGACTATCATATGCTTACCGTAACTTGAAAGTATTTCGATTTCTTGGCTTTATATATC
TTGTGGAAAGGACGAAACACCGGGTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAACGGGTCTTCGAGAAGACG
TTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAAC

U6-chimeric RNA-BbsI backbone (*Streptococcus pyogenes* SF370)

(SEQ ID NO: 42)

GAGGGCCTATTTCCCATGATTCCTTCATATTTGCATATACGATACAAGGCTGTTAGAGAGATAATTGGAATTAATT
TGACTGTAAACACAAAGATATTAGTACAAAATACGTGACGTAGAAAGTAATAATTTCTTGGGTAGTTTGCAGTTTT
AAAATTATGTTTTAAAATGGACTATCATATGCTTACCGTAACTTGAAAGTATTTCGATTTCTTGGCTTTATATATC
TTGTGGAAAGGACGAAACACCGGGTCTTCGAGAAGACCTGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTA
GTCCG

NLS-SpCas9-EGFP:

(SEQ ID NO: 43)

MDYKDHDGYKDHDIDYKDDDDKMAPKKKRKVGIHGVPAADKKYSIGLDIGTNSVGWAVITDEYKVPSKKFKVLGNT
DRHSIKKNLIGALLFDSGETAEATRLKRTARRYTRRKNRICYLQEIFSNEMAKVDDSFFHRLEESFLVEEDKKHE
RHPIFGNIVDEVAYHEKYPTIYHLRKKLVDSTDKADLRLIYLALAHMIKFRGHFLIEGDLNPDNSDVDKLFIQLVQ
TYNQLFEENPINASGVDAKAILSARLSKSRRLENLIAQLPGEKKNGLFGNLIALSLGLTPNFKSNFDLAEDAKLQL
SKDTYDDDLDNLLAQIGDQYADLFLAAKNLSDAILLSDILRVNTEITKAPLSASMIKRYDEHHQDLTLLKALVRQQ
LPEKYKEIFFDQSKNGYAGYIDGGASQEEFYKFIKPILEKMDGTEELLVKLNREDLLRKQRTFDNGSIPHQIHLGE
LHAILRRQEDFYPFLKDNREKIEKILTFRIPYYVGPLARGNSRFAWMTRKSEETITPWNFEEVVDKGASAQSFIER
MTNFDKNLPNEKVLPKHSLLYEYFTVYNELTKVKYVTEGMRKPAFLSGEQKKAIVDLLFKTNRKVTVKQLKEDYFK
KIECFDSVEISGVEDRFNASLGTYHDLLKIIKDKDFLDNEENEDILEDIVLTLTLFEDREMIEERLKTYAHLFDDK
VMKQLKRRRYTGWGRLSRKLINGIRDKQSGKTILDFLKSDGFANRNFMQLIHDDLSTFKEDIQKAQVSGQGDSLHE
HIANLAGSPAIKKGILQTVKVVDELVKVMGRHKPENIVIEMARENQTTQKGQKNSRERMKRIEEGIKELGSQILKE
HPVENTQLQNEKLYLYYLQNGRDMYVDQELDINRLSDYDVDHIVPQSFLKDDSIDNKVLTRSDKNRGKSDNVPEEV
VKKMKNYWRQLLNAKLITQRKFDNLTKAERGGLSELDKAGFIKRQLVETRQITKHVAQILDSRMNTKYDENDKLIR
EVKVITLKSKLVSDFRKDFQFYKVREINNYHHAHDAYLNAVVGTALIKKYPKLESEFVYGDYKVYDVRKMIAKSEQ
EIGKATAKYFFYSNIMNFFKTEITLANGEIRKRPLIETNGETGEIVWDKGRDFATVRKVLSMPQVNIVKKTEVQTG
GFSKESILPKRNSDKLIARKKDWDPKKYGGFDSPTVAYSVLVVAKVEKGKSKKLKSVKELLGITIMERSSFEKNPI
DFLEAKGYKEVKKDLIIKLPKYSLFELENGRKRMLASAGELQKGNELALPSKYVNFLYLASHYEKLKGSPEDNEQK
QLFVEQHKHYLDEIIEQISEFSKRVILADANLDKVLSAYNKHRDKPIREQAENIIHLFTLTNLGAPAAFKYFDTTI
DRKRYTSTKEVLDATLIHQSITGLYETRIDLSQLGGDAAAVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGD
ATYGKLTLKFICTTGKLPVPWPTLVTTLTYGVQCFSRYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEV
KGEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIEDGSVQLADHYQQNTPI
GDGPVLLPDNHYLSTQSALSKDPNEKRDHMVLLEFVTAAGITLGMDELYK

SpCas9-EGFP-NLS:

(SEQ ID NO: 44)

MDKKYSIGLDIGTNSVGWAVITDEYKVPSKKFKVLGNTDRHSIKKNLIGALLFDSGETAEATRLKRTARRYTRRK
NRICYLQEIFSNEMAKVDDSFFHRLEESFLVEEDKKHERHPIFGNIVDEVAYHEKYPTIYHLRKKLVDSTDKADLR
LIYLALAHMIKFRGHFLIEGDLNPDNSDVDKLFIQLVQTYNQLFEENPINASGVDAKAILSARLSKSRRLENLIAQ
LPGEKKNGLFGNLIALSLGLTPNFKSNFDLAEDAKLQLSKDTYDDDLDNLLAQIGDQYADLFLAAKNLSDAILLSD
ILRVNTEITKAPLSASMIKRYDEHHQDLTLLKALVRQQLPEKYKEIFFDQSKNGYAGYIDGGASQEEFYKFIKPIL
EKMDGTEELLVKLNREDLLRKQRTFDNGSIPHQIHLGELHAILRRQEDFYPFLKDNREKIEKILTFRIPYYVVGPL
ARGNSRFAWMTRKSEETITPWNFEEVVDKGASAQSFIERMTNFDKNLPNEKVLPKHSLLYEYFTVYNELTKVKYVT
EGMRKPAFLSGEQKKAIVDLLFKTNRKVTVKQLKEDYFKKIECFDSVEISGVEDRFNASLGTYHDLLKIIKDKDFL
DNEEDNEDILEDIVLTLTLFEDREMIEERLKTYAHLFDDKVMKQLKRRRYTGWGRLSRKLINGIRDKQSGKTILDF
LKSDGFANRNFMQLIHDDSLTFKEDIQKAQVSGQGDSLHEHIANLAGSPAIKKGILQTVKVVDELVKVMGRHKPEN
IVIEMARENQTTQKGQKNSRERMKRIEEGIKELGSQILKEHPVENTQLQNEKLYLYYLQNGRDMYVDQELDINRLS
DYDVDHIVPQSFLKKSIDNKVLTRSDKNRGKSDNVPSEEVVKKMKNYWRQLLNAKLITQRKFDNLTKAERGGLSEL
DKAGFIKRQLVETRQITKHVAQILDSRMNTKYDENDKLIREVKVITLKSKLVSDFRKDFQFYKVREINNYHHAHDA
YLNAVVGTALIKKYPKLESEFVYGDYKVYDVRMIAKSEQEIGKATAKYFFYSNIMNFFKTEITLANGEIRKRPLIE
TNGETGEIVWDKGRDFATVRKVLSMPQVNIVKKTEVQTGGFSKESILPKRNSDKLIARKKDWDPKKYGGFDSPTVA
YSVLVVAKEKGKSKKLKSVKELLGITIMERSSFEKNPIDFLEAKGYKEVKKDLIIKLPKYSLEFELENGRKRMLAS
AGELQKGNELALPSKYVNFLYLASHYEKLKGSPEDNEQKQLFVEQHKHYLDEIIEQISEFSKRVILADANLDKVLS
AYNKHRDKPIREQAENIIHLFTLTNLGAPAAFKYFDTTIDRKRYTSTKEVLDATLIHQSITGLYETRIDLSQLGGD
AAAVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPWPTLVTTLTYGVQCFSR
YPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHN
VYIMADKQKNGIKVNFKIRHNIEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSALSKDPNEKRDHMVLLEFVT
AAGITLGMDELYKKRPAATKKAGQAKKKK

-continued

NLS-SpCas9-EGFP-NLS:

(SEQ ID NO: 45)

MDYKDHDGYKDHDIDYKDDDDKMAPKKKRKVGIHGVPAADKKYSIGLDIGTNSVGWAVITDYEKVPSKKFKVLGNT
DRHSIKKNLIGALLFDSGETAEATRLKTARRYTRRKNRICYLQEIFSNEMAKVDDSFFHRLEESFLVEEDKKHER
HPIFGNIVDEVAYHEKYPTIYHLRKKLVDSTDKADLRLIYLALAHMIKFRGHFLIEGDLNPDNSDVDKLFIQLVQT
YNQLFEENPINASGVDAKAILSARLSKSRRLENLIAQLPGEKKNGLFGNLIALSLGLTPNFKSNFDLAEDAKLQLS
KDTYDDDLDNLLAQIGDQYADLFLAAKNLSDAILLSDIRLVNTEITKAPLSASMIKRYDEHHQDLTLLKALVRQQL
PEKYKEIFFDQSKNGYAGYIDGGASQEEFYKFIKPILEKMDGTEELLVKLNREDLLRKQRTFDNGSIPHQIHLGEL
HAILRRQEDFYPFLKDNREKIEKLILTFRIPYYVGPLARGNSRFAWMTRKSEETITPWNFEEVVDKGASAQSFIER
MTNFDKNLPNEKVLPKHSLLYEYFTVYNELTKVKYVTEGMRKPAFLSGEQKKAIVDLLFKTNRKVTVKQLKEDYFK
KIECFDSVEISGVEDRFNASLGTYHDLLKIIKDKDFLDNEENEDILEDIVLTLTLFEDREMIEERLKTYAHLFDDK
VMKQLKRRRYTGWGRLSRKLINGIRDKQSGKTILDFLKSDGFANRNFMQLIHDDSLTFKEDIQKAQVSGQGDSLHE
HIANLAGSPAIKKGILQTVKVVDELVKVMGRHKPENIVIEMARENQTTQKGQKNSRERMKRIEEGIKELGSQILKE
HPVENTQLQNEKLYLYYLQNGRDMYVDQELDINRLSDYDVDHIVPQSFLKDDSIDNKVLTRSDKNRGKSDNVPSEE
VVKKMKNYWRQLLNAKLITQRKFDNLTKAERGGLSELDKAGFIKRQLVETRQITKHVAQILDSRMNTKYDENDKLI
REVKVITLKSKLVSDFRKDFQFYKVREINNYHHAHDAYLNAVVGTALIKKYPKLESEFVYGDYKVYDVRKMIAKSE
QEIGKATAKYFFYSNIMNFFKTEITLANGEIRKRPLIETNGETGEIVWDKGRDFATVRKVLSMPQVNIVKKTEVQT
GGFSKESILPKRNSDKLIARKKDWDPKKYGGFDSPTVAYSVLVVAKVEKGKSKKLKSVKELLGITIMERSSFEKNP
IDFLEAKGYKEVKKDLIIKLPKYSLFELENGRKRMLASAGELQKGNELALPSKYVNFLYLASHYEKLKGSPEDNEQ
KQLFVEQHKHYLDEIIEQISEFSKRVILADANLDKVLSAYNKHRDKPIREQAENIIHLFTLTNLGAPAAFKYFDTT
IDRKRYTSTKEVLDATLIHQSITGLYETRIDLSQLGGDAAAVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEG
DATYGKLTLKFICTTGKLPVPWPTLVTTLTYGVQCFSRYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAE
VKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIEDGSVQLADHYQQNTP
IGDGPVLLPDNHYLSTQSALSKDPNEKRDHMVLLEFVTAAGITLGMDELYKKRPAATKKAGQAKKKK

NLS-SpCas9-NLS:

(SEQ ID NO: 46)

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHGVPAADKKYSIGLDIGTNSVGWAVITDEYKVPSKKFKVLGN
TDRHSIKKNLIGALLFDSGETAEATRLKRTARRYTRRKNRICYLQEIFSNEMAKVDDSFFHRLEESFLVEEDKKH
ERHPIFGNIVDEVAYHEKYPTIYHLRKKLVDSTDKADLRLIYLALAHMIKFRGHFLIEGDLNPDNSDVDKLFIQLV
QTYNQLFEENPINASGVDAKAILSARLSKSRRLENLIAQLPGEKKNGLFGNLIALSLGLTPNFKSNFDLAEDAKLQ
LSKDTYDDDLDNLLAQIGDQYADLFLAAKNLSDAILLSDIRLVNTEITKAPLSASMIKRYDEHHQDLTLLKALVR
QQLPEKYKEIFFDQSKNGYAGYIDGGASQEEFYKFIKPILEKMDGTEELLVKLNREDLLRKQRTFDNGSIPHQIHL
GELHAILRRQEDFYPFLKDNREKIEKILTFRIPYYVGPLARGNSRFAWMTRKSEETITPWNFEEVVDKGASAQSFI
ERMTNFDKNLPNEKVLPKHSLLYEYFTVYNELTKVKYVTEGMRKPAFLSGEQKKAIVDLLFKTNRKVTVKQLKEDY
FKKIECFDSVEISGVEDRFNASLGTYHDLLKIIKDKDFLDNEENEDILEKTILDFLKSDGFANRNFMQLIHDDSLT
FKEDIQKAQVSGQGDSLHEHIANLAGSPAIKKGILQTVKVVDELVKVMGRHKPENIVIEMARENQTTQKGQKNSRE
RMKRIEEGIKELGSQILKEHPVENTQLQNEKLYLYYLQNGRDMYVDQELDINRLSDYDVDHIVPQSFLKDDSIDNK
VLTRSDKNRGKSDNVPSEEVVKKMKNYWRQLLNAKLITQRKFDNLTKAERGGLSELDKAGFIKRQLVETRQITKHV
AQILDSRMNTKYDENDKLIREVKVITLKSKLVSDFRKDFQFYKVREINNYHHAHDAYLNAVVGTALIKKYPKLESE
FVYGDYKVYDVRKMIAKSEQEIGKATAKYFFYSNIMNFFKTEITLANGEIRKRPLIETNGETGEIVWDKGRDFATV
RKVLSMPQVNIVKKTEVQTGGFSKESILPKRNSDKLIARKKDWDPKKYGGFDSPTVAYSVLVVAKVEKGKSKKLKS
VKELLGITIMERSSFEKNPIDFLEAKGYKEVKKDLIIKLPKYSLFELENGRKRMLASAGELQKGNELALPSKYVNF
LYLASHYEKLKGSPEDNEQKQLFVEQHKHYLDEIIEQISEFSKRVILADANLDKVLSAYNKHRDKPIREQAENIIH
LFTLTNLGAPAAFKYFDTTIDRKRYTSTKEVLDATLIHQSITGLYETRIDLSQLGGDKRPAATKKAGQAKKKK

NLS-mCherry-SpRNase3:

(SEQ ID NO: 47)

MFLFSLTSFLSSSRTLVSKGEEDNMAIIKEFMRFKVHMEGSVNGHEFEIEGEGEGRPYEGTQTAKLKVTKGGPLPF
AWDILSPQFMYGSKAYVKHPADIPDYLKLSFPEGFKWERVMNFEDGGVVTVTQDSSLQDGEFIYKVKLRGTNFPSD
GPVMQKKTMGWEASSSERMYPEDGALKGEIKQRLKLKDGGHYDAEVKTTYKAKKPVQLPGAYNVNIKLDITSHNED
YTIVEQYERAEGRHSTGGMDELYKGSKQLEELLSTSFDIQFNDLTLLETAFTHTSYANEHRLLNVSHNERLEFLGD
AVLQLIISEYLFAKYPKKTEGDMSKLRSMIVREESLAGFSRFCSFDAYIKLGKGEEKSGGRRRDTILGDLFEAFLG
ALLLDKGIDAVRRFLKQVMIPQVEKGNFERVKDYKTCLQEFLQTKGDVAIDYQVISEKGPAHAKQFEVSIVVNGAV
LSKGLGKSKKALEQDAAKNALALQLSEV

SpRNase3-mCherry-NLS:

(SEQ ID NO: 48)

MKQLEELLSTSFDIQFNDLTLLETAFTHTSYANEHRLLNVSHNERLEFLGDAVLQLIISEYLFAKYPKKTEGDMSK
LRSMIVREESLAGFSRFCSFDAYIKLGKGEEKSGGRRRDTILGDLFEAFLGALLLDKGIDAVRRFLKQVMIPQVEK
GNFERVKDYKTCLQEFLQTKGDVAIDYQVISEKGPAHAKQFEVSIVVNGAVLSKGLGKSKKLAEQDAAKNALAQLS
EVGSVSKGEEDNMAIIKEFMRFKVHMEGSVNGHEFEIEGEGEGRPYEGTQTAKLKVTKGGPLPFAWDILSPQFMYG
SKAYVKHPADIPDYLKLSFPEGFKWERVMNFEDGGVVTVTQDSSLQDGEFIYKVKLRGTNFPSDGPVMQKKTMGWE
ASSERMYPEDGALKGEIKQRLKLKDGGHYDAEVKTTYKAKKPVQLPGAYNVNIKLDITSHNEDYTIVEQYERAEGR
HSTGGMDELYKKRPAATKKAGQAKKKK

NLS-SpCas9n-NLS (the D10A nickase mutation is lowercase):

(SEQ ID NO: 49)

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHGVPAADKKYSIGLaIGTNSVGWAVITDEYKVPSKKFKVLGN
TDRHSIKKNLIGALLFDSGETAEATRLKRTARRYTRRKNRICYLQEIFSNEMAKVDDSFFHRLEESFLVEEDKKH
ERHPIFGNIVDEVAYHEKYPTIYHLRKKLVDSTDKADLRLIYLALAHMIKFRGHFLIEGDLNPDNSDVDKLFIQLV
QTYNQLFEENPINASGVDAKAILSARLSKSRRLENLIAQLPGEKKNGLFGNLIALSLGLTPNFKSNFDLAEDAKLQ
LSKDTYDDDLDNLLAQIGDQYADLFLAAKNLSDAILLSDILRVNTEITKAPLSASMIKRYDEHHQDLTLLKALVRQ
QLPEKYKEIFFDQSKNGYAGYIDGGASQEEFYKFIKPILEKMDGTEELLVKLNREDLLRKQRTFDNGSIPHQIHLG
ELHAILRRQEDFYPFLKDNREKIEKILTFRIPYYVGPLARGNSRFAWMTRKSEETITPWNFEEVVDKGASAQSFIE
RMTNFDKNLPNEKVLPKHSLLYEYFTVYNELTKVKYVTEGMRKPAFLSGEQKKAIVDLLFKTNRKVTVKQLKEDYF
KKIECFDSVEISGVEDRFNASLGTYHDLLKIIKDKDFLDNEEDEDILEDIVLTLTLFEDREMIEERLKTYAHLFDD

-continued

KVMKQLKRRRYTGWGRLSRKLINGIRDKQSGKTILDFLKSDGFANRNFMQLIHDDSLTFKEDIQKAQVSGQGDSLH
EHIANLAGSPAIKKGILQTVKVVDELVKVMGRHKPENIVIEMARENQTTQKGQKSRERMKRIEEGIKELGSQILKE
HPVENTQLQNEKLYLYYLQNGRDMYVDQELDINRLSDYDVDHIVPQSFLKDDSIDNKVLTRSDKNRGKSDNVPSEE
VVKKMKNYWRQLLNAKLITQRKFDNLTKAERGGLSELDKAGFIKRQLVETRQITKHVAQILDSRMNTKYDENDKLI
REVKVITLKSKLVSDFRKDFQFYKVREINNYHHAHDAYLNAVVGTALIKKYPKLESEFVYGDYKVYDVRKMIAKSE
QEIGKATAKYFFYSNIMNFFKTEITLANGEIRKRPLIETNGETGEIVWDKGRDFATVRKVLSMPQVNIVKKTEVQT
GGFSKESILPKRNSDKLIARKKDWDPKKYGGFDSPTVAYSVLVVAKVEKGKSKKLKSVKELLGITIMERSSFEKNP
IDFLEAKGYKEVKKDLIIKLPKYSLFELENGRKRMLASAGELQKGNELALPSKYVNFLYLASHYEKLKGSPEDNEQ
KQLFVEQHKHYLDEIIEQISEFSKRVILADANLDKVLSAYNKHRDKPIREQAENIIHLFTLTNLGAPAAFKYFDTT
IDRKRYTSTKEVLDATLIHQSITGLYETRIDLSQLGGDKRPAATKKAGQAKKKK

hEMX1-HR Template-HindIII-NheI:

(SEQ ID NO: 50)

GAATGCTGCCCTCAGACCCGCTTCCTCCCTGTCCTTGTCTGTCCAAGGAGAATGAGGTCTCACTGGTGGATTTCGG
ACTACCCTGAGGAGCTGGCACCTGAGGGACAAGGCCCCCCACCTGCCCAGCTCCAGCCTCTGATGAGGGGTGGGAG
AGAGCTACATGAGGTTGCTAAGAAAGCCTCCCCTGAAGGAGACCACACAGTGTGTGAGGTTGGAGTCTCTAGCAGC
GGGTTCTGTGCCCCCAGGGATAGTCTGGCTGTCCAGGCACTGCTCTTGATATAAACACCACCTCCTAGTTATGAAA
CCATGCCCATTCTGCCTCTCTGTATGGAAAAGAGCATGGGGCTGGCCCGTGGGGTGGTGTCCACTTTAGGCCCTGT
GGGAGATCATGGGAACCCACGCAGTGGGTCATAGGCTCTCTCATTTACTACTCACATCCACTCTGTGAAGAAGCGA
TTATGATCTCTCCTCTAGAAACTCGTAGAGTCCCATGTCTGCCGGCTTCCAGAGCCTGCACTCCTCCACCTTGGCT
TGGCTTTGCTGGGGCTAGAGGAGCTAGGATGCACAGCAGCTCTGTGACCCTTTGTTTGAGAGGAACAGGAAAACCA
CCCTTCTCTCTGGCCCACTGTGTCCTCTTCCTGCCCTGCCATCCCCTTCTGTGAATGTTAGACCCATGGGAGCAGC
TGGTCAGAGGGGACCCCGGCCTGGGGCCCCTAACCCTATGTAGCCTCAGTCTTCCCATCAGGCTCTCAGCTCAGCC
TGAGTGTTGAGGCCCCAGTGGCTGCTCTGGGGGCCTCCTGAGTTTCTCATCTGTGCCCCTCCCTCCCTGGCCCAGG
TGAAGGTGTGGTTCCAGAACCGGAGGACAAAGTACAAACGGCAGAAGCTGGAGGAGGAAGGGCCTGAGTCCGAGCA
GAAGAAGAAGGGCTCCCATCACATCAACCGGTGGCGCATTGCCACGAAGCAGGCCAATGGGGAGGACATCGATGTC
ACCTCCAATGACaagcttgctagcGGTGGGCAACCACAAACCCACGAGGGCAGAGTGCTGCTTGCTGCTGGCCAGG
CCCCTGCGTGGGCCCAAGCTGGACTCTGGCCACTCCCTGGCCAGGCTTTGGGGAGGCCTGGAGTCATGGCCCCACA
GGGCTTGAAGCCCGGGGCCGCCATTGACAGAGGGACAAGCAATGGGCTGGCTGAGGCCTGGGACCACTTGGCCTTC
TCCTCGGAGAGCCTGCCTGCCTGGGCGGGCCCGCCCGCCACCGCAGCCTCCCAGCTGCTCTCCGTGTCTCCAATCT
CCCTTTTGTTTTGATGCATTTCTGTTTTAATTTATTTTCCAGGCACCACTGTAGTTTAGTGATCCCCAGTGTCCCC
CTTCCCTATGGGAATAATAAAAGTCTCTCTCTTAATGACACGGGCATCCAGCTCCAGCCCCAGAGCCTGGGGTGGT
AGATTCCGGCTCTGAGGGCCAGTGGGGGCTGGTAGAGCAAACGCGTTCAGGGCCTGGGAGCCTGGGGTGGGGTACT
GGTGGAGGGGGTCAAGGGTAATTCATTAACTCCTCTCTTTTGTTGGGGGACCCTGGTCTCTACCTCCAGCTCCACA
GCAGGAGAAACAGGCTAGACATAGGGAAGGGCCATCCTGTATCTTGAGGGAGGACAGGCCCAGGTCTTTCTTAACG
TATTGAGAGGTGGGAATCAGGCCCAGGTAGTTCAATGGGAGAGGGAGAGTGCTTCCCTCTGCCTAGAGACTCTGGT
GGCTTCTCCAGTTGAGGAGAAACCAGAGGAAAGGGGAGGATTGGGGTCTGGGGGAGGGAACACCATTCACAAAGGC
TGACGGTTCCAGTCCGAAGTCGTGGGCCCACCAGGATGCTCACCTGTCCTTGGAGAACCGCTGGGCAGGTTGAGAC
TGCAGAGACAGGGCTTAAGGCTGAGCCTGCAACCAGTCCCCAGTGACTCAGGGCCTCCTCAGCCCAAGAAAGAGCA
ACGTGCCAGGGCCCGCTGAGCTCTTGTGTTCACCTG

NLS-StCsn1-NLS:

(SEQ ID NO: 51)

MKRPAATKKAGQAKKKKSDLVLGLDIGIGSVGVGILNKVTGEIIHKNSRIFPAAQAENNLVRRTNRQGRRLARRKK
HRRVRLNRLFEESGLITDFTKISINLNPYQLRVKGLTDELSNEELFIALKNMVKHRGISYLDDASDDGNSSVGDYA
QIVKENSKQLETKTPGQIQLERYQTYGQLRGDFTVEKDGKKHRLINVFPTSAYRSEALRILQTQQEFNPQITDEFI
NRYLEILTGKRKYYHGPGNEKSRTDYGRYRTSGETLDNIFGILIGKCRFYPDEFRAAKASYTAQEFNLLNDLNNLT
VPTETKKLSKEQKNQIINYVKNEKAMGPAKLFKYIAKLLSCDVADIKGYRIDKSGKAEIHTFEAYRKMKTLETLDI
EQMDRETLDKLAYVLTLNTEREGIQEALEHEFADGSFSQKQVDELVQFRKANSSIFGKGWHNFSVKLMMELIPELY
ETSEEQMTILTRLGKQKTTSSSNKTKYIDEKLLTEEIYNPVVAKSVRQAIKIVNAAIKEYGDFDNIVIEMARETNE
DDEKKAIQKIQKANKDEKDAAMLKAANQYNGKAELPHSVFHGHKQLATKIRLWHQQGERCLYTGKTISIHDLINNS
QFEVDHILPLSITFDDSLANKVLVYATANQEKGQRTPYQALDSMDDAWSFRELKAFVRESKTLSNKKKEYLLTEED
ISKFDVRKKFIERNLVDTRYASRVVLNALQEHFRAHKIDTKVSVVRGQFTSQLRRHWGIEKTRDTYHHHAVDALII
AASSQLNLWKKQKNTLVSYSEDQLLDIETGELISDDEYKESVFKAPYQHFVDTLKSKEFEDSILFSYQVDSKFNRK
ISDATIYATRQAKVGKDKADETYVLGKIKDIYTQDGYDAFMKIYKKDKSKFLMYRHDPQTFEKVIEPILENYPNKQ
INEKGKEVPCNPFLKYKEEHGYIRKYSKKGNGPEIKSLKYYDSKLGNHIDITPKDSNNKVVLQSVSPWRADVYFNK
TTGKYEILGLKYADLQFEKGTGTYKISQEKYNDIKKKEGVDSDSEFKFTLYKNDLLLVKDTETKEQQLFRFLSRTM
PKQKHYVELKPYDKQKFEGGEALIKVLGNVANSGQCKKGLGKSNISIYKVRTDVLGNQHIIKNEGDKPKLDFKRPA
ATKKAGQAKKKK

U6-St_tracrRNA(7-97):

(SEQ ID NO: 52)

GAGGGCCTATTTCCCATGATTCCTTCATATTTGCATATACGATACAAGGCTGTTAGAGAGATAATTGGAATTAATT
TGACTGTAAACACAAAGATATTAGTACAAAATACGTGACGTAGAAAGTAATAATTTCTTGGGTAGTTTGCAGTTTT
AAAATTATGTTTTAAAATGGACTATCATATGCTTACCGTAACTTGAAAGTATTTCGATTTCTTGGCTTTATATATC
TTGTGGAAAGGACGAAACACCGTTACTTAAATCTTGCAGAAGCTACAAAGATAAGGCTTCATGCCGAAATCAACAC
CCTGTCATTTTATGGCAGGGTGTTTTCGTTATTTAA

U6-DR-spacer-DR (*S. pyogenes* SF370)

(SEQ ID NO: 53)

gagggcctatttcccatgattccttcatatttgcatatacgatacaaggctgttagagagataattggaattaatt
tgactgtaaacacaaagatattagtacaaaatacgtgacgtagaaagtaataatttcttgggtagtttgcagtttt
aaaattatgttttaaaatggactatcatatgcttaccgtaacttgaaagtatttcgatttcttggctttatatatc
ttgtggaaaggacgaaacaccgggttttagagctatgctgttttgaatggtcccaaaacNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNgttttagagctatgctgttttgaatggtcccaaaac**TTTTTTT**

(lowercase underline = direct repeat; N = guide sequence; bold = terminator)

-continued

Chimeric RNA containing +48 tracr RNA (*S. pyogenes* SF370)

(SEQ ID NO: 54)

gagggcctatttcccatgattccttcatatttgcatatacgatacaaggctgttagagagataattggaattaatt
tgactgtaaacacaaagatattagtacaaaatacgtgacgtagaaagtaataatttcttgggtagtttgcagtttt
aaaattatgttttaaaatggactatcatatgcttaccgtaacttgaaagtatttcgatttcttggctttatatatc
ttgtggaaaggacgaaacaccNNNNNNNNNNNNNNNNNNNNgttttagagctagaaatagcaagttaaaataaggc
tagtccgTTTTTTT

(N = guide sequence; first underline = tracr mate sequence; second underline = tracr sequence; bold = terminator)

Chimeric RNA containing +54 tracr RNA (*S. pyogenes* SF370)

(SEQ ID NO: 55)

gagggcctatttcccatgattccttcatatttgcatatacgatacaaggctgttagagagataattggaattaatt
tgactgtaaacacaaagatattagtacaaaatacgtgacgtagaaagtaataatttcttgggtagtttgcagtttt
aaaattatgttttaaaatggactatcatatgcttaccgtaacttgaaagtatttcgatttcttggctttatatatc
ttgtggaaaggacgaaacaccNNNNNNNNNNNNNNNNNNNNgttttagagctagaaatagcaagttaaaataaggc
tagtccgttatcaTTTTTTT

(N = guide sequence; first underline = tracr mate sequence; second underline = tracr sequence; bold = terminator)

Chimeric RNA containing +67 tracr RNA (*S. pyogenes* SF370)

(SEQ ID NO: 56)

gagggcctatttcccatgattccttcatatttgcatatacgatacaaggctgttagagagataattggaattaatt
tgactgtaaacacaaagatattagtacaaaatacgtgacgtagaaagtaataatttcttgggtagtttgcagtttt
aaaattatgttttaaaatggactatcatatgcttaccgtaacttgaaagtatttcgatttcttggctttatatatc
ttgtggaaaggacgaaacaccNNNNNNNNNNNNNNNNNNNNgttttagagctagaaatagcaagttaaaataaggc
tagtccgttatcaacttgaaaaagtgTTTTTTT

(N = guide sequence; first underline = tracr mate sequence; second underline = tracr sequence; bold = terminator)

Chimeric RNA containing +85 tracr RNA (*S. pyogenes* SF370)

(SEQ ID NO: 57)

gagggcctatttcccatgattccttcatatttgcatatacgatacaaggctgttagagagataattggaattaatt
tgactgtaaacacaaagatattagtacaaaatacgtgacgtagaaagtaataatttcttgggtagtttgcagtttt
aaaattatgttttaaaatggactatcatatgcttaccgtaacttgaaagtatttcgatttcttggctttatatatc
ttgtggaaaggacgaaacaccNNNNNNNNNNNNNNNNNNNNgttttagagctagaaatagcaagttaaaataaggc
tagtccgttatcaacttgaaaaagtggcaccgagtcggtgcTTTTTTT

(N = guide sequence; first underline = tracr mate sequence; second underline = tracr sequence; bold = terminator)

CBh-NLS-SpCas9-NLS

(SEQ ID NO: 58)

CGTTACATAACTTACGGTAAATGGCCCGCCTGGCTGACCGCCCAACGACCCCCGCCCATTGACGTCAATAATGACG
TATGTTCCCATAGTAACGCCAATAGGGACTTTCCATTGACGTCAATGGGTGGAGTATTTACGGTAAACTGCCCACT
TGGCAGTACATCAAGTGTATCATATGCCAAGTACGCCCCCTATTGACGTCAATGACGGTAAATGGCCCGCCTGGCA
TTATGCCCAGTACATGACCTTATGGGACTTTCCTACTTGGCAGTACATCTACGTATTAGTCATCGCTATTACCATG
GTCGAGGTGAGCCCCACGTTCTGCTTCACTCTCCCCATCTCCCCCCCCTCCCCACCCCCAATTTTGTATTTATTTA
TTTTTTAATTATTTTGTGCAGCGATGGGGGCGGGGGGGGGGGGGGGCGCGCGCCAGGCGGGGCGGGGCGGGGCGA
GGGGCGGGGCGGGGCGAGGCGGAGAGGTGCGGCGGCAGCCAATCAGAGCGGCGCGCAGCGCGCGGCGGGCGGGAGT
CGCTGCGACGCTGCCTTCGCCCCGTGCCCCGCTCCGCCGCCGCCTCGCGCCGCCCGCCCCGGCTCTGACTGACCGC
GTTACTCCCACAGGTGAGCGGGCGGGACGGCCCTTCTCCTCCGGGCTGTAATTAGCTGAGCAAGAGGTAAGGGTTT
AAGGGATGGTTGGTTGGTGGGGTATTAATGTTTAATTACCTGGAGCACCTGCCTGAAATCACTTTTTTTCAGGTTG
Gaccggtgccacc**ATG**GACTATAAGGACCACGACGGAGACTACAAGGATCATGATATTGATTACAAAGACGATGAC
GATAAGATGGCCCCAAAGAAGAAGCGGAAGGTCGGTATCCACGGAGTCCCAGCAGCCGACAAGAAGTACAGCATCG
GCCTGGACATCGGCACCAACTCTGTGGGCTGGGCCGTGATCACCGACGAGTACAAGGTGCCCAGCAAGAAATTCAA
GGTGCTGGGCAACACCGACCGGCACAGCATCAAGAAGAACCTGATCGGAGCCCTGCTGTTCGACAGCGGCGAAACA
GCCGAGGCCACCCGGCTGAAGAGAACCGCCAGAAGAAGATACACCAGACGGAAGAACCGGATCTGCTATCTGCAAG
AGATCTTCAGCAACGAGATGGCCAAGGTGGACGACAGCTTCTTCCACAGACTGGAAGAGTCCTTCCTGGTGGAAGA
GGATAAGAAGCACGAGCGGCACCCCATCTTCGGCAACATCGTGGACGAGGTGGCCTACCACGAGAAGTACCCCACC
ATCTACCACCTGAGAAAGAAACTGGTGGACAGCACCGACAAGGCCGACCTGCGGCTGATCTATCTGGCCCTGGCCC
ACATGATCAAGTTCCGGGGCCACTTCCTGATCGAGGGCGACCTGAACCCCGACAACAGCGACGTGGACAAGCTGTT
CATCCAGCTGGTGCAGACCTACAACCAGCTGTTCGAGGAAAACCCCATCAACGCCAGCGGCGTGGACGCCAAGGCC
ATCCTGTCTGCCAGACTGAGCAAGAGCAGACGGCTGGAAAATCTGATCGCCCAGCTGCCCGGCGAGAAGAAGAATG
GCCTGTTCGGCAACCTGATTGCCCTGAGCCTGGGCCTGACCCCCAACTTCAAGAGCAACTTCGACCTGGCCGAGGA
TGCCAAACTGCAGCTGAGCAAGGACACCTACGACGACGACCTGGACAACCTGCTGGCCCAGATCGGCGACCAGTAC
GCCGACCTGTTTCTGGCCGCCAAGAACCTGTCCGACGCCATCCTGCTGAGCGACATCCTGAGAGTGAACACCGAGA
TCACCAAGGCCCCCCTGAGCGCCTCTATGATCAAGAGATACGACGAGCACCACCAGGACCTGACCCTGCTGAAAGC
TCTCGTGCGGCAGCAGCTGCCTGAGAAGTACAAAGAGATTTTCTTCGACCAGAGCAAGAACGGCTACGCCGGCTAC
ATTGACGGCGGAGCCAGCCAGGAAGAGTTCTACAAGTTCATCAAGCCCATCCTGGAAAAGATGGACGGCACCGAGG
AACTGCTCGTGAAGCTGAACAGAGAGGACCTGCTGCGGAAGCAGCGGACCTTCGACAACGGCAGCATCCCCCACCA
GATCCACCTGGGAGAGCTGCACGCCATTCTGCGGCGGCAGGAAGATTTTTACCCATTCCTGAAGGACAACCGGGAA
AAGATCGAGAAGATCCTGACCTTCCGCATCCCCTACTACGTGGGCCCTCTGGCCAGGGGAAACAGCAGATTCGCCT
GGATGACCAGAAAGAGCGAGGAAACCATCACCCCCTGGAACTTCGAGGAAGTGGTGGACAAGGGCGCTTCCGCCCA
GAGCTTCATCGAGCGGATGACCAACTTCGATAAGAACCTGCCCAACGAGAAGGTGCTGCCCAAGCACAGCCTGCTG
TACGAGTACTTCACCGTGTATAACGAGCTGACCAAAGTGAAATACGTGACCGAGGGAATGAGAAAGCCCGCCTTCC
TGAGCGGCGAGCAGAAAAAGGCCATCGTGGACCTGCTGTTCAAGACCAACCGGAAAGTGACCGTGAAGCAGCTGAA
AGAGGACTACTTCAAGAAAATCGAGTGCTTCGACTCCGTGGAAATCTCCGGCGTGGAAGATCGGTTCAACGCCTCC

-continued

CTGGGCACATACCACGATCTGCTAAAATTATCAAGGACAAGGACTTCCTGGACAATGAGGAAAACGAGGACATTCT
GGAAGATATCGTGCTGACCCTGACACTGTTTGAGGACAGAGAGATGATCGAGGAACGGCTGAAAACCTATGCCCAC
CTGTTCGACGACAAAGTGATGAAGCAGCTGAAGCGGCGGAGATACACCGGCTGGGGCAGGCTGAGCCGGAAGCTGA
TCAACGGCATCCGGGACAAGCAGTCCGGCAAGACAATCCTGGATTTCCTGAAGTCCGACGGCTTCGCCAACAGAAA
CTTCATGCAGCTGATCCACGACGACAGCCTGACCTTTAAAGAGGACATCCAGAAAGCCCAGGTGTCCGGCCAGGGC
GATAGCCTGCACGAGCACATTGCCAATCTGGCCGGCAGCCCCGCCATTAAGAAGGGCATCCTGCAGACAGTGAAGG
TGGTGGACGAGCTCGTGAAAGTGATGGGCCGGCACAAGCCCGAGAACATCGTGATCGAAATGGCCAGAGAGAACCA
GACCACCCAGAAGGGACAGAAGAACAGCCGCGAGAGAATGAAGCGGATCGAAGAGGGCATCAAAGAGCTGGGCAGC
CAGATCCTGAAAGAACACCCCGTGGAAAACACCCAGCTGCAGAACGAGAAGCTGTACCTGTACTACCTGCAGAATG
GGCGGGATATGTACGTGGACCAGGAACTGGACATCAACCGGCTGTCCGACTACGATGTGGACCATATCGTGCCTCA
GAGCTTTCTGAAGGACGACTCCATCGACAACAAGGTGCTGACCAGAAGCGACAAGAACCGGGGCAAGAGCGACAAC
GTGCCCTCCGAAGAGGTCGTGAAGAAGATGAAGAACTACTGGCGGCAGCTGCTGAACGCCAAGCTGATTACCCAGA
GAAAGTTCGACAATCTGACCAAGGCCGAGAGAGGCGGCCTGAGCGAACTGGATAAGGCCGGCTTCATCAAGAGACA
GCTGGTGGAAACCCGGCAGATCACAAAGCACGTGGCACAGATCCTGGACTCCCGGATGAACACTAAGTACGACGAG
AATGACAAGCTGATCCGGGAAGTGAAAGTGATCACCCTGAAGTCCAAGCTGGTGTCCGATTTCCGGAAGGATTTCC
AGTTTTACAAAGTGCGCGAGATCAACAACTACCACCACGCCCACGACGCCTACCTGAACGCCGTCGTGGGAACCGC
CTGATCAAAAAGTACCCTAAGCTGGAAAGCGAGTTCGTGTACGGCGACTACAAGGTGTACGACGTGCGGAAGATGA
TCGCCAAGAGCGAGCAGGAAATCGGCAAGGCTACCGCCAAGTACTTCTTCTACAGCAACATCATGAACTTTTTCAA
GACCGAGATTACCCTGGCCAACGGCGAGATCCGGAAGCGGCCTCTGATCGAGACAAACGGCGAAACCGGGGAGATC
GTGTGGGATAAGGGCCGGGATTTTGCCACCGTGCGGAAAGTGCTGAGCATGCCCCAAGTGAATATCGTGAAAAAGA
CCGAGGTGCAGACAGGCGGCTTCAGCAAAGAGTCTATCCTGCCCAAGAGGAACAGCGATAAGCTGATCGCCAGAAA
GAAGGACTGGGACCCTAAGAAGTACGGCGGCTTCGACAGCCCCACCGTGGCCTATTCTGTGCTGGTGGTGGCCAAA
GTGGAAAAGGGCAAGTCCAAGAAACTGAAGAGTGTGAAAGAGCTGCTGGGGATCACCATCATGGAAAGAAGCAGCT
TCGAGAAGAATCCCATCGACTTTCTGGAAGCCAAGGGCTACAAAGAAGTGAAAAAGGACCTGATCATCAAGCTGCC
TAAGTACTCCCTGTTCGAGCTGGAAAACGGCCGGAAGAGAATGCTGGCCTCTGCCGGCGAACTGCAGAAGGGAAAC
GAACTGGCCCTGCCCTCCAAATATGTGAACTTCCTGTACCTGGCCAGCCACTATGAGAAGCTGAAGGGCTCCCCCG
AGGATAATGAGCAGAAACAGCTGTTTGTGGAACAGCACAAGCACTACCTGGACGAGATCATCGAGCAGATCAGCGA
GTTCTCCAAGAGAGTGATCCTGGCCGACGCTAATCTGGACAAAGTGCTGTCCGCCTACAACAAGCACCGGGATAAG
CCCATCAGAGAGCAGGCCGAGAATATCATCCACCTGTTTACCCTGACCAATCTGGGAGCCCCTGCCGCCTTCAAGT
ACTTTGACACCACCATCGACCGGAAGAGGTACACCAGCACCAAAGAGGTGCTGGACGCCACCCTGATCCACCAGAG
CATCACCGGCCTGTACGAGACACGGATCGACCTGTCTCAGCTGGGAGGCGACTTTCTTTTTCTTAGCTTGACCAGC
TTTCTTAGTAGCAGGACGCTTTAA

(underline = NLS-hSpCas9-NLS)

Example chimeric RNA for *S. thermophilus* LMD-9 CRISPR1 Cas9
(with PAM of NNAGAAW)

(SEQ ID NO: 59)

NNNNNNNNNNNNNNNNNNNNgtttttgtactctcaagatttaGAAAtaaatcttgcagaagctacaaagataaggc
ttcatgccgaaatcaacaccctgtcattttatggcagggtgttttcgttatttaa**TTTTTT**

(N = guide sequence; first underline = tracr mate sequence; second
underline = tracr sequence; bold = terminator)

Example chimeric RNA for *S. thermophilus* LMD-9 CRISPR1 Cas9
(with PAM of NNAGAAW)

(SEQ ID NO: 60)

NNNNNNNNNNNNNNNNNNNNgtttttgtactctcaGAAAtgcagaagctacaaagataaggcttcatgccgaaatc
aacaccctgtcattttatggcagggtgttttcgttatttaa**TTTTTT**

(N = guide sequence; first underline = tracr mate sequence; second
underline = tracr sequence; bold = terminator)

Example chimeric RNA for *S. thermophilus* LMD-9 CRISPR1 Cas9
(with PAM of NNAGAAW)

(SEQ ID NO: 61)

NNNNNNNNNNNNNNNNNNNNgtttttgtactctcaGAAAtgcagaagctacaaagataaggcttcatgccgaaatc
aacaccctgtcattttatggcagggtgt**TTTTTT**

(N = guide sequence; first underline = tracr mate sequence; second
underline = tracr sequence; bold = terminator)

Example chimeric RNA for *S. thermophilus* LMD-9 CRISPR1 Cas9
(with PAM of NNAGAAW)

(SEQ ID NO: 62)

NNNNNNNNNNNNNNNNNNNNgttattgtactctcaagatttaGAAAtaaatcttgcagaagctacaaagataaggc
ttcatgccgaaatcaacaccctgtcattttatggcagggtgttttcgttatttaa**TTTTTT**

(N = guide sequence; first underline = tracr mate sequence; second
underline = tracr sequence; bold = terminator)

Example chimeric RNA for *S. thermophilus* LMD-9 CRISPR1 Cas9
(with PAM of NNAGAAW)

(SEQ ID NO: 63)

NNNNNNNNNNNNNNNNNNNNgttattgtactctcaGAAAtgcagaagctacaaagataaggcttcatgccgaaatc
aacaccctgtcattttatggcagggtgttttcgttatttaa**TTTTTT**

(N = guide sequence; first underline = tracr mate sequence; second
underline = tracr sequence; bold = terminator)

-continued

Example chimeric RNA for *S. thermophilus* LMD-9 CRISPR1 Cas9
(with PAM of NNAGAAW)

(SEQ ID NO: 64)

NNNNNNNNNNNNNNNNNNNNgttattgtactctcaGAAAtgcagaagctacaaagataaggcttcatgccgaaatc
aacaccctgtcattttatggcagggtgt**TTTTTT**

(N = guide sequence; first underline = tracr mate sequence; second
underline = tracr sequence; bold = terminator)

Example chimeric RNA for *S. thermophilus* LMD-9 CRISPR1 Cas9
(with PAM of NNAGAAW)

(SEQ ID NO: 65)

NNNNNNNNNNNNNNNNNNNNgttattgtactctcaagatttaGAAAtaaatcttgcagaagctacaatgataaggc
ttcatgccgaaatcaacaccctgtcattttatggcagggtgttttcgttatttaa**TTTTTT**

(N = guide sequence; first underline = tracr mate sequence; second
underline = tracr sequence; bold = terminator)

Example chimeric RNA for *S. thermophilus* LMD-9 CRISPR1 Cas9
(with PAM of NNAGAAW)

(SEQ ID NO: 66)

NNNNNNNNNNNNNNNNNNNNgttattgtactctcaGAAAtgcagaagctacaatgataaggcttcatgccgaaatc
aacaccctgtcattttatggcagggtgttttcgttatttaa**TTTTTT**

(N = guide sequence; first underline = tracr mate sequence; second
underline = tracr sequence; bold = terminator)

Example chimeric RNA for *S. thermophilus* LMD-9 CRISPR1 Cas9
(with PAM of NNAGAAW)

(SEQ ID NO: 67)

NNNNNNNNNNNNNNNNNNNNgttattgtactctcaGAAAtgcagaagctacaatgataaggcttcatgccgaaatc
aacaccctgtcattttatggcagggtgt**TTTTTT**

(N = guide sequence; first underline = tracr mate sequence; second
underline = tracr sequence; bold = terminator)

Example chimeric RNA for *S. thermophilus* LMD-9 CRISPR1 Cas9
(with PAM of NNAGAAW)

(SEQ ID NO: 68)

NNNNNNNNNNNNNNNNNNNNgttttagagctgtgGAAAcacagcgagttaaaataaggcttagtccgtactcaact
tgaaaaggtggcaccgattcggtgt**TTTTTT**

(N = guide sequence; first underline = tracr mate sequence; second
underline = tracr sequence; bold = terminator)

Codon-optimized version of Cas9 from *S. thermophilus* LMD-9 CRISPR3 locus
(with NLS at both 5' and 3' ends)

(SEQ ID NO: 69)

ATGAAAAGGCCGGCGGCCACGAAAAAGGCCGGCCAGGCAAAAAAGAAAAAGACCAAGCCCTACAGCATCGGCCTGG
ACATCGGCACCAATAGCGTGGGCTGGGCCGTGACCACCGACAACTACAAGGTGCCCAGCAAGAAAATGAAGGTGCT
GGGCAACACCTCCAAGAAGTACATCAAGAAAAACCTGCTGGGCGTGCTGCTGTTCGACAGCGGCATTACAGCCGAG
GGCAGACGGCTGAAGAGAACCGCCAGACGGCGGTACACCCGGCGGAGAAACAGAATCCTGTATCTGCAAGAGATCT
TCAGCACCGAGATGGCTACCCTGGACGACGCCTTCTTCCAGCGGCTGGACGACAGCTTCCTGGTGCCCGACGACAA
GCGGGACAGCAAGTACCCCATCTTCGGCAACCTGGTGGAAGAGAAGGCCTACCACGACGAGTTCCCCACCATCTAC
CACCTGAGAAAGTACCTGGCCGACAGCACCAAGAAGGCCGACCTGAGACTGGTGTATCTGGCCCTGGCCCACATGA
TCAAGTACCGGGGCCACTTCCTGATCGAGGGCGAGTTCAACAGCAAGAACAACGACATCCAGAAGAACTTCCAGGA
CTTCCTGGACACCTACAACGCCATCTTCGAGAGCGACCTGTCCCTGGAAAACAGCAAGCAGCTGGAAGAGATCGTG
AAGGACAAGATCAGCAAGCTGGAAAAGAAGGACCGCATCCTGAAGCTGTTCCCCGGCGAGAAGAACAGCGGAATCT
TCAGCGAGTTTCTGAAGCTGATCGTGGGCAACCAGGCCGACTTCAGAAAGTGCTTCAACCTGGACGAGAAAGCCAG
CCTGCACTTCAGCAAAGAGAGCTACGACGAGGACCTGGAAACCCTGCTGGGATATATCGGCGACGACTACAGCGAC
GTGTTCCTGAAGGCCAAGAAGCTGTACGACGCTATCCTGCTGAGCGGCTTCCTGACCGTGACCGACAACGAGACAG
AGGCCCCACTGAGCAGCGCCATGATTAAGCGGTACAACGAGCACAAAGAGGATCTGGCTCTGCTGAAAGAGTACAT
CCGGAACATCAGCCTGAAAACCTACAATGAGGTGTTCAAGGACGACACCAAGAACGGCTACGCCGGCTACATCGAC
GGCAAGACCAACCAGGAAGATTTCTATGTGTACCTGAAGAAGCTGCTGGCCGAGTTCGAGGGGGCCGACTACTTTC
TGGAAAAAATCGACCGCGAGGATTTCCTGCGGAAGCAGCGGACCTTCGACAACGGCAGCATCCCCTACCAGATCCA
TCTGCAGGAAATGCGGGCCATCCTGGACAAGCAGGCCAAGTTCTACCCTTCCTGGCCAAGAACAAAGAGCGGATCG
AGAAGATCCTGACCTTCCGCATCCCTTACTACGTGGGCCCCCTGGCCAGAGGCAACAGCGATTTTGCCTGGTCCAT
CCGGAAGCGCAATGAGAAGATCACCCCCTGGAACTTCGAGGACGTGATCGACAAAGAGTCCAGCGCCGAGGCCTTC
ATCAACCGGATGACCAGCTTCGACCTGTACCTGCCCGAGGAAAAGGTGCTGCCCAAGCACAGCCTGCTGTACGAGA
CATTCAATGTGTATAACGAGCTGACCAAAGTGCGGTTTATCGCCGAGTCTATGCGGGACTACCAGTTCCTGGACTC
CAAGCAGAAAAAGGACATCGTGCGGCTGTACTTCAAGGACAAGCGGAAAGTGACCGATAAGGACATCATCGAGTAC
CTGCACGCCATCTACGGCTACGATGGCATCGAGCTGAAGGGCATCGAGAAGCAGTTCAACTCCAGCCTGAGCACAT
ACCACGACCTGCTGAACATTATCAACGACAAAGAATTTCTGGACGACTCCAGCAACGAGGCCATCATCGAAGAGAT
CATCCACACCCTGACCATCTTTGAGGACCGCGAGATGATCAAGCAGCGGCTGAGCAAGTTCGAGAACATCTTCGAC
AAGAGCGTGCTGAAAAAGCTGAGCAGACGGCACTACACCGGCTGGGGCAAGCTGAGCGCCAAGCTGATCAACGGCA
TCCGGGACGAGAAGTCCGGCAACACAATCCTGGACTACCTGATCGACGACGGCATCAGCAACCGGAACTTCATGCA
GCTGATCCACGACGACGCCCTGAGCTTCAAGAAGAAGATCCAGAAGGCCCAGATCATCGGGGACGAGGACAAGGGC
AACATCAAAGAAGTCGTGAAGTCCCTGCCCGGCAGCCCCGCCATCAAGAAGGGAATCCTGCAGAGCATCAAGATCG
TGGACGAGCTCGTGAAAGTGATGGGCGGCAGAAAGCCCGAGAGCATCGTGGTGGAAATGGCTAGAGAGAACCAGTA
CACCAATCAGGGCAAGAGCAACAGCCAGCAGAGACTGAAGAGACTGGAAAAGTCCCTGAAAGAGCTGGGCAGCAAG
ATTCTGAAAGAGAATATCCCTGCCAAGCTGTCCAAGATCGACAACAACGCCCTGCAGAACGACCGGCTGTACCTGT
ACTACCTGCAGAATGGCAAGGACATGTATACAGGCGACGACCTGGATATCGACCGCCTGAGCAACTACGACATCGA

-continued

CCATATTATCCCCCAGGCCTTCCTGAAAGACAACAGCATTGACAACAAAGTGCTGGTGTCCTCCGCCAGCAACCGC
GGCAAGTCCGATGATGTGCCCAGCCTGGAAGTCGTGAAAAAGAGAAAGACCTTCTGGTATCAGCTGCTGAAAAGCA
AGCTGATTAGCCAGAGGAAGTTCGACAACCTGACCAAGGCCGAGAGAGGCGGCCTGAGCCCTGAAGATAAGGCCGG
CTTCATCCAGAGACAGCTGGTGGAAACCCGGCAGATCACCAAGCACGTGGCCAGACTGCTGGATGAGAAGTTTAAC
AACAAGAAGGACGAGAACAACCGGGCCGTGCGGACCGTGAAGATCATCACCCTGAAGTCCACCCTGGTGTCCCAGT
TCCGGAAGGACTTCGAGCTGTATAAAGTGCGCGAGATCAATGACTTTCACCACGCCCACGACGCCTACCTGAATGC
CGTGGTGGCTTCCGCCCTGCTGAAGAAGTACCCTAAGCTGGAACCCGAGTTCGTGTACGGCGACTACCCCAAGTAC
AACTCCTTCAGAGAGCGGAAGTCCGCCACGAGAAGGTGTACTTCTACTCCAACATCATGAATATCTTTAAGAAGTC
CATCTCCCTGGCCGATGGCAGAGTGATCGAGCGGCCCCTGATCGAAGTGAACGAAGAGACAGGCGAGAGCGTGTGG
AACAAAGAAAGCGACCTGGCCACCGTGCGGCGGGTGCTGAGTTATCCTCAAGTGAATGTCGTGAAGAAGGTGGAAG
AACAGAACCACGGCCTGGATCGGGGCAAGCCCAAGGGCCTGTTCAACGCCAACCTGTCCAGCAAGCCTAAGCCCAA
CTCCAACGAGAATCTCGTGGGGGCCAAAGAGTACCTGGACCCTAAGAAGTACGGCGGATACGCCGGCATCTCCAAT
AGCTTCACCGTGCTCGTGAAGGGCACAATCGAGAAGGGCGCTAAGAAAAAGATCACAAACGTGCTGGAATTTCAGG
GGATCTCTATCCTGGACCGGATCAACTACCGGAAGGATAAGCTGAACTTTCTGCTGGAAAAAGGCTACAAGGACAT
TGAGCTGATTATCGAGCTGCCTAAGTACTCCCTGTTCGAACTGAGCGACGGCTCCAGACGGATGCTGGCCTCCATC
CTGTCCACCAACAACAAGCGGGGCGAGATCCACAAGGGAAACCAGATCTTCCTGAGCCAGAAATTTGTGAAACTGC
TGTACCACGCCAAGCGGATCTCCAACACCATCAATGAGAACCACCGGAAATACGTGGAAAACCACAAGAAAGAGTT
TGAGGAACTGTTCTACTACATCCTGGAGTTCAACGAGAACTATGTGGGAGCCAAGAAGAACGGCAAACTGCTGAAC
TCCGCCTTCCAGAGCTGGCAGAACCACAGCATCGACGAGCTGTGCAGCTCCTTCATCGGCCCTACCGGCAGCGAGC
GGAAGGGACTGTTTGAGCTGACCTCCAGAGGCTCTGCCGCCGACTTTGAGTTCCTGGGAGTGAAGATCCCCCGGTA
CAGAGACTACACCCCCTCTAGTCTGCTGAAGGACGCCACCCTGATCCACCAGAGCGTGACCGGCCTGTACGAAACC
CGGATCGACCTGGCTAAGCTGGGCGAGGGAAAGCGTCCTGCTGCTACTAAGAAAGCTGGTCAAGCTAAGAAAAAGA
AATAA

Example 5: Optimization of the Guide RNA for *Streptococcus pyogenes* Cas9 (Referred to as SpCas9)

[0185] Applicants mutated the tracrRNA and direct repeat sequences, or mutated the chimeric guide RNA to enhance the RNAs in cells.

[0186] The optimization is based on the observation that there were stretches of thymines (Ts) in the tracrRNA and guide RNA, which might lead to early transcription termination by the pol 3 promoter. Therefore Applicants generated the following optimized sequences. Optimized tracrRNA and corresponding optimized direct repeat are presented in pairs.

Optimized tracrRNA 1 (mutation underlined):
(SEQ ID NO: 70)
GGAACCATTCAtAACAGCATAGCAAGTTAtAATAAGGCTAGTCCGTTA
TCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTT

Optimized direct repeat 1 (mutation underlined):
(SEQ ID NO: 71)
GTTaTAGAGCTATGCTGTTaTGAATGGTCCCAAAAC

Optimized tracrRNA 2 (mutation underlined):
(SEQ ID NO: 72)
GGAACCATTCAAtACAGCATAGCAAGTTAAtATAAGGCTAGTCCGTTA
TCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTT

Optimized direct repeat 2 (mutation underlined):
(SEQ ID NO: 73)
GTaTTAGAGCTATGCTGTaTTGAATGGTCCCAAAAC

Applicants also optimized the chimeric guideRNA for optimal activity in eukaryotic cells.

Original guide RNA:
(SEQ ID NO: 74)
NNNNNNNNNNNNNNNNNNNNGTTTTAGAGCTAGAAATAGCAAGTTAAA
ATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGC
TTTTTTT

Optimized chimeric guide RNA sequence 1:
(SEQ ID NO: 75)
NNNNNNNNNNNNNNNNNNNNGTATTAGAGCTAGAAATAGCAAGTTAAT
ATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGC
TTTTTTT

-continued

Optimized chimeric guide RNA sequence 2:
(SEQ ID NO: 76)
NNNNNNNNNNNNNNNNNNNNGTTTTAGAGCTATGCTGTTTTGGAAACA
AAACAGCATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAA
AAGTGGCACCGAGTCGGTGCTTTTTTT

Optimized chimeric guide RNA sequence 3:
(SEQ ID NO: 77)
NNNNNNNNNNNNNNNNNNNNGTATTAGAGCTATGCTGTATTGGAAACA
ATACAGCATAGCAAGTTAATATAAGGCTAGTCCGTTATCAACTTGAAA
AAGTGGCACCGAGTCGGTGCTTTTTTT

[0187] Applicants showed that optimized chimeric guide RNA works better as indicated in FIG. **9**. The experiment was conducted by co-transfecting 293FT cells with Cas9 and a U6-guide RNA DNA cassette to express one of the four RNA forms shown above. The target of the guide RNA is the same target site in the human Emx1 locus: "GTCACCTC-CAATGACTAGGG" (SEQ ID NO: 78)

Example 6: Optimization of *Streptococcus thermophilus* LMD-9 CRISPR1 Cas9 (Referred to as St1Cas9)

[0188] Applicants designed guide chimeric RNAs as shown in FIG. **12**.

[0189] The St1Cas9 guide RNAs can under go the same type of optimization as for SpCas9 guide RNAs, by breaking the stretches of poly thymines (Ts).

Example 7: Improvement of the Cas9 System for In Vivo Application

[0190] Applicants conducted a Metagenomic search for a Cas9 with small molecular weight. Most Cas9 homologs are fairly large. For example the SpCas9 is around 1368aa long, which is too large to be easily packaged into viral vectors for delivery. Some of the sequences may have been mis-annotated and therefore the exact frequency for each length may not necessarily be accurate. Nevertheless it provides a glimpse at distribution of Cas9 proteins and suggest that there are shorter Cas9 homologs.

[0191] Through computational analysis, Applicants found that in the bacterial strain *Campylobacter*, there are two Cas9 proteins with less than 1000 amino acids. The sequence for one Cas9 from *Campylobacter jejuni* is pre-

sented below. At this length, CjCas9 can be easily packaged into AAV, lentiviruses, Adenoviruses, and other viral vectors for robust delivery into primary cells and in vivo in animal models.

>*Campylobacter jejuni* Cas9 (CjCas91)

(SEQ ID NO: 79)

MARILAFDIGISSIGWAFSENDILKGCGVRIFTKVENPKTGESLALPRR

LARSARKRLARRKARLNHLKHLIANEFKLNYEDYQSFDESLAKAYKGSL

ISPYELRFRALNELLSKQDFARVILHIAKRRGYDDIKNSDDKEKGAILK

AIKQNEEKLANYQSVGEYLYKEYFQKFKENSKEFTNVRNKKESYERCIA

QSFLKDELKLIFKKQREFGFSFSKKFEEEVLSVAFYKRALKDFSHLVGN

CSFFTDEKRAPKNSPLAFMFVALTRIINLLNNLKNTEGILYTKDDLNAL

LENEVLKNGTLTYKQTKKLLGLSDDYEFKGEKGTYFIEFKKYKEFIKAL

GEHNLSQDDLNEIAKDITLIKDEIKLKKALAKYDLNQNQIDSLSKLEFK

DHLNISFKALKLVTPLMLEGKKYDEACNELNLKVAINEDKKDFLPAFNE

TYYKDEVTNPVVLRAIKEYRKVLNALLKKYGKVHKINIELAREVGKNHS

QRAKIEKEQNENYKAKKDAELECEKLGLKINSKNILKLRLFKEQKEFCA

YSGEKIKISDLQDEKMLEIDHIYPYSRSFDDSYMNKVLVFTKQNQEKLN

QTPFEAFGNDSAKWQKIEVLKNLPTKKQKRILDKNYKDKEQKNFKDNLN

DTRYIARLVLNYTKDYLDFLPLSDDENTKLNDTQKGSKVHVEAKSGMLT

SALRHTWGFSAKDRNNHLHHAIDAVIIAYANNSIVKAFSDFKKEQESNS

AELYAKKISELDYKNKRKFFEPFSGFRQKVLDKIDEIFVSKPERKKPSG

ALHEETFRKEEEFYQSYGGKEGVLKALELGKIRKVNGKIVKNGDMFRVD

IFKHKKTNKFYAVPIYTMDFALKVLPNKAVARSKKGEIKDWILMDENYE

FCFSLYKDSLILIQTKDMQEPEFVYYNAFTSSTVSLIVSKHDNKFETLS

KNQKILFKNANEKEVIAKSIGIQNLKVFEKYIVSALGEVTKAEFRQRED

FKK

[0192] The putative tracrRNA element for this CjCas9 is:

(SEQ ID NO: 80)

TATAATCTCATAAGAAATTTAAAAAGGGACTAAAATAAAGAGTTTGCGG

GACTCTGCGGGGTTACAATCCCCTAAAACCGCTTTTAAAATT

[0193] The Direct Repeat sequence is:

(SEQ ID NO: 81)

ATTTTACCATAAAGAAATTTAAAAAGGGACTAAAAC

[0194] The co-fold structure of the tracrRNA and direct repeat is provided in FIG. **6**.

[0195] An example of a chimeric guideRNA for CjCas9 is:

(SEQ ID NO: 82)

NNNNNNNNNNNNNNNNNNNGUUUUAGUCCCGAAAGGGACUAAAAUAAA

GAGUUUGCGGGACUCUGCGGGGUUACAAUCCCCUAAAACCGCUUUU

[0196] Applicants have also optimized Cas9 guide RNA using in vitro methods. FIG. **18** shows data from the St1Cas9 chimeric guide RNA optimization in vitro.

[0197] While preferred embodiments of the present invention have been shown and described herein, it will be obvious to those skilled in the art that such embodiments are provided by way of example only. Numerous variations, changes, and substitutions will now occur to those skilled in the art without departing from the invention. It should be understood that various alternatives to the embodiments of the invention described herein may be employed in practicing the invention. It is intended that the following claims define the scope of the invention and that methods and structures within the scope of these claims and their equivalents be covered thereby.

Example 8: Sa sgRNA Optimization

[0198] Applicants designed five sgRNA variants for SaCas9 for an optimal truncated architecture with highest cleavage efficiency. In addition, the native direct repeat:tracr duplex system was tested alongside sgRNAs. Guides with indicated lengths were co-transfected with SaCas9 and tested in HEK 293FT cells for activity. A total of 100 ng sgRNA U6-PCR amplicon (or 50 ng of direct repeat and 50 ng of tracrRNA) and 400 ng of SaCas9 plasmid were co-transfected into 200,000 Hepa1-6 mouse hepatocytes, and DNA was harvested 72-hours post-transfection for SURVEYOR analysis. The results are shown in FIG. **23**.

REFERENCES

[0199] 1. Urnov, F. D., Rebar, E. J., Holmes, M. C., Zhang, H. S. & Gregory, P. D. Genome editing with engineered zinc finger nucleases. *Nat. Rev. Genet.* 11, 636-646 (2010).

[0200] 2. Bogdanove, A. J. & Voytas, D. F. TAL effectors: customizable proteins for DNA targeting. *Science* 333, 1843-1846 (2011).

[0201] 3. Stoddard, B. L. Homing endonuclease structure and function. *Q. Rev. Biophys.* 38, 49-95 (2005).

[0202] 4. Bae, T. & Schneewind, O. Allelic replacement in *Staphylococcus aureus* with inducible counter-selection. *Plasmid* 55, 58-63 (2006).

[0203] 5. Sung, C. K., Li, H., Claverys, J. P. & Morrison, D. A. An rpsL cassette, janus, for gene replacement through negative selection in *Streptococcus pneumoniae*. *Appl. Environ. Microbiol.* 67, 5190-5196 (2001).

[0204] 6. Sharan, S. K., Thomason, L. C., Kuznetsov, S. G. & Court, D. L. Recombineering: a homologous recombination-based method of genetic engineering. *Nat. Protoc.* 4, 206-223 (2009).

[0205] 7. Jinek, M. et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* 337, 816-821 (2012).

[0206] 8. Deveau, H., Garneau, J. E. & Moineau, S. CRISPR-Cas system and its role in phage-bacteria interactions. *Annu. Rev. Microbiol.* 64, 475-493 (2010).

[0207] 9. Horvath, P. & Barrangou, R. CRISPR-Cas, the immune system of bacteria and archaea. *Science* 327, 167-170 (2010).

[0208] 10. Terns, M. P. & Terns, R. M. CRISPR-based adaptive immune systems. *Curr. Opin. Microbiol.* 14, 321-327 (2011).

[0209] 11. van der Oost, J., Jore, M. M., Westra, E. R., Lundgren, M. & Brouns, S. J. CRISPR-based adaptive and heritable immunity in prokaryotes. *Trends. Biochem. Sci.* 34, 401-407 (2009).

[0210] 12. Brouns, S. J. et al. Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* 321, 960-964 (2008).

[0211] 13. Carte, J., Wang, R., Li, H., Terns, R. M. & Terns, M. P. Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. *Genes Dev.* 22, 3489-3496 (2008).

[0212] 14. Deltcheva, E. et al. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* 471, 602-607 (2011).

[0213] 15. Hatoum-Aslan, A., Maniv, I. & Marraffini, L. A. Mature clustered, regularly interspaced, short palindromic repeats RNA (crRNA) length is measured by a ruler mechanism anchored at the precursor processing site. *Proc. Natl. Acad. Sci. U.S.A.* 108, 21218-21222 (2011).

[0214] 16. Haurwitz, R. E., Jinek, M., Wiedenheft, B., Zhou, K. & Doudna, J. A. Sequence- and structure-specific RNA processing by a CRISPR endonuclease. *Science* 329, 1355-1358 (2010).

[0215] 17. Deveau, H. et al. Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. *J. Bacteriol.* 190, 1390-1400 (2008).

[0216] 18. Gasiunas, G., Barrangou, R., Horvath, P. & Siksnys, V. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proc. Natl. Acad. Sci. U.S.A.* (2012).

[0217] 19. Makarova, K. S., Aravind, L., Wolf, Y. I. & Koonin, E. V. Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. *Biol. Direct.* 6, 38 (2011).

[0218] 20. Barrangou, R. RNA-mediated programmable DNA cleavage. *Nat. Biotechnol.* 30, 836-838 (2012).

[0219] 21. Brouns, S. J. Molecular biology. A Swiss army knife of immunity. *Science* 337, 808-809 (2012).

[0220] 22. Carroll, D. A CRISPR Approach to Gene Targeting. *Mol. Ther.* 20, 1658-1660 (2012).

[0221] 23. Bikard, D., Hatoum-Aslan, A., Mucida, D. & Marraffini, L. A. CRISPR interference can prevent natural transformation and virulence acquisition during in vivo bacterial infection. *Cell Host Microbe* 12, 177-186 (2012).

[0222] 24. Sapranauskas, R. et al. The *Streptococcus thermophilus* CRISPR-Cas system provides immunity in *Escherichia coli*. *Nucleic Acids Res.* (2011).

[0223] 25. Semenova, E. et al. Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. *Proc. Natl. Acad. Sci. U.S.A.* (2011).

[0224] 26. Wiedenheft, B. et al. RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. *Proc. Natl. Acad. Sci. U.S.A.* (2011).

[0225] 27. Zahner, D. & Hakenbeck, R. The *Streptococcus pneumoniae* beta-galactosidase is a surface protein. *J. Bacteriol.* 182, 5919-5921 (2000).

[0226] 28. Marraffini, L. A., Dedent, A. C. & Schneewind, O. Sortases and the art of anchoring proteins to the envelopes of gram-positive bacteria. *Microbiol. Mol. Biol. Rev.* 70, 192-221 (2006).

[0227] 29. Motamedi, M. R., Szigety, S. K. & Rosenberg, S. M. Double-strand-break repair recombination in *Escherichia coli*: physical evidence for a DNA replication mechanism in vivo. *Genes Dev.* 13, 2889-2903 (1999).

[0228] 30. Hosaka, T. et al. The novel mutation K87E in ribosomal protein S12 enhances protein synthesis activity during the late growth phase in *Escherichia coli*. *Mol. Genet. Genomics* 271, 317-324 (2004).

[0229] 31. Costantino, N. & Court, D. L. Enhanced levels of lambda Red-mediated recombinants in mismatch repair mutants. *Proc. Natl. Acad. Sci. U.S.A.* 100, 15748-15753 (2003).

[0230] 32. Edgar, R. & Qimron, U. The *Escherichia coli* CRISPR system protects from lambda lysogenization, lysogens, and prophage induction. *J. Bacteriol.* 192, 6291-6294 (2010).

[0231] 33. Marraffini, L. A. & Sontheimer, E. J. Self versus non-self discrimination during CRISPR RNA-directed immunity. *Nature* 463, 568-571 (2010).

[0232] 34. Fischer, S. et al. An archaeal immune system can detect multiple Protospacer Adjacent Motifs (PAMs) to target invader DNA. *J. Biol. Chem.* 287, 33351-33363 (2012).

[0233] 35. Gudbergsdottir, S. et al. Dynamic properties of the *Sulfolobus* CRISPR-Cas and CRISPR/Cmr systems when challenged with vector-borne viral and plasmid genes and protospacers. *Mol. Microbiol.* 79, 35-49 (2011).

[0234] 36. Wang, H. H. et al. Genome-scale promoter engineering by coselection MAGE. *Nat Methods* 9, 591-593 (2012).

[0235] 37. Cong, L. et al. Multiplex Genome Engineering Using CRISPR-Cas Systems. *Science* In press (2013).

[0236] 38. Mali, P. et al. RNA-Guided Human Genome Engineering via Cas9. *Science* In press (2013).

[0237] 39. Hoskins, J. et al. Genome of the bacterium *Streptococcus pneumoniae* strain R6. *J. Bacteriol.* 183, 5709-5717 (2001).

[0238] 40. Havarstein, L. S., Coomaraswamy, G. & Morrison, D. A. An unmodified heptadecapeptide pheromone induces competence for genetic transformation in *Streptococcus pneumoniae*. *Proc. Natl. Acad. Sci. U.S.A.* 92, 11140-11144 (1995).

[0239] 41. Horinouchi, S. & Weisblum, B. Nucleotide sequence and functional map of pC194, a plasmid that specifies inducible chloramphenicol resistance. *J. Bacteriol.* 150, 815-825 (1982).

[0240] 42. Horton, R. M. In Vitro Recombination and Mutagenesis of DNA: SOEing Together Tailor-Made Genes. *Methods Mol. Biol.* 15, 251-261 (1993).

[0241] 43. Podbielski, A., Spellerberg, B., Woischnik, M., Pohl, B. & Lutticken, R. Novel series of plasmid vectors for gene inactivation and expression analysis in group A streptococci (GAS). *Gene* 177, 137-147 (1996).

[0242] 44. Husmann, L. K., Scott, J. R., Lindahl, G. & Stenberg, L. Expression of the Arp protein, a member of the M protein family, is not sufficient to inhibit phagocytosis of *Streptococcus pyogenes*. *Infection and immunity* 63, 345-348 (1995).

[0243] 45. Gibson, D. G. et al. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods* 6, 343-345 (2009).

[0244] 46. Tangri S, et al. ("Rationally engineered therapeutic proteins with reduced immunogenicity" J Immunol. 2005 Mar. 15; 174(6):3187-96.

[0245] While preferred embodiments of the present invention have been shown and described herein, it will be obvious to those skilled in the art that such embodiments are provided by way of example only. Numerous variations, changes, and substitutions will now occur to those skilled in the art without departing from the invention. It should be understood that various alternatives to the embodiments of the invention described herein may be employed in practicing the invention.

SEQUENCE LISTING

Sequence total quantity: 272

SEQ ID NO: 1 moltype = DNA length = 15
FEATURE Location/Qualifiers
misc_feature 1..15
note = Description of Artificial Sequence: Synthetic oligonucleotide
source 1..15
mol_type = other DNA
organism = synthetic construct
SEQUENCE: 1
aggacgaagt cctaa 15

SEQ ID NO: 2 moltype = AA length = 7
FEATURE Location/Qualifiers
source 1..7
mol_type = protein
organism = Simian virus 40
SEQUENCE: 2
PKKKRKV 7

SEQ ID NO: 3 moltype = AA length = 16
FEATURE Location/Qualifiers
REGION 1..16
note = Description of Unknown: Nucleoplasmin bipartite NLS sequence
source 1..16
mol_type = protein
organism = unidentified
SEQUENCE: 3
KRPAATKKAG QAKKKK 16

SEQ ID NO: 4 moltype = AA length = 9
FEATURE Location/Qualifiers
REGION 1..9
note = Description of Unknown: C-myc NLS sequence
source 1..9
mol_type = protein
organism = unidentified
SEQUENCE: 4
PAAKRVKLD 9

SEQ ID NO: 5 moltype = AA length = 11
FEATURE Location/Qualifiers
REGION 1..11
note = Description of Unknown: C-myc NLS sequence
source 1..11
mol_type = protein
organism = unidentified
SEQUENCE: 5
RQRRNELKRS P 11

SEQ ID NO: 6 moltype = AA length = 38
FEATURE Location/Qualifiers
source 1..38
mol_type = protein
organism = Homo sapiens
SEQUENCE: 6
NQSSNFGPMK GGNFGGRSSG PYGGGGQYFA KPRNQGGY 38

SEQ ID NO: 7 moltype = AA length = 42
FEATURE Location/Qualifiers
REGION 1..42
note = Description of Unknown: IBB domain from importin-alpha sequence

-continued

```
source                  1..42
                        mol_type = protein
                        organism = unidentified
SEQUENCE: 7
RMRIZFKNKG KDTAELRRRR VEVSVELRKA KKDEQILKRR NV                    42

SEQ ID NO: 8            moltype = AA  length = 8
FEATURE                 Location/Qualifiers
REGION                  1..8
                        note = Description of Unknown: Myoma T protein sequence
source                  1..8
                        mol_type = protein
                        organism = unidentified
SEQUENCE: 8
VSRKRPRP                                                          8

SEQ ID NO: 9            moltype = AA  length = 8
FEATURE                 Location/Qualifiers
REGION                  1..8
                        note = Description of Unknown: Myoma T protein sequence
source                  1..8
                        mol_type = protein
                        organism = unidentified
SEQUENCE: 9
PPKKARED                                                          8

SEQ ID NO: 10           moltype = AA  length = 8
FEATURE                 Location/Qualifiers
source                  1..8
                        mol_type = protein
                        organism = Homo sapiens
SEQUENCE: 10
PQPKKKPL                                                          8

SEQ ID NO: 11           moltype = AA  length = 12
FEATURE                 Location/Qualifiers
source                  1..12
                        mol_type = protein
                        organism = Mus musculus
SEQUENCE: 11
SALIKKKKKM AP                                                     12

SEQ ID NO: 12           moltype = AA  length = 5
FEATURE                 Location/Qualifiers
source                  1..5
                        mol_type = protein
                        organism = Influenza virus
SEQUENCE: 12
DRLRR                                                             5

SEQ ID NO: 13           moltype = AA  length = 7
FEATURE                 Location/Qualifiers
source                  1..7
                        mol_type = protein
                        organism = Influenza virus
SEQUENCE: 13
PKQKKRK                                                           7

SEQ ID NO: 14           moltype = AA  length = 10
FEATURE                 Location/Qualifiers
source                  1..10
                        mol_type = protein
                        organism = Hepatitis delta virus
SEQUENCE: 14
RKLKKKIKKL                                                        10

SEQ ID NO: 15           moltype = AA  length = 10
FEATURE                 Location/Qualifiers
source                  1..10
                        mol_type = protein
                        organism = Mus musculus
SEQUENCE: 15
REKKKFLKRR                                                        10

SEQ ID NO: 16           moltype = AA  length = 20
FEATURE                 Location/Qualifiers
```

-continued

```
source                  1..20
                        mol_type = protein
                        organism = Homo sapiens
SEQUENCE: 16
KRKGDEVDGV DEVAKKKSKK                                          20

SEQ ID NO: 17           moltype = AA  length = 17
FEATURE                 Location/Qualifiers
source                  1..17
                        mol_type = protein
                        organism = Homo sapiens
SEQUENCE: 17
RKCLQAGMNL EARKTKK                                             17

SEQ ID NO: 18           moltype =   length =
SEQUENCE: 18
000

SEQ ID NO: 19           moltype =   length =
SEQUENCE: 19
000

SEQ ID NO: 20           moltype =   length =
SEQUENCE: 20
000

SEQ ID NO: 21           moltype =   length =
SEQUENCE: 21
000

SEQ ID NO: 22           moltype = DNA  length = 137
FEATURE                 Location/Qualifiers
misc_feature            1..137
                        note = Description of Artificial Sequence: Synthetic
                         polynucleotide
misc_difference         1..20
                        note = a, c, t, g, unknown or other
source                  1..137
                        mol_type = other DNA
                        organism = synthetic construct
SEQUENCE: 22
nnnnnnnnnn nnnnnnnnnn gtttttgtac tctcaagatt tagaaataaa tcttgcagaa  60
gctacaaaga taaggcttca tgccgaaatc aacaccctgt cattttatgg cagggtgttt  120
tcgttattta atttttt                                             137

SEQ ID NO: 23           moltype = DNA  length = 123
FEATURE                 Location/Qualifiers
misc_feature            1..123
                        note = Description of Artificial Sequence: Synthetic
                         polynucleotide
misc_difference         1..20
                        note = a, c, t, g, unknown or other
source                  1..123
                        mol_type = other DNA
                        organism = synthetic construct
SEQUENCE: 23
nnnnnnnnnn nnnnnnnnnn gtttttgtac tctcagaaat gcagaagcta caaagataag  60
gcttcatgcc gaaatcaaca ccctgtcatt ttatggcagg gtgttttcgt tatttaattt  120
ttt                                                            123

SEQ ID NO: 24           moltype = DNA  length = 110
FEATURE                 Location/Qualifiers
misc_feature            1..110
                        note = Description of Artificial Sequence: Synthetic
                         polynucleotide
misc_difference         1..20
                        note = a, c, t, g, unknown or other
source                  1..110
                        mol_type = other DNA
                        organism = synthetic construct
SEQUENCE: 24
nnnnnnnnnn nnnnnnnnnn gtttttgtac tctcagaaat gcagaagcta caaagataag  60
gcttcatgcc gaaatcaaca ccctgtcatt ttatggcagg gtgtttttttt             110

SEQ ID NO: 25           moltype = DNA  length = 102
FEATURE                 Location/Qualifiers
```

-continued

```
misc_feature           1..102
                       note = Description of Artificial Sequence: Synthetic
                        polynucleotide
misc_difference        1..20
                       note = a, c, t, g, unknown or other
source                 1..102
                       mol_type = other DNA
                       organism = synthetic construct
SEQUENCE: 25
nnnnnnnnnn nnnnnnnnnn gttttagagc tagaaatagc aagttaaaat aaggctagtc  60
cgttatcaac ttgaaaaagt ggcaccgagt cggtgctttt tt                    102

SEQ ID NO: 26          moltype = DNA  length = 88
FEATURE                Location/Qualifiers
misc_feature           1..88
                       note = Description of Artificial Sequence: Synthetic
                        oligonucleotide
misc_difference        1..20
                       note = a, c, t, g, unknown or other
source                 1..88
                       mol_type = other DNA
                       organism = synthetic construct
SEQUENCE: 26
nnnnnnnnnn nnnnnnnnnn gttttagagc tagaaatagc aagttaaaat aaggctagtc  60
cgttatcaac ttgaaaaagt gttttttt                                    88

SEQ ID NO: 27          moltype = DNA  length = 76
FEATURE                Location/Qualifiers
misc_feature           1..76
                       note = Description of Artificial Sequence: Synthetic
                        oligonucleotide
misc_difference        1..20
                       note = a, c, t, g, unknown or other
source                 1..76
                       mol_type = other DNA
                       organism = synthetic construct
SEQUENCE: 27
nnnnnnnnnn nnnnnnnnnn gttttagagc tagaaatagc aagttaaaat aaggctagtc  60
cgttatcatt tttttt                                                 76

SEQ ID NO: 28          moltype = RNA  length = 12
FEATURE                Location/Qualifiers
misc_feature           1..12
                       note = Description of Artificial Sequence: Synthetic
                        oligonucleotide
source                 1..12
                       mol_type = other RNA
                       organism = synthetic construct
SEQUENCE: 28
gttttagagc ta                                                     12

SEQ ID NO: 29          moltype = DNA  length = 33
FEATURE                Location/Qualifiers
source                 1..33
                       mol_type = unassigned DNA
                       organism = Homo sapiens
SEQUENCE: 29
ggacatcgat gtcacctcca atgactaggg tgg                              33

SEQ ID NO: 30          moltype = DNA  length = 33
FEATURE                Location/Qualifiers
source                 1..33
                       mol_type = unassigned DNA
                       organism = Homo sapiens
SEQUENCE: 30
cattggaggt gacatcgatg tcctccccat tgg                              33

SEQ ID NO: 31          moltype = DNA  length = 33
FEATURE                Location/Qualifiers
source                 1..33
                       mol_type = unassigned DNA
                       organism = Homo sapiens
SEQUENCE: 31
ggaagggcct gagtccgagc agaagaagaa ggg                              33

SEQ ID NO: 32          moltype = DNA  length = 33
```

-continued

```
FEATURE                 Location/Qualifiers
source                  1..33
                        mol_type = unassigned DNA
                        organism = Homo sapiens
SEQUENCE: 32
ggtggcgaga ggggccgaga ttgggtgttc agg                                33

SEQ ID NO: 33           moltype = DNA  length = 33
FEATURE                 Location/Qualifiers
source                  1..33
                        mol_type = unassigned DNA
                        organism = Homo sapiens
SEQUENCE: 33
atgcaggagg gtggcgagag gggccgagat tgg                                33

SEQ ID NO: 34           moltype = DNA  length = 21
FEATURE                 Location/Qualifiers
misc_feature            1..21
                        note = Description of Artificial Sequence: Synthetic primer
source                  1..21
                        mol_type = other DNA
                        organism = synthetic construct
SEQUENCE: 34
aaaaccaccc ttctctctgg c                                             21

SEQ ID NO: 35           moltype = DNA  length = 21
FEATURE                 Location/Qualifiers
misc_feature            1..21
                        note = Description of Artificial Sequence: Synthetic primer
source                  1..21
                        mol_type = other DNA
                        organism = synthetic construct
SEQUENCE: 35
ggagattgga gacacggaga g                                             21

SEQ ID NO: 36           moltype = DNA  length = 20
FEATURE                 Location/Qualifiers
misc_feature            1..20
                        note = Description of Artificial Sequence: Synthetic primer
source                  1..20
                        mol_type = other DNA
                        organism = synthetic construct
SEQUENCE: 36
ctggaaagcc aatgcctgac                                               20

SEQ ID NO: 37           moltype = DNA  length = 20
FEATURE                 Location/Qualifiers
misc_feature            1..20
                        note = Description of Artificial Sequence: Synthetic primer
source                  1..20
                        mol_type = other DNA
                        organism = synthetic construct
SEQUENCE: 37
ggcagcaaac tccttgtcct                                               20

SEQ ID NO: 38           moltype = DNA  length = 12
FEATURE                 Location/Qualifiers
misc_feature            1..12
                        note = Description of Artificial Sequence: Synthetic
                         oligonucleotide
source                  1..12
                        mol_type = other DNA
                        organism = synthetic construct
SEQUENCE: 38
gttttagagc ta                                                       12

SEQ ID NO: 39           moltype = DNA  length = 335
FEATURE                 Location/Qualifiers
misc_feature            1..335
                        note = Description of Artificial Sequence: Synthetic
                         polynucleotide
source                  1..335
                        mol_type = other DNA
                        organism = synthetic construct
SEQUENCE: 39
gagggcctat ttcccatgat tccttcatat ttgcatatac gatacaaggc tgttagagag  60
```

-continued

```
ataattggaa ttaatttgac tgtaaacaca aagatattag tacaaaatac gtgacgtaga  120
aagtaataat ttcttgggta gtttgcagtt ttaaaattat gttttaaaat ggactatcat  180
atgcttaccg taacttgaaa gtatttcgat ttcttggctt tatatatctt gtggaaagga  240
cgaaacaccg gaaccattca aaacagcata gcaagttaaa ataaggctag tccgttatca  300
acttgaaaaa gtggcaccga gtcggtgctt ttttt                              335

SEQ ID NO: 40          moltype = DNA  length = 423
FEATURE                Location/Qualifiers
misc_feature           1..423
                       note = Description of Artificial Sequence: Synthetic
                        polynucleotide
source                 1..423
                       mol_type = other DNA
                       organism = synthetic construct
SEQUENCE: 40
gagggcctat ttcccatgat tccttcatat ttgcatatac gatacaaggc tgttagagag  60
ataattggaa ttaatttgac tgtaaacaca aagatattag tacaaaatac gtgacgtaga  120
aagtaataat ttcttgggta gtttgcagtt ttaaaattat gttttaaaat ggactatcat  180
atgcttaccg taacttgaaa gtatttcgat ttcttggctt tatatatctt gtggaaagga  240
cgaaacaccg gtagtattaa gtattgtttt atggctgata aatttctttg aatttctcct  300
tgattatttg ttataaaagt tataaaataa tcttgttgga accattcaaa acagcatagc  360
aagttaaaat aaggctagtc cgttatcaac ttgaaaaagt ggcaccgagt cggtgctttt  420
ttt                                                                423

SEQ ID NO: 41          moltype = DNA  length = 339
FEATURE                Location/Qualifiers
misc_feature           1..339
                       note = Description of Artificial Sequence: Synthetic
                        polynucleotide
source                 1..339
                       mol_type = other DNA
                       organism = synthetic construct
SEQUENCE: 41
gagggcctat ttcccatgat tccttcatat ttgcatatac gatacaaggc tgttagagag  60
ataattggaa ttaatttgac tgtaaacaca aagatattag tacaaaatac gtgacgtaga  120
aagtaataat ttcttgggta gtttgcagtt ttaaaattat gttttaaaat ggactatcat  180
atgcttaccg taacttgaaa gtatttcgat ttcttggctt tatatatctt gtggaaagga  240
cgaaacaccg ggttttagag ctatgctgtt ttgaatggtc ccaaaacggg tcttcgagaa  300
gacgttttag agctatgctg ttttgaatgg tcccaaaac                          339

SEQ ID NO: 42          moltype = DNA  length = 309
FEATURE                Location/Qualifiers
misc_feature           1..309
                       note = Description of Artificial Sequence: Synthetic
                        polynucleotide
source                 1..309
                       mol_type = other DNA
                       organism = synthetic construct
SEQUENCE: 42
gagggcctat ttcccatgat tccttcatat ttgcatatac gatacaaggc tgttagagag  60
ataattggaa ttaatttgac tgtaaacaca aagatattag tacaaaatac gtgacgtaga  120
aagtaataat ttcttgggta gtttgcagtt ttaaaattat gttttaaaat ggactatcat  180
atgcttaccg taacttgaaa gtatttcgat ttcttggctt tatatatctt gtggaaagga  240
cgaaacaccg ggtcttcgag aagacctgtt ttagagctag aaatagcaag ttaaaataag  300
gctagtccg                                                          309

SEQ ID NO: 43          moltype = AA  length = 1648
FEATURE                Location/Qualifiers
REGION                 1..1648
                       note = Description of Artificial Sequence: Synthetic
                        polypeptide
source                 1..1648
                       mol_type = protein
                       organism = synthetic construct
SEQUENCE: 43
MDYKDHDGDY KDHDIDYKDD DDKMAPKKKR KVGIHGVPAA DKKYSIGLDI GTNSVGWAVI  60
TDEYKVPSKK FKVLGNTDRH SIKKNLIGAL LFDSGETAEA TRLKRTARRR YTRRKNRICY  120
LQEIFSNEMA KVDDSFFHRL EESFLVEEDK KHERHPIFGN IVDEVAYHEK YPTIYHLRKK  180
LVDSTDKADL RLIYLALAHM IKFRGHFLIE GDLNPDNSDV DKLFIQLVQT YNQLFEENPI  240
NASGVDAKAI LSARLSKSRR LENLIAQLPG EKKNGLFGNL IALSLGLTPN FKSNFDLAED  300
AKLQLSKDTY DDDLDNLLAQ IGDQYADLFL AAKNLSDAIL LSDILRVNTE ITKAPLSASM  360
IKRYDEHHQD LTLLKALVRQ QLPEKYKEIF FDQSKNGYAG YIDGGASQEE FYKFIKPILE  420
KMDGTEELLV KLNREDLLRK QRTFDNGSIP HQIHLGELHA ILRRQEDFYP FLKDNREKIE  480
KILTFRIPYY VGPLARGNSR FAWMTRKSEE TITPWNFEEV VDKGASAQSF IERMTNFDKN  540
LPNEKVLPKH SLLYEYFTVY NELTKVKYVT EGMRKPAFLS GEQKKAIVDL LFKTNRKVTV  600
KQLKEDYFKK IECFDSVEIS GVEDRFNASL GTYHDLLKII KDKDFLDNEE NEDILEDIVL  660
```

-continued

```
TLTLFEDREM IEERLKTYAH LFDDKVMKQL KRRRYTGWGR LSRKLINGIR DKQSGKTILD  720
FLKSDGFANR NFMQLIHDDS LTFKEDIQKA QVSGQGDSLH EHIANLAGSP AIKKGILQTV  780
KVVDELVKVM GRHKPENIVI EMARENQTTQ KGQKNSRERM KRIEEGIKEL GSQILKEHPV  840
ENTQLQNEKL YLYYLQNGRD MYVDQELDIN RLSDYDVDHI VPQSFLKDDS IDNKVLTRSD  900
KNRGKSDNVP SEEVVKKMKN YWRQLLNAKL ITQRKFDNLT KAERGGLSEL DKAGFIKRQL  960
VETRQITKHV AQILDSRMNT KYDENDKLIR EVKVITLKSK LVSDFRKDFQ FYKVREINNY  1020
HHAHDAYLNA VVGTALIKKY PKLESEFVYG DYKVYDVRKM IAKSEQEIGK ATAKYFFYSN  1080
IMNFFKTEIT LANGEIRKRP LIETNGETGE IVWDKGRDFA TVRKVLSMPQ VNIVKKTEVQ  1140
TGGFSKESIL PKRNSDKLIA RKKDWDPKKY GGFDSPTVAY SVLVVAKVEK GKSKKLKSVK  1200
ELLGITIMER SSFEKNPIDF LEAKGYKEVK KDLIIKLPKY SLFELENGRK RMLASAGELQ  1260
KGNELALPSK YVNFLYLASH YEKLKGSPED NEQKQLFVEQ HKHYLDEIIE QISEFSKRVI  1320
LADANLDKVL SAYNKHRDKP IREQAENIIH LFTLTNLGAP AAFKYFDTTI DRKRYTSTKE  1380
VLDATLIHQS ITGLYETRID LSQLGGDAAA VSKGEELFTG VVPILVELDG DVNGHKFSVS  1440
GEGEGDATYG KLTLKFICTT GKLPVPWPTL VTTLTYGVQC FSRYPDHMKQ HDFFKSAMPE  1500
GYVQERTIFF KDDGNYKTRA EVKFEGDTLV NRIELKGIDF KEDGNILGHK LEYNYNSHNV  1560
YIMADKQKNG IKVNFKIRHN IEDGSVQLAD HYQQNTPIGD GPVLLPDNHY LSTQSALSKD  1620
PNEKRDHMVL LEFVTAAGIT LGMDELYK                                     1648

SEQ ID NO: 44          moltype = AA  length = 1625
FEATURE                Location/Qualifiers
REGION                 1..1625
                       note = Description of Artificial Sequence: Synthetic
                        polypeptide
source                 1..1625
                       mol_type = protein
                       organism = synthetic construct
SEQUENCE: 44
MDKKYSIGLD IGTNSVGWAV ITDEYKVPSK KFKVLGNTDR HSIKKNLIGA LLFDSGETAE  60
ATRLKRTARR RYTRRKNRIC YLQEIFSNEM AKVDDSFFHR LEESFLVEED KKHERHPIFG  120
NIVDEVAYHE KYPTIYHLRK KLVDSTDKAD LRLIYLALAH MIKFRGHFLI EGDLNPDNSD  180
VDKLFIQLVQ TYNQLFEENP INASGVDAKA ILSARLSKSR RLENLIAQLP GEKKNGLFGN  240
LIALSLGLTP NFKSNFDLAE DAKLQLSKDT YDDDLDNLLA QIGDQYADLF LAAKNLSDAI  300
LLSDILRVNT EITKAPLSAS MIKRYDEHHQ DLTLLKALVR QQLPEKYKEI FFDQSKNGYA  360
GYIDGGASQE EFYKFIKPIL EKMDGTEELL VKLNREDLLR KQRTFDNGSI PHQIHLGELH  420
AILRRQEDFY PFLKDNREKI EKILTFRIPY YVGPLARGNS RFAWMTRKSE ETITPWNFEE  480
VVDKGASAQS FIERMTNFDK NLPNEKVLPK HSLLYEYFTV YNELTKVKYV TEGMRKPAFL  540
SGEQKKAIVD LLFKTNRKVT VKQLKEDYFK KIECFDSVEI SGVEDRFNAS LGTYHDLLKI  600
IKDKDFLDNE ENEDILEDIV LTLTLFEDRE MIEERLKTYA HLFDDKVMKQ LKRRRYTGWG  660
RLSRKLINGI RDKQSGKTIL DFLKSDGFAN RNFMQLIHDD SLTFKEDIQK AQVSGQGDSL  720
HEHIANLAGS PAIKKGILQT VKVVDELVKV MGRHKPENIV IEMARENQTT QKGQKNSRER  780
MKRIEEGIKE LGSQILKEHP VENTQLQNEK LYLYYLQNGR DMYVDQELDI NRLSDYDVDH  840
IVPQSFLKDD SIDNKVLTRS DKNRGKSDNV PSEEVVKKMK NYWRQLLNAK LITQRKFDNL  900
TKAERGGLSE LDKAGFIKRQ LVETRQITKH VAQILDSRMN TKYDENDKLI REVKVITLKS  960
KLVSDFRKDF QFYKVREINN YHHAHDAYLN AVVGTALIKK YPKLESEFVY GDYKVYDVRK  1020
MIAKSEQEIG KATAKYFFYS NIMNFFKTEI TLANGEIRKR PLIETNGETG EIVWDKGRDF  1080
ATVRKVLSMP QVNIVKKTEV QTGGFSKESI LPKRNSDKLI ARKKDWDPKK YGGFDSPTVA  1140
YSVLVVAKVE KGKSKKLKSV KELLGITIME RSSFEKNPID FLEAKGYKEV KKDLIIKLPK  1200
YSLFELENGR KRMLASAGEL QKGNELALPS KYVNFLYLAS HYEKLKGSPE DNEQKQLFVE  1260
QHKHYLDEII EQISEFSKRV ILADANLDKV LSAYNKHRDK PIREQAENII HLFTLTNLGA  1320
PAAFKYFDTT IDRKRYTSTK EVLDATLIHQ SITGLYETRI DLSQLGGDAA AVSKGEELFT  1380
GVVPILVELD GDVNGHKFSV SGEGEGDATY GKLTLKFICT TGKLPVPWPT LVTTLTYGVQ  1440
CFSRYPDHMK QHDFFKSAMP EGYVQERTIF FKDDGNYKTR AEVKFEGDTL VNRIELKGID  1500
FKEDGNILGH KLEYNYNSHN VYIMADKQKN GIKVNFKIRH NIEDGSVQLA DHYQQNTPIG  1560
DGPVLLPDNH YLSTQSALSK DPNEKRDHMV LLEFVTAAGI TLGMDELYKK RPAATKKAGQ  1620
AKKKK                                                              1625

SEQ ID NO: 45          moltype = AA  length = 1664
FEATURE                Location/Qualifiers
REGION                 1..1664
                       note = Description of Artificial Sequence: Synthetic
                        polypeptide
source                 1..1664
                       mol_type = protein
                       organism = synthetic construct
SEQUENCE: 45
MDYKDHDGDY KDHDIDYKDD DDKMAPKKKR KVGIHGVPAA DKKYSIGLDI GTNSVGWAVI  60
TDEYKVPSKK FKVLGNTDRH SIKKNLIGAL LFDSGETAEA TRLKRTARRR YTRRKNRICY  120
LQEIFSNEMA KVDDSFFHRL EESFLVEEDK KHERHPIFGN IVDEVAYHEK YPTIYHLRKK  180
LVDSTDKADL RLIYLALAHM IKFRGHFLIE GDLNPDNSDV DKLFIQLVQT YNQLFEENPI  240
NASGVDAKAI LSARLSKSRR LENLIAQLPG EKKNGLFGNL IALSLGLTPN FKSNFDLAED  300
AKLQLSKDTY DDDLDNLLAQ IGDQYADLFL AAKNLSDAIL LSDILRVNTE ITKAPLSASM  360
IKRYDEHHQD LTLLKALVRQ QLPEKYKEIF FDQSKNGYAG YIDGGASQEE FYKFIKPILE  420
KMDGTEELLV KLNREDLLRK QRTFDNGSIP HQIHLGELHA ILRRQEDFYP FLKDNREKIE  480
KILTFRIPYY VGPLARGNSR FAWMTRKSEE TITPWNFEEV VDKGASAQSF IERMTNFDKN  540
LPNEKVLPKH SLLYEYFTVY NELTKVKYVT EGMRKPAFLS GEQKKAIVDL LFKTNRKVTV  600
KQLKEDYFKK IECFDSVEIS GVEDRFNASL GTYHDLLKII KDKDFLDNEE NEDILEDIVL  660
```

-continued

```
TLTLFEDREM IEERLKTYAH LFDDKVMKQL KRRRYTGWGR LSRKLINGIR DKQSGKTILD  720
FLKSDGFANR NFMQLIHDDS LTFKEDIQKA QVSGQGDSLH EHIANLAGSP AIKKGILQTV  780
KVVDELVKVM GRHKPENIVI EMARENQTTQ KGQKNSRERM KRIEEGIKEL GSQILKEHPV  840
ENTQLQNEKL YLYYLQNGRD MYVDQELDIN RLSDYDVDHI VPQSFLKDDS IDNKVLTRSD  900
KNRGKSDNVP SEEVVKKMKN YWRQLLNAKL ITQRKFDNLT KAERGGLSEL DKAGFIKRQL  960
VETRQITKHV AQILDSRMNT KYDENDKLIR EVKVITLKSK LVSDFRKDFQ FYKVREINNY  1020
HHAHDAYLNA VVGTALIKKY PKLESEFVYG DYKVYDVRKM IAKSEQEIGK ATAKYFFYSN  1080
IMNFFKTEIT LANGEIRKRP LIETNGETGE IVWDKGRDFA TVRKVLSMPQ VNIVKKTEVQ  1140
TGGFSKESIL PKRNSDKLIA RKKDWDPKKY GGFDSPTVAY SVLVVAKVEK GKSKKLKSVK  1200
ELLGITIMER SSFEKNPIDF LEAKGYKEVK KDLIIKLPKY SLFELENGRK RMLASAGELQ  1260
KGNELALPSK YVNFLYLASH YEKLKGSPED NEQKQLFVEQ HKHYLDEIIE QISEFSKRVI  1320
LADANLDKVL SAYNKHRDKP IREQAENIIH LFTLTNLGAP AAFKYFDTTI DRKRYTSTKE  1380
VLDATLIHQS ITGLYETRID LSQLGGDAAA VSKGEELFTG VVPILVELDG DVNGHKFSVS  1440
GEGEGDATYG KLTLKFICTT GKLPVPWPTL VTTLTYGVQC FSRYPDHMKQ HDFFKSAMPE  1500
GYVQERTIFF KDDGNYKTRA EVKFEGDTLV NRIELKGIDF KEDGNILGHK LEYNYNSHNV  1560
YIMADKQKNG IKVNFKIRHN IEDGSVQLAD HYQQNTPIGD GPVLLPDNHY LSTQSALSKD  1620
PNEKRDHMVL LEFVTAAGIT LGMDELYKKR PAATKKAGQA KKKK                    1664

SEQ ID NO: 46          moltype = AA  length = 1423
FEATURE                Location/Qualifiers
REGION                 1..1423
                       note = Description of Artificial Sequence: Synthetic
                        polypeptide
source                 1..1423
                       mol_type = protein
                       organism = synthetic construct
SEQUENCE: 46
MDYKDHDGDY KDHDIDYKDD DDKMAPKKKR KVGIHGVPAA DKKYSIGLDI GTNSVGWAVI  60
TDEYKVPSKK FKVLGNTDRH SIKKNLIGAL LFDSGETAEA TRLKRTARRR YTRRKNRICY  120
LQEIFSNEMA KVDDSFFHRL EESFLVEEDK KHERHPIFGN IVDEVAYHEK YPTIYHLRKK  180
LVDSTDKADL RLIYLALAHM IKFRGHFLIE GDLNPDNSDV DKLFIQLVQT YNQLFEENPI  240
NASGVDAKAI LSARLSKSRR LENLIAQLPG EKKNGLFGNL IALSLGLTPN FKSNFDLAED  300
AKLQLSKDTY DDDLDNLLAQ IGDQYADLFL AAKNLSDAIL LSDILRVNTE ITKAPLSASM  360
IKRYDEHHQD LTLLKALVRQ QLPEKYKEIF FDQSKNGYAG YIDGGASQEE FYKFIKPILE  420
KMDGTEELLV KLNREDLLRK QRTFDNGSIP HQIHLGELHA ILRRQEDFYP FLKDNREKIE  480
KILTFRIPYY VGPLARGNSR FAWMTRKSEE TITPWNFEEV VDKGASAQSF IERMTNFDKN  540
LPNEKVLPKH SLLYEYFTVY NELTKVKYVT EGMRKPAFLS GEQKKAIVDL LFKTNRKVTV  600
KQLKEDYFKK IECFDSVEIS GVEDRFNASL GTYHDLLKII KDKDFLDNEE NEDILEDIVL  660
TLTLFEDREM IEERLKTYAH LFDDKVMKQL KRRRYTGWGR LSRKLINGIR DKQSGKTILD  720
FLKSDGFANR NFMQLIHDDS LTFKEDIQKA QVSGQGDSLH EHIANLAGSP AIKKGILQTV  780
KVVDELVKVM GRHKPENIVI EMARENQTTQ KGQKNSRERM KRIEEGIKEL GSQILKEHPV  840
ENTQLQNEKL YLYYLQNGRD MYVDQELDIN RLSDYDVDHI VPQSFLKDDS IDNKVLTRSD  900
KNRGKSDNVP SEEVVKKMKN YWRQLLNAKL ITQRKFDNLT KAERGGLSEL DKAGFIKRQL  960
VETRQITKHV AQILDSRMNT KYDENDKLIR EVKVITLKSK LVSDFRKDFQ FYKVREINNY  1020
HHAHDAYLNA VVGTALIKKY PKLESEFVYG DYKVYDVRKM IAKSEQEIGK ATAKYFFYSN  1080
IMNFFKTEIT LANGEIRKRP LIETNGETGE IVWDKGRDFA TVRKVLSMPQ VNIVKKTEVQ  1140
TGGFSKESIL PKRNSDKLIA RKKDWDPKKY GGFDSPTVAY SVLVVAKVEK GKSKKLKSVK  1200
ELLGITIMER SSFEKNPIDF LEAKGYKEVK KDLIIKLPKY SLFELENGRK RMLASAGELQ  1260
KGNELALPSK YVNFLYLASH YEKLKGSPED NEQKQLFVEQ HKHYLDEIIE QISEFSKRVI  1320
LADANLDKVL SAYNKHRDKP IREQAENIIH LFTLTNLGAP AAFKYFDTTI DRKRYTSTKE  1380
VLDATLIHQS ITGLYETRID LSQLGGDKRP AATKKAGQAK KKK                    1423

SEQ ID NO: 47          moltype = AA  length = 483
FEATURE                Location/Qualifiers
REGION                 1..483
                       note = Description of Artificial Sequence: Synthetic
                        polypeptide
source                 1..483
                       mol_type = protein
                       organism = synthetic construct
SEQUENCE: 47
MFLFLSLTSF LSSSRTLVSK GEEDNMAIIK EFMRFKVHME GSVNGHEFEI EGEGEGRPYE  60
GTQTAKLKVT KGGPLPFAWD ILSPQFMYGS KAYVKHPADI PDYLKLSFPE GFKWERVMNF  120
EDGGVVTVTQ DSSLQDGEFI YKVKLRGTNF PSDGPVMQKK TMGWEASSER MYPEDGALKG  180
EIKQRLKLKD GGHYDAEVKT TYKAKKPVQL PGAYNVNIKL DITSHNEDYT IVEQYERAEG  240
RHSTGGMDEL YKGSKQLEEL LSTSFDIQFN DLTLLETAFT HTSYANEHRL LNVSHNERLE  300
FLGDAVLQLI ISEYLFAKYP KKTEGDMSKL RSMIVREESL AGFSRFCSFD AYIKLGKGEE  360
KSGGRRRDTI LGDLFEAFLG ALLLDKGIDA VRRFLKQVMI PQVEKGNFER VKDYKTCLQE  420
FLQTKGDVAI DYQVISEKGP AHAKQFEVSI VVNGAVLSKG LGKSKKLAEQ DAAKNALAQL  480
SEV                                                                483

SEQ ID NO: 48          moltype = AA  length = 483
FEATURE                Location/Qualifiers
REGION                 1..483
                       note = Description of Artificial Sequence: Synthetic
                        polypeptide
```

-continued

```
source                  1..483
                        mol_type = protein
                        organism = synthetic construct
SEQUENCE: 48
MKQLEELLST SFDIQFNDLT LLETAFTHTS YANEHRLLNV SHNERLEFLG DAVLQLIISE  60
YLFAKYPKKT EGDMSKLRSM IVREESLAGF SRFCSFDAYI KLGKGEEKSG GRRRDTILGD  120
LFEAFLGALL LDKGIDAVRR FLKQVMIPQV EKGNFERVKD YKTCLQEFLQ TKGDVAIDYQ  180
VISEKGPAHA KQFEVSIVVN GAVLSKGLGK SKKLAEQDAA KNALAQLSEV GSVSKGEEDN  240
MAIIKEFMRF KVHMEGSVNG HEFEIEGEGE GRPYEGTQTA KLKVTKGGPL PFAWDILSPQ  300
FMYGSKAYVK HPADIPDYLK LSFPEGFKWE RVMNFEDGGV VTVTQDSSLQ DGEFIYKVKL  360
RGTNFPSDGP VMQKKTMGWE ASSERMYPED GALKGEIKQR LKLKDGGHYD AEVKTTYKAK  420
KPVQLPGAYN VNIKLDITSH NEDYTIVEQY ERAEGRHSTG GMDELYKKRP AATKKAGQAK  480
KKK                                                                483

SEQ ID NO: 49           moltype = AA  length = 1423
FEATURE                 Location/Qualifiers
REGION                  1..1423
                        note = Description of Artificial Sequence: Synthetic
                         polypeptide
source                  1..1423
                        mol_type = protein
                        organism = synthetic construct
SEQUENCE: 49
MDYKDHDGDY KDHDIDYKDD DDKMAPKKKR KVGIHGVPAA DKKYSIGLAI GTNSVGWAVI  60
TDEYKVPSKK FKVLGNTDRH SIKKNLIGAL LFDSGETAEA TRLKRTARRR YTRRKNRICY  120
LQEIFSNEMA KVDDSFFHRL EESFLVEEDK KHERHPIFGN IVDEVAYHEK YPTIYHLRKK  180
LVDSTDKADL RLIYLALAHM IKFRGHFLIE GDLNPDNSDV DKLFIQLVQT YNQLFEENPI  240
NASGVDAKAI LSARLSKSRR LENLIAQLPG EKKNGLFGNL IALSLGLTPN FKSNFDLAED  300
AKLQLSKDTY DDDLDNLLAQ IGDQYADLFL AAKNLSDAIL LSDILRVNTE ITKAPLSASM  360
IKRYDEHHQD LTLLKALVRQ QLPEKYKEIF FDQSKNGYAG YIDGGASQEE FYKFIKPILE  420
KMDGTEELLV KLNREDLLRK QRTFDNGSIP HQIHLGELHA ILRRQEDFYP FLKDNREKIE  480
KILTFRIPYY VGPLARGNSR FAWMTRKSEE TITPWNFEEV VDKGASAQSF IERMTNFDKN  540
LPNEKVLPKH SLLYEYFTVY NELTKVKYVT EGMRKPAFLS GEQKKAIVDL LFKTNRKVTV  600
KQLKEDYFKK IECFDSVEIS GVEDRFNASL GTYHDLLKII KDKDFLDNEE NEDILEDIVL  660
TLTLFEDREM IEERLKTYAH LFDDKVMKQL KRRRYTGWGR LSRKLINGIR DKQSGKTILD  720
FLKSDGFANR NFMQLIHDDS LTFKEDIQKA QVSGQGDSLH EHIANLAGSP AIKKGILQTV  780
KVVDELVKVM GRHKPENIVI EMARENQTTQ KGQKNSRERM KRIEEGIKEL GSQILKEHPV  840
ENTQLQNEKL YLYYLQNGRD MYVDQELDIN RLSDYDVDHI VPQSFLKDDS IDNKVLTRSD  900
KNRGKSDNVP SEEVVKKMKN YWRQLLNAKL ITQRKFDNLT KAERGGLSEL DKAGFIKRQL  960
VETRQITKHV AQILDSRMNT KYDENDKLIR EVKVITLKSK LVSDFRKDFQ FYKVREINNY  1020
HHAHDAYLNA VVGTALIKKY PKLESEFVYG DYKVYDVRKM IAKSEQEIGK ATAKYFFYSN  1080
IMNFFKTEIT LANGEIRKRP LIETNGETGE IVWDKGRDFA TVRKVLSMPQ VNIVKKTEVQ  1140
TGGFSKESIL PKRNSDKLIA RKKDWDPKKY GGFDSPTVAY SVLVVAKVEK GKSKKLKSVK  1200
ELLGITIMER SSFEKNPIDF LEAKGYKEVK KDLIIKLPKY SLFELENGRK RMLASAGELQ  1260
KGNELALPSK YVNFLYLASH YEKLKGSPED NEQKQLFVEQ HKHYLDEIIE QISEFSKRVI  1320
LADANLDKVL SAYNKHRDKP IREQAENIIH LFTLTNLGAP AAFKYFDTTI DRKRYTSTKE  1380
VLDATLIHQS ITGLYETRID LSQLGGDKRP AATKKAGQAK KKK                    1423

SEQ ID NO: 50           moltype = DNA  length = 2012
FEATURE                 Location/Qualifiers
misc_feature            1..2012
                        note = Description of Artificial Sequence: Synthetic
                         polynucleotide
source                  1..2012
                        mol_type = other DNA
                        organism = synthetic construct
SEQUENCE: 50
gaatgctgcc ctcagacccg cttcctccct gtccttgtct gtccaaggag aatgaggtct  60
cactggtgga tttcggacta ccctgaggag ctggcacctg agggacaagg ccccccacct  120
gcccagctcc agcctctgat gaggggtggg agagagctac atgaggttgc taagaaagcc  180
tcccctgaag gagaccacac agtgtgtgag gttggagtct ctagcagcgg gttctgtgcc  240
cccagggata gtctggctgt ccaggcactg ctcttgatat aaacaccacc tcctagttat  300
gaaaccatgc ccattctgcc tctctgtatg gaaaagagca tggggctggc ccgtggggtg  360
gtgtccactt taggccctgt gggagatcat gggaacccac gcagtgggtc ataggctctc  420
tcatttacta ctcacatcca ctctgtgaag aagcgattat gatctctcct ctagaaactc  480
gtagagtccc atgtctgccg gcttccagag cctgcactcc tccaccttgg cttggctttg  540
ctggggctag aggagctagg atgcacagca gctctgtgac cctttgtttg agaggaacag  600
gaaaaccacc cttctctctg gcccactgtg tcctcttcct gccctgccat ccccttctgt  660
gaatgttaga cccatgggag cagctggtca gaggggaccc cggcctgggg cccctaaccc  720
tatgtagcct cagtcttccc atcaggctct cagctcagcc tgagtgttga ggccccagtg  780
gctgctctgg gggcctcctg agtttctcat ctgtgccccc ccctccctgg cccaggtgaa  840
ggtgtggttc cagaaccgga ggacaaagta caaacggcag aagctggagg aggaagggcc  900
tgagtccgag cagaagaaga agggctccca tcacatcaac cggtggcgca ttgccacgaa  960
gcaggccaat ggggaggaca tcgatgtcac ctccaatgac aagcttgcta gcggtgggca  1020
accacaaacc cacgagggca gagtgctgct tgctgctggc caggcccctg cgtgggccca  1080
agctggactc tggccactcc ctggccaggc tttggggagg cctggagtca tggccccaca  1140
```

-continued

```
gggcttgaag cccggggccg ccattgacag agggacaagc aatgggctgg ctgaggcctg  1200
ggaccacttg gccttctcct cggagagcct gcctgcctgg gcgggcccgc ccgccaccgc  1260
agcctcccag ctgctctccg tgtctccaat ctccctttg ttttgatgca tttctgtttt  1320
aatttatttt ccaggcacca ctgtagttta gtgatcccca gtgtcccct tccctatggg  1380
aataataaaa gtctctctct taatgacacg ggcatccagc tccagcccca gagcctgggg  1440
tggtagattc cggctctgag ggccagtggg ggctggtaga gcaaacgcgt tcagggcctg  1500
ggagcctggg gtggggtact ggtggagggg gtcaagggta attcattaac tcctctcttt  1560
tgttggggga ccctggtctc tacctccagc tccacagcag gagaaacagg ctagacatag  1620
ggaagggcca tcctgtatct tgagggagga caggcccagg tctttcttaa cgtattgaga  1680
ggtgggaatc aggcccaggt agttcaatgg gagagggaga gtgcttccct ctgcctagag  1740
actctggtgg cttctccagt tgaggagaaa ccagaggaaa ggggaggatt ggggtctggg  1800
ggagggaaca ccattcacaa aggctgacgg ttccagtccg aagtcgtggg cccaccagga  1860
tgctcacctg tccttggaga accgctgggc aggttgagac tgcagagaca gggcttaagg  1920
ctgagcctgc aaccagtccc cagtgactca gggcctcctc agcccaagaa agagcaacgt  1980
gccagggccc gctgagctct tgtgttcacc tg                                2012

SEQ ID NO: 51          moltype = AA  length = 1153
FEATURE                Location/Qualifiers
REGION                 1..1153
                       note = Description of Artificial Sequence: Synthetic
                        polypeptide
source                 1..1153
                       mol_type = protein
                       organism = synthetic construct
SEQUENCE: 51
MKRPAATKKA GQAKKKKSDL VLGLDIGIGS VGVGILNKVT GEIIHKNSRI FPAAQAENNL  60
VRRTNRQGRR LARRKKHRRV RLNRLFEESG LITDFTKISI NLNPYQLRVK GLTDELSNEE  120
LFIALKNMVK HRGISYLDDA SDDGNSSVGD YAQIVKENSK QLETKTPGQI QLERYQTYGQ  180
LRGDFTVEKD GKKHRLINVF PTSAYRSEAL RILQTQQEFN PQITDEFINR YLEILTGKRK  240
YYHGPGNEKS RTDYGRYRTS GETLDNIFGI LIGKCTFYPD EFRAAKASYT AQEFNLLNDL  300
NNLTVPTETK KLSKEQKNQI INYVKNEKAM GPAKLFKYIA KLLSCDVADI KGYRIDKSGK  360
AEIHTFEAYR KMKTLETLDI EQMDRETLDK LAYVLTLNTE REGIQEALEH EFADGSFSQK  420
QVDELVQFRK ANSSIFGKGW HNFSVKLMME LIPELYETSE EQMTILTRLG KQKTTSSSNK  480
TKYIDEKLLT EEIYNPVVAK SVRQAIKIVN AAIKEYGDFD NIVIEMARET NEDDEKKAIQ  540
KIQKANKDEK DAAMLKAANQ YNGKAELPHS VFHGHKQLAT KIRLWHQQGE RCLYTGKTIS  600
IHDLINNSNQ FEVDHILPLS ITFDDSLANK VLVYATANQE KGQRTPYQAL DSMDDAWSFR  660
ELKAFVRESK TLSNKKKEYL LTEEDISKFD VRKKFIERNL VDTRYASRVV LNALQEHFRA  720
HKIDTKVSVV RGQFTSQLRR HWGIEKTRDT YHHHAVDALI IAASSQLNLW KKQKNTLVSY  780
SEDQLLDIET GELISDDEYK ESVFKAPYQH FVDTLKSKEF EDSILFSYQV DSKFNRKISD  840
ATIYATRQAK VGKDKADETY VLGKIKDIYT QDGYDAFMKI YKKDKSKFLM YRHDPQTFEK  900
VIEPILENYP NKQINEKGKE VPCNPFLKYK EEHGYIRKYS KKGNGPEIKS LKYYDSKLGN  960
HIDITPKDSN NKVVLQSVSP WRADVYFNKT TGKYEILGLK YADLQFEKGT GTYKISQEKY  1020
NDIKKKEGVD SDSEFKFTLY KNDLLLVKDT ETKEQQLFRF LSRTMPKQKH YVELKPYDKQ  1080
KFEGGEALIK VLGNVANSGQ CKKGLGKSNI SIYKVRTDVL GNQHIIKNEG DKPKLDFKRP  1140
AATKKAGQAK KKK                                                     1153

SEQ ID NO: 52          moltype = DNA  length = 340
FEATURE                Location/Qualifiers
misc_feature           1..340
                       note = Description of Artificial Sequence: Synthetic
                        polynucleotide
source                 1..340
                       mol_type = other DNA
                       organism = synthetic construct
SEQUENCE: 52
gagggcctat ttcccatgat tccttcatat ttgcatatac gatacaaggc tgttagagag  60
ataattggaa ttaatttgac tgtaaacaca aagatattag tacaaaatac gtgacgtaga  120
aagtaataat ttcttgggta gtttgcagtt ttaaaattat gttttaaaat ggactatcat  180
atgcttaccg taacttgaaa gtatttcgat ttcttggctt tatatatctt gtggaaagga  240
cgaaacaccg ttacttaaat cttgcagaag ctacaaagat aaggcttcat gccgaaatca  300
acaccctgtc attttatggc agggtgtttt cgttatttaa                        340

SEQ ID NO: 53          moltype = DNA  length = 360
FEATURE                Location/Qualifiers
misc_feature           1..360
                       note = Description of Artificial Sequence: Synthetic
                        polynucleotide
misc_difference        288..317
                       note = a, c, t, g, unknown or other
source                 1..360
                       mol_type = other DNA
                       organism = synthetic construct
SEQUENCE: 53
gagggcctat ttcccatgat tccttcatat ttgcatatac gatacaaggc tgttagagag  60
ataattggaa ttaatttgac tgtaaacaca aagatattag tacaaaatac gtgacgtaga  120
aagtaataat ttcttgggta gtttgcagtt ttaaaattat gttttaaaat ggactatcat  180
```

-continued

```
atgcttaccg taacttgaaa gtatttcgat ttcttggctt tatatatctt gtggaaagga  240
cgaaacaccg ggttttagag ctatgctgtt ttgaatggtc ccaaaacnnn nnnnnnnnnn  300
nnnnnnnnnn nnnnnnngtt ttagagctat gctgttttga atggtcccaa aacttttttt  360

SEQ ID NO: 54          moltype = DNA  length = 318
FEATURE                Location/Qualifiers
misc_feature           1..318
                       note = Description of Artificial Sequence: Synthetic
                        polynucleotide
misc_difference        250..269
                       note = a, c, t, g, unknown or other
source                 1..318
                       mol_type = other DNA
                       organism = synthetic construct
SEQUENCE: 54
gagggcctat ttcccatgat tccttcatat ttgcatatac gatacaaggc tgttagagag  60
ataattggaa ttaatttgac tgtaaacaca aagatattag tacaaaatac gtgacgtaga  120
aagtaataat ttcttgggta gtttgcagtt ttaaaattat gttttaaaat ggactatcat  180
atgcttaccg taacttgaaa gtatttcgat ttcttggctt tatatatctt gtggaaagga  240
cgaaacaccn nnnnnnnnnn nnnnnnnnng ttttagagct agaaatagca agttaaaata  300
aggctagtcc gttttttt                                               318

SEQ ID NO: 55          moltype = DNA  length = 325
FEATURE                Location/Qualifiers
misc_feature           1..325
                       note = Description of Artificial Sequence: Synthetic
                        polynucleotide
misc_difference        250..269
                       note = a, c, t, g, unknown or other
source                 1..325
                       mol_type = other DNA
                       organism = synthetic construct
SEQUENCE: 55
gagggcctat ttcccatgat tccttcatat ttgcatatac gatacaaggc tgttagagag  60
ataattggaa ttaatttgac tgtaaacaca aagatattag tacaaaatac gtgacgtaga  120
aagtaataat ttcttgggta gtttgcagtt ttaaaattat gttttaaaat ggactatcat  180
atgcttaccg taacttgaaa gtatttcgat ttcttggctt tatatatctt gtggaaagga  240
cgaaacaccn nnnnnnnnnn nnnnnnnnng ttttagagct agaaatagca agttaaaata  300
aggctagtcc gttatcattt ttttt                                       325

SEQ ID NO: 56          moltype = DNA  length = 337
FEATURE                Location/Qualifiers
misc_feature           1..337
                       note = Description of Artificial Sequence: Synthetic
                        polynucleotide
misc_difference        250..269
                       note = a, c, t, g, unknown or other
source                 1..337
                       mol_type = other DNA
                       organism = synthetic construct
SEQUENCE: 56
gagggcctat ttcccatgat tccttcatat ttgcatatac gatacaaggc tgttagagag  60
ataattggaa ttaatttgac tgtaaacaca aagatattag tacaaaatac gtgacgtaga  120
aagtaataat ttcttgggta gtttgcagtt ttaaaattat gttttaaaat ggactatcat  180
atgcttaccg taacttgaaa gtatttcgat ttcttggctt tatatatctt gtggaaagga  240
cgaaacaccn nnnnnnnnnn nnnnnnnnng ttttagagct agaaatagca agttaaaata  300
aggctagtcc gttatcaact tgaaaaagtg ttttttt                           337

SEQ ID NO: 57          moltype = DNA  length = 352
FEATURE                Location/Qualifiers
misc_feature           1..352
                       note = Description of Artificial Sequence: Synthetic
                        polynucleotide
misc_difference        250..269
                       note = a, c, t, g, unknown or other
source                 1..352
                       mol_type = other DNA
                       organism = synthetic construct
SEQUENCE: 57
gagggcctat ttcccatgat tccttcatat ttgcatatac gatacaaggc tgttagagag  60
ataattggaa ttaatttgac tgtaaacaca aagatattag tacaaaatac gtgacgtaga  120
aagtaataat ttcttgggta gtttgcagtt ttaaaattat gttttaaaat ggactatcat  180
atgcttaccg taacttgaaa gtatttcgat ttcttggctt tatatatctt gtggaaagga  240
cgaaacaccn nnnnnnnnnn nnnnnnnnng ttttagagct agaaatagca agttaaaata  300
aggctagtcc gttatcaact tgaaaaagtg gcaccgagtc ggtgcttttt tt           352
```

-continued

SEQ ID NO: 58 moltype = DNA length = 5101
FEATURE Location/Qualifiers
misc_feature 1..5101
note = Description of Artificial Sequence: Synthetic polynucleotide
source 1..5101
mol_type = other DNA
organism = synthetic construct

SEQUENCE: 58

```
cgttacataa cttacggtaa atggcccgcc tggctgaccg cccaacgacc cccgcccatt  60
gacgtcaata atgacgtatg ttcccatagt aacgccaata gggactttcc attgacgtca  120
atgggtggag tatttacggt aaactgccca cttggcagta catcaagtgt atcatatgcc  180
aagtacgccc cctattgacg tcaatgacgg taaatggccc gcctggcatt atgcccagta  240
catgacctta tgggactttc ctacttggca gtacatctac gtattagtca tcgctattac  300
catggtcgag gtgagcccca cgttctgctt cactctcccc atctccccccc cctccccacc  360
cccaattttg tatttattta ttttttaatt attttgtgca gcgatggggg cgggggggg  420
ggggggcgc gcgccaggcg gggcggggcg gggcgagggg cggggcgggg cgaggcggag  480
aggtgcggcg gcagccaatc agagcggcgc gctccgaaag tttcctttta tggcgaggcg  540
gcggcggcgg cggccctata aaaagcgaag cgcgcggcgg gcgggagtcg ctgcgacgct  600
gccttcgccc cgtgccccgc tccgccgccg cctcgcgccg cccgccccgg ctctgactga  660
ccgcgttact cccacaggtg agcgggcggg acggcccttc tcctccgggc tgtaattagc  720
tgagcaagag gtaagggttt aagggatggt tggttggtgg ggtattaatg tttaattacc  780
tggagcacct gcctgaaatc actttttttc aggttggacc ggtgccacca tggactataa  840
ggaccacgac ggagactaca aggatcatga tattgattac aaagacgatg acgataagat  900
ggccccaaag aagaagcgga aggtcggtat ccacggagtc ccagcagccg acaagaagta  960
cagcatcggc ctggacatcg gcaccaactc tgtgggctgg gccgtgatca ccgacgagta  1020
caaggtgccc agcaagaaat tcaaggtgct gggcaacacc gaccggcaca gcatcaagaa  1080
gaacctgatc ggagccctgc tgttcgacag cggcgaaaca gccgaggcca cccggctgaa  1140
gagaaccgcc agaagaagat acaccagacg gaagaaccgg atctgctatc tgcaagagat  1200
cttcagcaac gagatggcca aggtggacga cagcttcttc cacagactgg aagagtcctt  1260
cctggtggaa gaggataaga agcacgagcg gcaccccatc ttcggcaaca tcgtggacga  1320
ggtggcctac cacgagaagt accccaccat ctaccacctg agaaagaaac tggtggacag  1380
caccgacaag gccgacctgc ggctgatcta tctggccctg gcccacatga tcaagttccg  1440
gggccacttc ctgatcgagg gcgacctgaa ccccgacaac agcgacgtgg acaagctgtt  1500
catccagctg gtgcagacct acaaccagct gttcgaggaa aaccccatca acgccagcgg  1560
cgtggacgcc aaggccatcc tgtctgccag actgagcaag agcagacggc tggaaaatct  1620
gatcgcccag ctgcccggcg agaagaagaa tggcctgttc ggcaacctga ttgccctgag  1680
cctgggcctg acccccaact tcaagagcaa cttcgacctg gccgaggatg ccaaactgca  1740
gctgagcaag gacacctacg acgacgacct ggacaacctg ctggcccaga tcggcgacca  1800
gtacgccgac ctgtttctgg ccgccaagaa cctgtccgac gccatcctgc tgagcgacat  1860
cctgagagtg aacaccgaga tcaccaaggc ccccctgagc gcctctatga tcaagagata  1920
cgacgagcac caccaggacc tgaccctgct gaaagctctc gtgcggcagc agctgcctga  1980
gaagtacaaa gagattttct tcgaccagag caagaacggc tacgccggct acattgacgg  2040
cggagccagc caggaagagt tctacaagtt catcaagccc atcctggaaa agatggacgg  2100
caccgaggaa ctgctcgtga agctgaacag agaggacctg ctgcggaagc agcggacctt  2160
cgacaacggc agcatccccc accagatcca cctgggagag ctgcacgcca ttctgcggcg  2220
gcaggaagat ttttacccat tcctgaagga caaccgggaa aagatcgaga agatcctgac  2280
cttccgcatc ccctactacg tgggccctct ggccagggga aacagcagat tcgcctggat  2340
gaccagaaag agcgaggaaa ccatcacccc ctggaacttc gaggaagtgg tggacaaggg  2400
cgcttccgcc cagagcttca tcgagcggat gaccaacttc gataagaacc tgcccaacga  2460
gaaggtgctg cccaagcaca gcctgctgta cgagtacttc accgtgtata acgagctgac  2520
caaagtgaaa tacgtgaccg agggaatgag aaagcccgcc ttcctgagcg gcgagcagaa  2580
aaaggccatc gtggacctgc tgttcaagac caaccggaaa gtgaccgtga agcagctgaa  2640
agaggactac ttcaagaaaa tcgagtgctt cgactccgtg gaaatctccg gcgtggaaga  2700
tcggttcaac gcctccctgg gcacatacca cgatctgctg aaaattatca aggacaagga  2760
cttcctggac aatgaggaaa acgaggacat tctggaagat atcgtgctga ccctgacact  2820
gtttgaggac agagagatga tcgaggaacg gctgaaaacc tatgcccacc tgttcgacga  2880
caaagtgatg aagcagctga agcggcggag atacaccggc tggggcaggc tgagccggaa  2940
gctgatcaac ggcatccggg acaagcagtc cggcaagaca atcctggatt tcctgaagtc  3000
cgacggcttc gccaacagaa acttcatgca gctgatccac gacgacagcc tgacctttaa  3060
agaggacatc cagaaagccc aggtgtccgg ccagggcgat agcctgcacg agcacattgc  3120
caatctggcc ggcagccccg ccattaagaa gggcatcctg cagacagtga aggtggtgga  3180
cgagctcgtg aaagtgatgg gccggcacaa gcccgagaac atcgtgatcg aaatggccag  3240
agagaaccag accacccaga agggacagaa gaacagccgc gagagaatga agcggatcga  3300
agagggcatc aaagagctgg gcagccagat cctgaaagaa caccccgtgg aaaacaccca  3360
gctgcagaac gagaagctgt acctgtacta cctgcagaat gggcgggata tgtacgtgga  3420
ccaggaactg gacatcaacc ggctgtccga ctacgatgtg gaccatatcg tgcctcagag  3480
ctttctgaag gacgactcca tcgacaacaa ggtgctgacc agaagcgaca agaaccgggg  3540
caagagcgac aacgtgccct ccgaagaggt cgtgaagaag atgaagaact actggcggca  3600
gctgctgaac gccaagctga ttacccagag aaagttcgac aatctgacca aggccgagag  3660
aggcggcctg agcgaactgg ataaggccgg cttcatcaag agacagctgg tggaaacccg  3720
gcagatcaca aagcacgtgg cacagatcct ggactcccgg atgaacacta agtacgacga  3780
gaatgacaag ctgatccggg aagtgaaagt gatcaccctg aagtccaagc tggtgtccga  3840
tttccggaag gatttccagt tttacaaagt gcgcgagatc aacaactacc accacgccca  3900
cgacgcctac ctgaacgccg tcgtgggaac cgccctgatc aaaaagtacc ctaagctgga  3960
aagcgagttc gtgtacggcg actacaaggt gtacgacgtg cggaagatga tcgccaagag  4020
```

-continued

```
cgagcaggaa atcggcaagg ctaccgccaa gtacttcttc tacagcaaca tcatgaactt  4080
tttcaagacc gagattaccc tggccaacgg cgagatccgg aagcggcctc tgatcgagac  4140
aaacggcgaa accggggaga tcgtgtggga taagggccgg gattttgcca ccgtgcggaa  4200
agtgctgagc atgccccaag tgaatatcgt gaaaaagacc gaggtgcaga caggcggctt  4260
cagcaaagag tctatcctgc ccaagaggaa cagcgataag ctgatcgcca gaaagaagga  4320
ctgggaccct aagaagtacg gcggcttcga cagccccacc gtggcctatt ctgtgctggt  4380
ggtggccaaa gtggaaaagg gcaagtccaa gaaactgaag agtgtgaaag agctgctggg  4440
gatcaccatc atggaaagaa gcagcttcga gaagaatccc atcgactttc tggaagccaa  4500
gggctacaaa gaagtgaaaa aggacctgat catcaagctg cctaagtact ccctgttcga  4560
gctggaaaac ggccggaaga gaatgctggc ctctgccggc gaactgcaga agggaaacga  4620
actggccctg ccctccaaat atgtgaactt cctgtacctg gccagccact atgagaagct  4680
gaagggctcc cccgaggata atgagcagaa acagctgttt gtggaacagc acaagcacta  4740
cctggacgag atcatcgagc agatcagcga gttctccaag agagtgatcc tggccgacgc  4800
taatctggac aaagtgctgt ccgcctacaa caagcaccgg gataagccca tcagagagca  4860
ggccgagaat atcatccacc tgtttaccct gaccaatctg ggagcccctg ccgccttcaa  4920
gtactttgac accaccatcg accggaagag gtacaccagc accaaagagg tgctggacgc  4980
caccctgatc caccagagca tcaccggcct gtacgagaca cggatcgacc tgtctcagct  5040
gggaggcgac tttctttttc ttagcttgac cagctttctt agtagcagca ggacgcttta  5100
a                                                                  5101

SEQ ID NO: 59          moltype = DNA  length = 137
FEATURE                Location/Qualifiers
misc_feature           1..137
                       note = Description of Artificial Sequence: Synthetic
                        polynucleotide
misc_difference        1..20
                       note = a, c, t, g, unknown or other
source                 1..137
                       mol_type = other DNA
                       organism = synthetic construct
SEQUENCE: 59
nnnnnnnnnn nnnnnnnnnn gtttttgtac tctcaagatt tagaaataaa tcttgcagaa  60
gctacaaaga taaggcttca tgccgaaatc aacaccctgt cattttatgg cagggtgttt  120
tcgttattta atttttt                                                137

SEQ ID NO: 60          moltype = DNA  length = 123
FEATURE                Location/Qualifiers
misc_feature           1..123
                       note = Description of Artificial Sequence: Synthetic
                        polynucleotide
misc_difference        1..20
                       note = a, c, t, g, unknown or other
source                 1..123
                       mol_type = other DNA
                       organism = synthetic construct
SEQUENCE: 60
nnnnnnnnnn nnnnnnnnnn gtttttgtac tctcagaaat gcagaagcta caaagataag  60
gcttcatgcc gaaatcaaca ccctgtcatt ttatggcagg gtgttttcgt tatttaattt  120
ttt                                                               123

SEQ ID NO: 61          moltype = DNA  length = 110
FEATURE                Location/Qualifiers
misc_feature           1..110
                       note = Description of Artificial Sequence: Synthetic
                        polynucleotide
misc_difference        1..20
                       note = a, c, t, g, unknown or other
source                 1..110
                       mol_type = other DNA
                       organism = synthetic construct
SEQUENCE: 61
nnnnnnnnnn nnnnnnnnnn gtttttgtac tctcagaaat gcagaagcta caaagataag  60
gcttcatgcc gaaatcaaca ccctgtcatt ttatggcagg gtgtttttt              110

SEQ ID NO: 62          moltype = DNA  length = 137
FEATURE                Location/Qualifiers
misc_feature           1..137
                       note = Description of Artificial Sequence: Synthetic
                        polynucleotide
misc_difference        1..20
                       note = a, c, t, g, unknown or other
source                 1..137
                       mol_type = other DNA
                       organism = synthetic construct
SEQUENCE: 62
nnnnnnnnnn nnnnnnnnnn gttattgtac tctcaagatt tagaaataaa tcttgcagaa  60
```

-continued

```
gctacaaaga taaggcttca tgccgaaatc aacaccctgt cattttatgg cagggtgttt  120
tcgttattta atttttt                                                 137

SEQ ID NO: 63          moltype = DNA  length = 123
FEATURE                Location/Qualifiers
misc_feature           1..123
                       note = Description of Artificial Sequence: Synthetic
                        polynucleotide
misc_difference        1..20
                       note = a, c, t, g, unknown or other
source                 1..123
                       mol_type = other DNA
                       organism = synthetic construct
SEQUENCE: 63
nnnnnnnnnn nnnnnnnnnn gttattgtac tctcagaaat gcagaagcta caaagataag  60
gcttcatgcc gaaatcaaca ccctgtcatt ttatggcagg gtgttttcgt tatttaattt  120
ttt                                                                123

SEQ ID NO: 64          moltype = DNA  length = 110
FEATURE                Location/Qualifiers
misc_feature           1..110
                       note = Description of Artificial Sequence: Synthetic
                        polynucleotide
misc_difference        1..20
                       note = a, c, t, g, unknown or other
source                 1..110
                       mol_type = other DNA
                       organism = synthetic construct
SEQUENCE: 64
nnnnnnnnnn nnnnnnnnnn gttattgtac tctcagaaat gcagaagcta caaagataag  60
gcttcatgcc gaaatcaaca ccctgtcatt ttatggcagg gtgttttttt              110

SEQ ID NO: 65          moltype = DNA  length = 137
FEATURE                Location/Qualifiers
misc_feature           1..137
                       note = Description of Artificial Sequence: Synthetic
                        polynucleotide
misc_difference        1..20
                       note = a, c, t, g, unknown or other
source                 1..137
                       mol_type = other DNA
                       organism = synthetic construct
SEQUENCE: 65
nnnnnnnnnn nnnnnnnnnn gttattgtac tctcaagatt tagaaataaa tcttgcagaa  60
gctacaatga taaggcttca tgccgaaatc aacaccctgt cattttatgg cagggtgttt  120
tcgttattta atttttt                                                 137

SEQ ID NO: 66          moltype = DNA  length = 123
FEATURE                Location/Qualifiers
misc_feature           1..123
                       note = Description of Artificial Sequence: Synthetic
                        polynucleotide
misc_difference        1..20
                       note = a, c, t, g, unknown or other
source                 1..123
                       mol_type = other DNA
                       organism = synthetic construct
SEQUENCE: 66
nnnnnnnnnn nnnnnnnnnn gttattgtac tctcagaaat gcagaagcta caatgataag  60
gcttcatgcc gaaatcaaca ccctgtcatt ttatggcagg gtgttttcgt tatttaattt  120
ttt                                                                123

SEQ ID NO: 67          moltype = DNA  length = 110
FEATURE                Location/Qualifiers
misc_feature           1..110
                       note = Description of Artificial Sequence: Synthetic
                        polynucleotide
misc_difference        1..20
                       note = a, c, t, g, unknown or other
source                 1..110
                       mol_type = other DNA
                       organism = synthetic construct
SEQUENCE: 67
nnnnnnnnnn nnnnnnnnnn gttattgtac tctcagaaat gcagaagcta caatgataag  60
gcttcatgcc gaaatcaaca ccctgtcatt ttatggcagg gtgttttttt              110
```

-continued

SEQ ID NO: 68 moltype = DNA length = 107
FEATURE Location/Qualifiers
misc_feature 1..107
note = Description of Artificial Sequence: Synthetic polynucleotide
misc_difference 1..20
note = a, c, t, g, unknown or other
source 1..107
mol_type = other DNA
organism = synthetic construct
SEQUENCE: 68
nnnnnnnnnn nnnnnnnnnn gttttagagc tgtggaaaca cagcgagtta aaataaggct 60
tagtccgtac tcaacttgaa aaggtggcac cgattcggtg ttttttt 107

SEQ ID NO: 69 moltype = DNA length = 4263
FEATURE Location/Qualifiers
misc_feature 1..4263
note = Description of Artificial Sequence: Synthetic polynucleotide
source 1..4263
mol_type = other DNA
organism = synthetic construct
SEQUENCE: 69
atgaaaaggc cggcggccac gaaaaaggcc ggccaggcaa aaaagaaaaa gaccaagccc 60
tacagcatcg gcctggacat cggcaccaat agcgtgggct gggccgtgac caccgacaac 120
tacaaggtgc ccagcaagaa aatgaaggtg ctgggcaaca cctccaagaa gtacatcaag 180
aaaaacctgc tgggcgtgct gctgttcgac agcggcatta cagccgaggg cagacggctg 240
aagagaaccg ccagacggcg gtacacccgg cggagaaaca gaatcctgta tctgcaagag 300
atcttcagca ccgagatggc taccctggac gacgccttct tccagcggct ggacgacagc 360
ttcctggtgc ccgacgacaa gcgggacagc aagtacccca tcttcggcaa cctggtggaa 420
gagaaggcct accacgacga gttccccacc atctaccacc tgagaaagta cctggccgac 480
agcaccaaga aggccgacct gagactggtg tatctggccc tggcccacat gatcaagtac 540
cggggccact tcctgatcga gggcgagttc aacagcaaga acaacgacat ccagaagaac 600
ttccaggact tcctggacac ctacaacgcc atcttcgaga gcgacctgtc cctggaaaac 660
agcaagcagc tggaagagat cgtgaaggac aagatcagca agctggaaaa gaaggaccgc 720
atcctgaagc tgttccccgg cgagaagaac agcggaatct tcagcgagtt tctgaagctg 780
atcgtgggca accaggccga cttcagaaag tgcttcaacc tggacgagaa agccagcctg 840
cacttcagca aagagagcta cgacgaggac ctggaaaccc tgctgggata tatcggcgac 900
gactacagcg acgtgttcct gaaggccaag aagctgtacg acgctatcct gctgagcggc 960
ttcctgaccg tgaccgacaa cgagacagag gccccactga gcagcgccat gattaagcgg 1020
tacaacgagc acaaagagga tctggctctg ctgaaagagt acatccggaa catcagcctg 1080
aaaacctaca atgaggtgtt caaggacgac accaagaacg gctacgccgg ctacatcgac 1140
ggcaagacca accaggaaga tttctatgtg tacctgaaga agctgctggc cgagttcgag 1200
ggggccgact actttctgga aaaaatcgac cgcgaggatt tcctgcggaa gcagcggacc 1260
ttcgacaacg gcagcatccc ctaccagatc catctgcagg aaatgcgggc catcctggac 1320
aagcaggcca agttctaccc attcctggcc aagaacaaag agcggatcga gaagatcctg 1380
accttccgca tcccttacta cgtgggcccc ctggccagag gcaacagcga ttttgcctgg 1440
tccatccgga agcgcaatga gaagatcacc ccctggaact tcgaggacgt gatcgacaaa 1500
gagtccagcg ccgaggcctt catcaaccgg atgaccagct tcgacctgta cctgcccgag 1560
gaaaaggtgc tgcccaagca cagcctgctg tacgagacat tcaatgtgta taacgagctg 1620
accaaagtgc ggtttatcgc cgagtctatg cgggactacc agttcctgga ctccaagcag 1680
aaaaaggaca tcgtgcggct gtacttcaag gacaagcgga aagtgaccga taaggacatc 1740
atcgagtacc tgcacgccat ctacggctac gatggcatcg agctgaaggg catcgagaag 1800
cagttcaact ccagcctgag cacataccac gacctgctga acattatcaa cgacaaagaa 1860
tttctggacg actccagcaa cgaggccatc atcgaagaga tcatccacac cctgaccatc 1920
tttgaggacc gcgagatgat caagcagcgg ctgagcaagt tcgagaacat cttcgacaag 1980
agcgtgctga aaaagctgag cagacggcac tacaccggct ggggcaagct gagcgccaag 2040
ctgatcaacg gcatccggga cgagaagtcc ggcaacacaa tcctggacta cctgatcgac 2100
gacggcatca gcaaccggaa cttcatgcag ctgatccacg acgacgccct gagcttcaag 2160
aagaagatcc agaaggccca gatcatcggg gacgaggaca agggcaacat caaagaagtc 2220
gtgaagtccc tgcccggcag ccccgccatc aagaagggaa tcctgcagag catcaagatc 2280
gtggacgagc tcgtgaaagt gatgggcggc agaaagcccg agagcatcgt ggtggaaatg 2340
gctagagaga accagtacac caatcagggc aagagcaaca gccagcagag actgaagaga 2400
ctggaaaagt ccctgaaaga gctgggcagc aagattctga aagagaatat ccctgccaag 2460
ctgtccaaga tcgacaacaa cgccctgcag aacgaccggc tgtacctgta ctacctgcag 2520
aatggcaagg acatgtatac aggcgacgac ctggatatcg accgcctgag caactacgac 2580
atcgaccata ttatccccca ggccttcctg aaagacaaca gcattgacaa caaagtgctg 2640
gtgtcctccg ccagcaaccg cggcaagtcc gatgatgtgc ccagcctgga agtcgtgaaa 2700
aagagaaaga ccttctggta tcagctgctg aaaagcaagc tgattagcca gaggaagttc 2760
gacaacctga ccaaggccga gagaggcggc ctgagccctg aagataaggc cggcttcatc 2820
cagagacagc tggtggaaac ccggcagatc accaagcacg tggccagact gctggatgag 2880
aagtttaaca acaagaagga cgagaacaac cgggccgtgc ggaccgtgaa gatcatcacc 2940
ctgaagtcca ccctggtgtc ccagttccgg aaggacttcg agctgtataa agtgcgcgag 3000
atcaatgact ttcaccacgc ccacgacgcc tacctgaatg ccgtggtggc ttccgccctg 3060
ctgaagaagt accctaagct ggaacccgag ttcgtgtacg gcgactaccc caagtacaac 3120
tccttcagag agcggaagtc cgccaccgag aaggtgtact tctactccaa catcatgaat 3180

-continued

```
atctttaaga agtccatctc cctggccgat ggcagagtga tcgagcggcc cctgatcgaa  3240
gtgaacgaag agacaggcga gagcgtgtgg aacaaagaaa gcgacctggc caccgtgcgg  3300
cgggtgctga gttatcctca agtgaatgtc gtgaagaagg tggaagaaca gaaccacggc  3360
ctggatcggg gcaagcccaa gggcctgttc aacgccaacc tgtccagcaa gcctaagccc  3420
aactccaacg agaatctcgt gggggccaaa gagtacctgg accctaagaa gtacggcgga  3480
tacgccggca tctccaatag cttcaccgtg ctcgtgaagg gcacaatcga gaagggcgct  3540
aagaaaaaga tcacaaacgt gctggaattt caggggatct ctatcctgga ccggatcaac  3600
taccggaagg ataagctgaa ctttctgctg gaaaaaggct acaaggacat tgagctgatt  3660
atcgagctgc ctaagtactc cctgttcgaa ctgagcgacg gctccagacg gatgctggcc  3720
tccatcctgt ccaccaacaa caagcggggc gagatccaca agggaaacca gatcttcctg  3780
agccagaaat ttgtgaaact gctgtaccac gccaagcgga tctccaacac catcaatgag  3840
aaccaccgga aatacgtgga aaaccacaag aaagagtttg aggaactgtt ctactacatc  3900
ctggagttca acgagaacta tgtgggagcc aagaagaacg gcaaactgct gaactccgcc  3960
ttccagagct ggcagaacca cagcatcgac gagctgtgca gctccttcat cggccctacc  4020
ggcagcgagc ggaagggact gtttgagctg acctccagag gctctgccgc cgactttgag  4080
ttcctgggag tgaagatccc ccggtacaga gactacaccc cctctagtct gctgaaggac  4140
gccaccctga tccaccagag cgtgaccggc ctgtacgaaa cccggatcga cctggctaag  4200
ctgggcgagg gaaagcgtcc tgctgctact aagaaagctg gtcaagctaa gaaaaagaaa  4260
taa                                                               4263

SEQ ID NO: 70          moltype = DNA  length = 84
FEATURE                Location/Qualifiers
misc_feature           1..84
                       note = Description of Artificial Sequence: Synthetic
                        oligonucleotide
source                 1..84
                       mol_type = other DNA
                       organism = synthetic construct
SEQUENCE: 70
ggaaccattc ataacagcat agcaagttat aataaggcta gtccgttatc aacttgaaaa  60
agtggcaccg agtcggtgct tttt                                        84

SEQ ID NO: 71          moltype = DNA  length = 36
FEATURE                Location/Qualifiers
misc_feature           1..36
                       note = Description of Artificial Sequence: Synthetic
                        oligonucleotide
source                 1..36
                       mol_type = other DNA
                       organism = synthetic construct
SEQUENCE: 71
gttatagagc tatgctgtta tgaatggtcc caaaac                           36

SEQ ID NO: 72          moltype = DNA  length = 84
FEATURE                Location/Qualifiers
misc_feature           1..84
                       note = Description of Artificial Sequence: Synthetic
                        oligonucleotide
source                 1..84
                       mol_type = other DNA
                       organism = synthetic construct
SEQUENCE: 72
ggaaccattc aatacagcat agcaagttaa tataaggcta gtccgttatc aacttgaaaa  60
agtggcaccg agtcggtgct tttt                                        84

SEQ ID NO: 73          moltype = DNA  length = 36
FEATURE                Location/Qualifiers
misc_feature           1..36
                       note = Description of Artificial Sequence: Synthetic
                        oligonucleotide
source                 1..36
                       mol_type = other DNA
                       organism = synthetic construct
SEQUENCE: 73
gtattagagc tatgctgtat tgaatggtcc caaaac                           36

SEQ ID NO: 74          moltype = DNA  length = 103
FEATURE                Location/Qualifiers
misc_feature           1..103
                       note = Description of Artificial Sequence: Synthetic
                        polynucleotide
misc_difference        1..20
                       note = a, c, t, g, unknown or other
source                 1..103
                       mol_type = other DNA
                       organism = synthetic construct
```

-continued

```
SEQUENCE: 74
nnnnnnnnnn nnnnnnnnnn gttttagagc tagaaatagc aagttaaaat aaggctagtc  60
cgttatcaac ttgaaaaagt ggcaccgagt cggtgctttt ttt                   103

SEQ ID NO: 75          moltype = DNA  length = 103
FEATURE                Location/Qualifiers
misc_feature           1..103
                       note = Description of Artificial Sequence: Synthetic
                        polynucleotide
misc_difference        1..20
                       note = a, c, t, g, unknown or other
source                 1..103
                       mol_type = other DNA
                       organism = synthetic construct
SEQUENCE: 75
nnnnnnnnnn nnnnnnnnnn gtattagagc tagaaatagc aagttaatat aaggctagtc  60
cgttatcaac ttgaaaaagt ggcaccgagt cggtgctttt ttt                   103

SEQ ID NO: 76          moltype = DNA  length = 123
FEATURE                Location/Qualifiers
misc_feature           1..123
                       note = Description of Artificial Sequence: Synthetic
                        polynucleotide
misc_difference        1..20
                       note = a, c, t, g, unknown or other
source                 1..123
                       mol_type = other DNA
                       organism = synthetic construct
SEQUENCE: 76
nnnnnnnnnn nnnnnnnnnn gttttagagc tatgctgttt tggaaacaaa acagcatagc  60
aagttaaaat aaggctagtc cgttatcaac ttgaaaaagt ggcaccgagt cggtgctttt  120
ttt                                                               123

SEQ ID NO: 77          moltype = DNA  length = 123
FEATURE                Location/Qualifiers
misc_feature           1..123
                       note = Description of Artificial Sequence: Synthetic
                        polynucleotide
misc_difference        1..20
                       note = a, c, t, g, unknown or other
source                 1..123
                       mol_type = other DNA
                       organism = synthetic construct
SEQUENCE: 77
nnnnnnnnnn nnnnnnnnnn gtattagagc tatgctgtat tggaaacaat acagcatagc  60
aagttaatat aaggctagtc cgttatcaac ttgaaaaagt ggcaccgagt cggtgctttt  120
ttt                                                               123

SEQ ID NO: 78          moltype = DNA  length = 20
FEATURE                Location/Qualifiers
source                 1..20
                       mol_type = unassigned DNA
                       organism = Homo sapiens
SEQUENCE: 78
gtcacctcca atgactaggg                                             20

SEQ ID NO: 79          moltype = AA  length = 984
FEATURE                Location/Qualifiers
source                 1..984
                       mol_type = protein
                       organism = Campylobacter jejuni
SEQUENCE: 79
MARILAFDIG ISSIGWAFSE NDELKDCGVR IFTKVENPKT GESLALPRRL ARSARKRLAR  60
RKARLNHLKH LIANEFKLNY EDYQSFDESL AKAYKGSLIS PYELRFRALN ELLSKQDFAR  120
VILHIAKRRG YDDIKNSDDK EKGAILKAIK QNEEKLANYQ SVGEYLYKEY FQKFKENSKE  180
FTNVRNKKES YERCIAQSFL KDELKLIFKK QREFGFSFSK KFEEEVLSVA FYKRALKDFS  240
HLVGNCSFFT DEKRAPKNSP LAFMFVALTR IINLLNNLKN TEGILYTKDD LNALLNEVLK  300
NGTLTYKQTK KLLGLSDDYE FKGEKGTYFI EFKKYKEFIK ALGEHNLSQD DLNEIAKDIT  360
LIKDEIKLKK ALAKYDLNQN QIDSLSKLEF KDHLNISFKA LKLVTPLMLE GKKYDEACNE  420
LNLKVAINED KKDFLPAFNE TYYKDEVTNP VVLRAIKEYR KVLNALLKKY GKVHKINIEL  480
AREVGKNHSQ RAKIEKEQNE NYKAKKDAEL ECEKLGLKIN SKNILKLRLF KEQKEFCAYS  540
GEKIKISDLQ DEKMLEIDHI YPYSRSFDDS YMNKVLVFTK QNQEKLNQTP FEAFGNDSAK  600
WQKIEVLAKN LPTKKQKRIL DKNYKDKEQK NFKDRNLNDT RYIARLVLNY TKDYLDFLPL  660
SDDENTKLND TQKGSKVHVE AKSGMLTSAL RHTWGFSAKD RNNHLHHAID AVIIAYANNS  720
IVKAFSDFKK EQESNSAELY AKKISELDYK NKRKFFEPFS GFRQKVLDKI DEIFVSKPER  780
KKPSGALHEE TFRKEEEFYQ SYGGKEGVLK ALELGKIRKV NGKIVKNGDM FRVDIFKHKK  840
```

-continued

```
TNKFYAVPIY TMDFALKVLP NKAVARSKKG EIKDWILMDE NYEFCFSLYK DSLILIQTKD  900
MQEPEFVYYN AFTSSTVSLI VSKHDNKFET LSKNQKILFK NANEKEVIAK SIGIQNLKVF  960
EKYIVSALGE VTKAEFRQRE DFKK                                         984

SEQ ID NO: 80          moltype = DNA  length = 91
FEATURE                Location/Qualifiers
misc_feature           1..91
                       note = Description of Artificial Sequence: Synthetic
                        oligonucleotide
source                 1..91
                       mol_type = other DNA
                       organism = synthetic construct
SEQUENCE: 80
tataatctca taagaaattt aaaaagggac taaaataaag agtttgcggg actctgcggg  60
gttacaatcc cctaaaaccg cttttaaaat t                                 91

SEQ ID NO: 81          moltype = DNA  length = 36
FEATURE                Location/Qualifiers
misc_feature           1..36
                       note = Description of Artificial Sequence: Synthetic
                        oligonucleotide
source                 1..36
                       mol_type = other DNA
                       organism = synthetic construct
SEQUENCE: 81
attttaccat aaagaaattt aaaaagggac taaaac                            36

SEQ ID NO: 82          moltype = RNA  length = 95
FEATURE                Location/Qualifiers
misc_feature           1..95
                       note = Description of Artificial Sequence: Synthetic
                        oligonucleotide
misc_difference        1..20
                       note = a, c, u, g, unknown or other
source                 1..95
                       mol_type = other RNA
                       organism = synthetic construct
SEQUENCE: 82
nnnnnnnnnn nnnnnnnnnn gttttagtcc cgaaagggac taaaataaag agtttgcggg  60
actctgcggg gttacaatcc cctaaaaccg ctttt                             95

SEQ ID NO: 83          moltype = RNA  length = 69
FEATURE                Location/Qualifiers
misc_feature           1..69
                       note = Description of Artificial Sequence: Synthetic
                        oligonucleotide
source                 1..69
                       mol_type = other RNA
                       organism = synthetic construct
SEQUENCE: 83
gtcacctcca atgactaggg gttttagagc tagaaatagc aagttaaaat aaggctagtc  60
cgttttttt                                                          69

SEQ ID NO: 84          moltype = RNA  length = 69
FEATURE                Location/Qualifiers
misc_feature           1..69
                       note = Description of Artificial Sequence: Synthetic
                        oligonucleotide
source                 1..69
                       mol_type = other RNA
                       organism = synthetic construct
SEQUENCE: 84
gacatcgatg tcctccccat gttttagagc tagaaatagc aagttaaaat aaggctagtc  60
cgttttttt                                                          69

SEQ ID NO: 85          moltype = RNA  length = 69
FEATURE                Location/Qualifiers
misc_feature           1..69
                       note = Description of Artificial Sequence: Synthetic
                        oligonucleotide
source                 1..69
                       mol_type = other RNA
                       organism = synthetic construct
SEQUENCE: 85
gagtccgagc agaagaagaa gttttagagc tagaaatagc aagttaaaat aaggctagtc  60
cgttttttt                                                          69
```

-continued

```
SEQ ID NO: 86          moltype = RNA  length = 69
FEATURE                Location/Qualifiers
misc_feature           1..69
                       note = Description of Artificial Sequence: Synthetic
                        oligonucleotide
source                 1..69
                       mol_type = other RNA
                       organism = synthetic construct
SEQUENCE: 86
ggggccgaga ttgggtgttc gttttagagc tagaaatagc aagttaaaat aaggctagtc  60
cgttttttt                                                          69

SEQ ID NO: 87          moltype = RNA  length = 69
FEATURE                Location/Qualifiers
misc_feature           1..69
                       note = Description of Artificial Sequence: Synthetic
                        oligonucleotide
source                 1..69
                       mol_type = other RNA
                       organism = synthetic construct
SEQUENCE: 87
gtggcgagag gggccgagat gttttagagc tagaaatagc aagttaaaat aaggctagtc  60
cgttttttt                                                          69

SEQ ID NO: 88          moltype = RNA  length = 76
FEATURE                Location/Qualifiers
misc_feature           1..76
                       note = Description of Artificial Sequence: Synthetic
                        oligonucleotide
source                 1..76
                       mol_type = other RNA
                       organism = synthetic construct
SEQUENCE: 88
gtcacctcca atgactaggg gttttagagc tagaaatagc aagttaaaat aaggctagtc  60
cgttatcatt tttttt                                                  76

SEQ ID NO: 89          moltype = RNA  length = 76
FEATURE                Location/Qualifiers
misc_feature           1..76
                       note = Description of Artificial Sequence: Synthetic
                        oligonucleotide
source                 1..76
                       mol_type = other RNA
                       organism = synthetic construct
SEQUENCE: 89
gacatcgatg tcctccccat gttttagagc tagaaatagc aagttaaaat aaggctagtc  60
cgttatcatt tttttt                                                  76

SEQ ID NO: 90          moltype = RNA  length = 76
FEATURE                Location/Qualifiers
misc_feature           1..76
                       note = Description of Artificial Sequence: Synthetic
                        oligonucleotide
source                 1..76
                       mol_type = other RNA
                       organism = synthetic construct
SEQUENCE: 90
gagtccgagc agaagaagaa gttttagagc tagaaatagc aagttaaaat aaggctagtc  60
cgttatcatt tttttt                                                  76

SEQ ID NO: 91          moltype = RNA  length = 76
FEATURE                Location/Qualifiers
misc_feature           1..76
                       note = Description of Artificial Sequence: Synthetic
                        oligonucleotide
source                 1..76
                       mol_type = other RNA
                       organism = synthetic construct
SEQUENCE: 91
ggggccgaga ttgggtgttc gttttagagc tagaaatagc aagttaaaat aaggctagtc  60
cgttatcatt tttttt                                                  76

SEQ ID NO: 92          moltype = RNA  length = 76
FEATURE                Location/Qualifiers
misc_feature           1..76
```

-continued

```
                     note = Description of Artificial Sequence: Synthetic
                      oligonucleotide
source               1..76
                     mol_type = other RNA
                     organism = synthetic construct
SEQUENCE: 92
gtggcgagag gggccgagat gttttagagc tagaaatagc aagttaaaat aaggctagtc  60
cgttatcatt tttttt                                                  76

SEQ ID NO: 93        moltype = RNA  length = 88
FEATURE              Location/Qualifiers
misc_feature         1..88
                     note = Description of Artificial Sequence: Synthetic
                      oligonucleotide
source               1..88
                     mol_type = other RNA
                     organism = synthetic construct
SEQUENCE: 93
gtcacctcca atgactaggg gttttagagc tagaaatagc aagttaaaat aaggctagtc  60
cgttatcaac ttgaaaaagt gttttttt                                     88

SEQ ID NO: 94        moltype = RNA  length = 88
FEATURE              Location/Qualifiers
misc_feature         1..88
                     note = Description of Artificial Sequence: Synthetic
                      oligonucleotide
source               1..88
                     mol_type = other RNA
                     organism = synthetic construct
SEQUENCE: 94
gacatcgatg tcctccccat gttttagagc tagaaatagc aagttaaaat aaggctagtc  60
cgttatcaac ttgaaaaagt gttttttt                                     88

SEQ ID NO: 95        moltype = RNA  length = 88
FEATURE              Location/Qualifiers
misc_feature         1..88
                     note = Description of Artificial Sequence: Synthetic
                      oligonucleotide
source               1..88
                     mol_type = other RNA
                     organism = synthetic construct
SEQUENCE: 95
gagtccgagc agaagaagaa gttttagagc tagaaatagc aagttaaaat aaggctagtc  60
cgttatcaac ttgaaaaagt gttttttt                                     88

SEQ ID NO: 96        moltype = RNA  length = 88
FEATURE              Location/Qualifiers
misc_feature         1..88
                     note = Description of Artificial Sequence: Synthetic
                      oligonucleotide
source               1..88
                     mol_type = other RNA
                     organism = synthetic construct
SEQUENCE: 96
ggggccgaga ttgggtgttc gttttagagc tagaaatagc aagttaaaat aaggctagtc  60
cgttatcaac ttgaaaaagt gttttttt                                     88

SEQ ID NO: 97        moltype = RNA  length = 88
FEATURE              Location/Qualifiers
misc_feature         1..88
                     note = Description of Artificial Sequence: Synthetic
                      oligonucleotide
source               1..88
                     mol_type = other RNA
                     organism = synthetic construct
SEQUENCE: 97
gtggcgagag gggccgagat gttttagagc tagaaatagc aagttaaaat aaggctagtc  60
cgttatcaac ttgaaaaagt gttttttt                                     88

SEQ ID NO: 98        moltype = RNA  length = 103
FEATURE              Location/Qualifiers
misc_feature         1..103
                     note = Description of Artificial Sequence: Synthetic
                      oligonucleotide
source               1..103
                     mol_type = other RNA
```

-continued

```
                     organism = synthetic construct
SEQUENCE: 98
gtcacctcca atgactaggg gttttagagc tagaaatagc aagttaaaat aaggctagtc  60
cgttatcaac ttgaaaaagt ggcaccgagt cggtgctttt ttt                    103

SEQ ID NO: 99          moltype = RNA  length = 103
FEATURE                Location/Qualifiers
misc_feature           1..103
                       note = Description of Artificial Sequence: Synthetic
                        oligonucleotide
source                 1..103
                       mol_type = other RNA
                       organism = synthetic construct
SEQUENCE: 99
gacatcgatg tcctccccat gttttagagc tagaaatagc aagttaaaat aaggctagtc  60
cgttatcaac ttgaaaaagt ggcaccgagt cggtgctttt ttt                    103

SEQ ID NO: 100         moltype = RNA  length = 103
FEATURE                Location/Qualifiers
misc_feature           1..103
                       note = Description of Artificial Sequence: Synthetic
                        oligonucleotide
source                 1..103
                       mol_type = other RNA
                       organism = synthetic construct
SEQUENCE: 100
gagtccgagc agaagaagaa gttttagagc tagaaatagc aagttaaaat aaggctagtc  60
cgttatcaac ttgaaaaagt ggcaccgagt cggtgctttt ttt                    103

SEQ ID NO: 101         moltype = RNA  length = 103
FEATURE                Location/Qualifiers
misc_feature           1..103
                       note = Description of Artificial Sequence: Synthetic
                        oligonucleotide
source                 1..103
                       mol_type = other RNA
                       organism = synthetic construct
SEQUENCE: 101
ggggccgaga ttgggtgttc gttttagagc tagaaatagc aagttaaaat aaggctagtc  60
cgttatcaac ttgaaaaagt ggcaccgagt cggtgctttt ttt                    103

SEQ ID NO: 102         moltype = RNA  length = 103
FEATURE                Location/Qualifiers
misc_feature           1..103
                       note = Description of Artificial Sequence: Synthetic
                        oligonucleotide
source                 1..103
                       mol_type = other RNA
                       organism = synthetic construct
SEQUENCE: 102
gtggcgagag gggccgagat gttttagagc tagaaatagc aagttaaaat aaggctagtc  60
cgttatcaac ttgaaaaagt ggcaccgagt cggtgctttt ttt                    103

SEQ ID NO: 103         moltype = DNA  length = 102
FEATURE                Location/Qualifiers
misc_feature           1..102
                       note = Description of Artificial Sequence: Synthetic
                        polynucleotide
source                 1..102
                       mol_type = other DNA
                       organism = synthetic construct
SEQUENCE: 103
gttttagagc tatgctgttt tgaatggtcc caaaacggaa gggcctgagt ccgagcagaa  60
gaagaagttt tagagctatg ctgttttgaa tggtcccaaa ac                     102

SEQ ID NO: 104         moltype = DNA  length = 100
FEATURE                Location/Qualifiers
source                 1..100
                       mol_type = unassigned DNA
                       organism = Homo sapiens
SEQUENCE: 104
cggaggacaa agtacaaacg gcagaagctg gaggaggaag ggcctgagtc cgagcagaag  60
aagaagggct cccatcacat caaccggtgg cgcattgcca                        100

SEQ ID NO: 105         moltype = DNA  length = 50
FEATURE                Location/Qualifiers
```

-continued

```
source                  1..50
                        mol_type = unassigned DNA
                        organism = Homo sapiens
SEQUENCE: 105
agctggagga ggaagggcct gagtccgagc agaagaagaa gggctcccac              50

SEQ ID NO: 106          moltype = RNA  length = 30
FEATURE                 Location/Qualifiers
misc_feature            1..30
                        note = Description of Artificial Sequence: Synthetic
                         oligonucleotide
source                  1..30
                        mol_type = other RNA
                        organism = synthetic construct
SEQUENCE: 106
gagtccgagc agaagaagaa gttttagagc                                    30

SEQ ID NO: 107          moltype = DNA  length = 49
FEATURE                 Location/Qualifiers
misc_feature            1..49
                        note = Description of Artificial Sequence: Synthetic
                         oligonucleotide
source                  1..49
                        mol_type = other DNA
                        organism = synthetic construct
SEQUENCE: 107
agctggagga ggaagggcct gagtccgagc agaagagaag ggctcccat               49

SEQ ID NO: 108          moltype = DNA  length = 53
FEATURE                 Location/Qualifiers
source                  1..53
                        mol_type = unassigned DNA
                        organism = Homo sapiens
SEQUENCE: 108
ctggaggagg aagggcctga gtccgagcag aagaagaagg gctcccatca cat          53

SEQ ID NO: 109          moltype = DNA  length = 52
FEATURE                 Location/Qualifiers
source                  1..52
                        mol_type = unassigned DNA
                        organism = Homo sapiens
SEQUENCE: 109
ctggaggagg aagggcctga gtccgagcag aagagaaggg ctcccatcac at           52

SEQ ID NO: 110          moltype = DNA  length = 54
FEATURE                 Location/Qualifiers
source                  1..54
                        mol_type = unassigned DNA
                        organism = Homo sapiens
SEQUENCE: 110
ctggaggagg aagggcctga gtccgagcag aagaaagaag ggctcccatc acat         54

SEQ ID NO: 111          moltype = DNA  length = 50
FEATURE                 Location/Qualifiers
source                  1..50
                        mol_type = unassigned DNA
                        organism = Homo sapiens
SEQUENCE: 111
ctggaggagg aagggcctga gtccgagcag aagaagggct cccatcacat              50

SEQ ID NO: 112          moltype = DNA  length = 47
FEATURE                 Location/Qualifiers
source                  1..47
                        mol_type = unassigned DNA
                        organism = Homo sapiens
SEQUENCE: 112
ctggaggagg aagggcctga gcccgagcag aagggctccc atcacat                 47

SEQ ID NO: 113          moltype = RNA  length = 66
FEATURE                 Location/Qualifiers
misc_feature            1..66
                        note = Description of Artificial Sequence: Synthetic
                         oligonucleotide
misc_difference         1..20
                        note = a, c, t, g, unknown or other
source                  1..66
```

-continued

```
                     mol_type = other RNA
                     organism = synthetic construct
modified_base        56
                     mod_base = OTHER
                     note = Thymine
modified_base        59
                     mod_base = OTHER
                     note = Thymine
SEQUENCE: 113
nnnnnnnnnn nnnnnnnnnn gttttagagc tagaaatagc aagttaaaat aaggctagtc  60
cgtttt                                                            66

SEQ ID NO: 114       moltype = RNA  length = 20
FEATURE              Location/Qualifiers
misc_feature         1..20
                     note = Description of Artificial Sequence: Synthetic
                      oligonucleotide
source               1..20
                     mol_type = other RNA
                     organism = synthetic construct
SEQUENCE: 114
gagtccgagc agaagaagaa                                              20

SEQ ID NO: 115       moltype = RNA  length = 20
FEATURE              Location/Qualifiers
misc_feature         1..20
                     note = Description of Artificial Sequence: Synthetic
                      oligonucleotide
source               1..20
                     mol_type = other RNA
                     organism = synthetic construct
SEQUENCE: 115
gacatcgatg tcctccccat                                              20

SEQ ID NO: 116       moltype = RNA  length = 20
FEATURE              Location/Qualifiers
misc_feature         1..20
                     note = Description of Artificial Sequence: Synthetic
                      oligonucleotide
source               1..20
                     mol_type = other RNA
                     organism = synthetic construct
SEQUENCE: 116
gtcacctcca atgactaggg                                              20

SEQ ID NO: 117       moltype = RNA  length = 20
FEATURE              Location/Qualifiers
misc_feature         1..20
                     note = Description of Artificial Sequence: Synthetic
                      oligonucleotide
source               1..20
                     mol_type = other RNA
                     organism = synthetic construct
SEQUENCE: 117
attgggtgtt cagggcagag                                              20

SEQ ID NO: 118       moltype = RNA  length = 20
FEATURE              Location/Qualifiers
misc_feature         1..20
                     note = Description of Artificial Sequence: Synthetic
                      oligonucleotide
source               1..20
                     mol_type = other RNA
                     organism = synthetic construct
SEQUENCE: 118
gtggcgagag gggccgagat                                              20

SEQ ID NO: 119       moltype = RNA  length = 20
FEATURE              Location/Qualifiers
misc_feature         1..20
                     note = Description of Artificial Sequence: Synthetic
                      oligonucleotide
source               1..20
                     mol_type = other RNA
                     organism = synthetic construct
SEQUENCE: 119
```

-continued

```
ggggccgaga ttgggtgttc                                              20

SEQ ID NO: 120          moltype = RNA  length = 20
FEATURE                 Location/Qualifiers
misc_feature            1..20
                        note = Description of Artificial Sequence: Synthetic
                         oligonucleotide
source                  1..20
                        mol_type = other RNA
                        organism = synthetic construct
SEQUENCE: 120
gtgccattag ctaaatgcat                                              20

SEQ ID NO: 121          moltype = RNA  length = 20
FEATURE                 Location/Qualifiers
misc_feature            1..20
                        note = Description of Artificial Sequence: Synthetic
                         oligonucleotide
source                  1..20
                        mol_type = other RNA
                        organism = synthetic construct
SEQUENCE: 121
gtaccaccca caggtgccag                                              20

SEQ ID NO: 122          moltype = RNA  length = 20
FEATURE                 Location/Qualifiers
misc_feature            1..20
                        note = Description of Artificial Sequence: Synthetic
                         oligonucleotide
source                  1..20
                        mol_type = other RNA
                        organism = synthetic construct
SEQUENCE: 122
gaaagcctct gggccaggaa                                              20

SEQ ID NO: 123          moltype = DNA  length = 48
FEATURE                 Location/Qualifiers
source                  1..48
                        mol_type = unassigned DNA
                        organism = Homo sapiens
SEQUENCE: 123
ctggaggagg aagggcctga gtccgagcag aagaagaagg gctcccat               48

SEQ ID NO: 124          moltype = RNA  length = 20
FEATURE                 Location/Qualifiers
misc_feature            1..20
                        note = Description of Artificial Sequence: Synthetic
                         oligonucleotide
source                  1..20
                        mol_type = other RNA
                        organism = synthetic construct
SEQUENCE: 124
gagtccgagc agaagaagat                                              20

SEQ ID NO: 125          moltype = RNA  length = 20
FEATURE                 Location/Qualifiers
misc_feature            1..20
                        note = Description of Artificial Sequence: Synthetic
                         oligonucleotide
source                  1..20
                        mol_type = other RNA
                        organism = synthetic construct
SEQUENCE: 125
gagtccgagc agaagaagta                                              20

SEQ ID NO: 126          moltype = RNA  length = 20
FEATURE                 Location/Qualifiers
misc_feature            1..20
                        note = Description of Artificial Sequence: Synthetic
                         oligonucleotide
source                  1..20
                        mol_type = other RNA
                        organism = synthetic construct
SEQUENCE: 126
gagtccgagc agaagaacaa                                              20
```

-continued

```
SEQ ID NO: 127          moltype = RNA  length = 20
FEATURE                 Location/Qualifiers
misc_feature            1..20
                        note = Description of Artificial Sequence: Synthetic
                         oligonucleotide
source                  1..20
                        mol_type = other RNA
                        organism = synthetic construct
SEQUENCE: 127
gagtccgagc agaagatgaa                                          20

SEQ ID NO: 128          moltype = RNA  length = 20
FEATURE                 Location/Qualifiers
misc_feature            1..20
                        note = Description of Artificial Sequence: Synthetic
                         oligonucleotide
source                  1..20
                        mol_type = other RNA
                        organism = synthetic construct
SEQUENCE: 128
gagtccgagc agaagtagaa                                          20

SEQ ID NO: 129          moltype = RNA  length = 20
FEATURE                 Location/Qualifiers
misc_feature            1..20
                        note = Description of Artificial Sequence: Synthetic
                         oligonucleotide
source                  1..20
                        mol_type = other RNA
                        organism = synthetic construct
SEQUENCE: 129
gagtccgagc agatgaagaa                                          20

SEQ ID NO: 130          moltype = RNA  length = 20
FEATURE                 Location/Qualifiers
misc_feature            1..20
                        note = Description of Artificial Sequence: Synthetic
                         oligonucleotide
source                  1..20
                        mol_type = other RNA
                        organism = synthetic construct
SEQUENCE: 130
gagtccgagc acaagaagaa                                          20

SEQ ID NO: 131          moltype = RNA  length = 20
FEATURE                 Location/Qualifiers
misc_feature            1..20
                        note = Description of Artificial Sequence: Synthetic
                         oligonucleotide
source                  1..20
                        mol_type = other RNA
                        organism = synthetic construct
SEQUENCE: 131
gagtccgagg agaagaagaa                                          20

SEQ ID NO: 132          moltype = RNA  length = 20
FEATURE                 Location/Qualifiers
misc_feature            1..20
                        note = Description of Artificial Sequence: Synthetic
                         oligonucleotide
source                  1..20
                        mol_type = other RNA
                        organism = synthetic construct
SEQUENCE: 132
gagtccgtgc agaagaagaa                                          20

SEQ ID NO: 133          moltype = RNA  length = 20
FEATURE                 Location/Qualifiers
misc_feature            1..20
                        note = Description of Artificial Sequence: Synthetic
                         oligonucleotide
source                  1..20
                        mol_type = other RNA
                        organism = synthetic construct
SEQUENCE: 133
gagtcggagc agaagaagaa                                          20
```

-continued

```
SEQ ID NO: 134          moltype = RNA  length = 20
FEATURE                 Location/Qualifiers
misc_feature            1..20
                        note = Description of Artificial Sequence: Synthetic
                         oligonucleotide
source                  1..20
                        mol_type = other RNA
                        organism = synthetic construct
SEQUENCE: 134
gagaccgagc agaagaagaa                                          20

SEQ ID NO: 135          moltype = DNA  length = 24
FEATURE                 Location/Qualifiers
misc_feature            1..24
                        note = Description of Artificial Sequence: Synthetic
                         oligonucleotide
source                  1..24
                        mol_type = other DNA
                        organism = synthetic construct
SEQUENCE: 135
aatgacaagc ttgctagcgg tggg                                     24

SEQ ID NO: 136          moltype = DNA  length = 39
FEATURE                 Location/Qualifiers
misc_feature            1..39
                        note = Description of Artificial Sequence: Synthetic
                         oligonucleotide
source                  1..39
                        mol_type = other DNA
                        organism = synthetic construct
SEQUENCE: 136
aaaacggaag ggcctgagtc cgagcagaag aagaagttt                      39

SEQ ID NO: 137          moltype = DNA  length = 39
FEATURE                 Location/Qualifiers
misc_feature            1..39
                        note = Description of Artificial Sequence: Synthetic
                         oligonucleotide
source                  1..39
                        mol_type = other DNA
                        organism = synthetic construct
SEQUENCE: 137
aaacaggggc cgagattggg tgttcagggc agaggtttt                      39

SEQ ID NO: 138          moltype = DNA  length = 38
FEATURE                 Location/Qualifiers
misc_feature            1..38
                        note = Description of Artificial Sequence: Synthetic
                         oligonucleotide
source                  1..38
                        mol_type = other DNA
                        organism = synthetic construct
SEQUENCE: 138
aaaacggaag ggcctgagtc cgagcagaag aagaagtt                       38

SEQ ID NO: 139          moltype = DNA  length = 40
FEATURE                 Location/Qualifiers
misc_feature            1..40
                        note = Description of Artificial Sequence: Synthetic
                         oligonucleotide
source                  1..40
                        mol_type = other DNA
                        organism = synthetic construct
SEQUENCE: 139
aacggaggga ggggcacaga tgagaaactc agggttttag                     40

SEQ ID NO: 140          moltype = DNA  length = 38
FEATURE                 Location/Qualifiers
source                  1..38
                        mol_type = unassigned DNA
                        organism = Homo sapiens
SEQUENCE: 140
agcccttctt cttctgctcg gactcaggcc cttcctcc                       38

SEQ ID NO: 141          moltype = DNA  length = 40
```

-continued

```
FEATURE                 Location/Qualifiers
source                  1..40
                        mol_type = unassigned DNA
                        organism = Homo sapiens
SEQUENCE: 141
cagggaggga ggggcacaga tgagaaactc aggaggcccc                          40

SEQ ID NO: 142          moltype = DNA  length = 80
FEATURE                 Location/Qualifiers
misc_feature            1..80
                        note = Description of Artificial Sequence: Synthetic
                         oligonucleotide
source                  1..80
                        mol_type = other DNA
                        organism = synthetic construct
SEQUENCE: 142
ggcaatgcgc caccggttga tgtgatggga gcccttctag gaggccccca gagcagccac  60
tggggcctca acactcaggc                                              80

SEQ ID NO: 143          moltype = DNA  length = 98
FEATURE                 Location/Qualifiers
misc_feature            1..98
                        note = Description of Artificial Sequence: Synthetic
                         oligonucleotide
source                  1..98
                        mol_type = other DNA
                        organism = synthetic construct
SEQUENCE: 143
ggacgaaaca ccggaaccat tcaaaacagc atagcaagtt aaaataaggc tagtccgtta  60
tcaacttgaa aaagtggcac cgagtcggtg cttttttt                          98

SEQ ID NO: 144          moltype = DNA  length = 186
FEATURE                 Location/Qualifiers
misc_feature            1..186
                        note = Description of Artificial Sequence: Synthetic
                         polynucleotide
source                  1..186
                        mol_type = other DNA
                        organism = synthetic construct
SEQUENCE: 144
ggacgaaaca ccggtagtat taagtattgt tttatggctg ataaatttct ttgaatttct  60
ccttgattat ttgttataaa agttataaaa taatcttgtt ggaaccattc aaaacagcat  120
agcaagttaa aataaggcta gtccgttatc aacttgaaaa agtggcaccg agtcggtgct  180
tttttt                                                             186

SEQ ID NO: 145          moltype = RNA  length = 46
FEATURE                 Location/Qualifiers
misc_feature            1..46
                        note = Description of Artificial Sequence: Synthetic
                         oligonucleotide
misc_difference         1..19
                        note = a, c, u, g, unknown or other
source                  1..46
                        mol_type = other RNA
                        organism = synthetic construct
SEQUENCE: 145
nnnnnnnnnn nnnnnnnnng ttattgtact ctcaagattt attttt                  46

SEQ ID NO: 146          moltype = RNA  length = 91
FEATURE                 Location/Qualifiers
misc_feature            1..91
                        note = Description of Artificial Sequence: Synthetic
                         oligonucleotide
source                  1..91
                        mol_type = other RNA
                        organism = synthetic construct
SEQUENCE: 146
gttacttaaa tcttgcagaa gctacaaaga taaggcttca tgccgaaatc aacaccctgt  60
cattttatgg cagggtgttt tcgttattta a                                 91

SEQ ID NO: 147          moltype = DNA  length = 70
FEATURE                 Location/Qualifiers
source                  1..70
                        mol_type = unassigned DNA
                        organism = Homo sapiens
SEQUENCE: 147
```

-continued

```
ttttctagtg ctgagtttct gtgactcctc tacattctac ttctctgtgt ttctgtatac  60
tacctcctcc                                                        70

SEQ ID NO: 148         moltype = DNA  length = 122
FEATURE                Location/Qualifiers
source                 1..122
                       mol_type = unassigned DNA
                       organism = Homo sapiens
SEQUENCE: 148
ggaggaaggg cctgagtccg agcagaagaa gaagggctcc catcacatca accggtggcg  60
cattgccacg aagcaggcca atggggagga catcgatgtc acctccaatg actagggtgg  120
gc                                                                122

SEQ ID NO: 149         moltype = RNA  length = 48
FEATURE                Location/Qualifiers
misc_feature           1..48
                       note = Description of Artificial Sequence: Synthetic
                        oligonucleotide
misc_difference        3..32
                       note = a, c, u, g, unknown or other
source                 1..48
                       mol_type = other RNA
                       organism = synthetic construct
SEQUENCE: 149
acnnnnnnnn nnnnnnnnnn nnnnnnnnnn nngttttaga gctatgct               48

SEQ ID NO: 150         moltype = RNA  length = 67
FEATURE                Location/Qualifiers
misc_feature           1..67
                       note = Description of Artificial Sequence: Synthetic
                        oligonucleotide
source                 1..67
                       mol_type = other RNA
                       organism = synthetic construct
modified_base          24
                       mod_base = OTHER
                       note = Thymine
SEQUENCE: 150
agcatagcaa gttaaaataa ggctagtccg ttatcaactt gaaaaagtgg caccgagtcg  60
gtgcttt                                                           67

SEQ ID NO: 151         moltype = RNA  length = 62
FEATURE                Location/Qualifiers
misc_feature           1..62
                       note = Description of Artificial Sequence: Synthetic
                        oligonucleotide
misc_difference        1..20
                       note = a, c, u, g, unknown or other
source                 1..62
                       mol_type = other RNA
                       organism = synthetic construct
SEQUENCE: 151
nnnnnnnnnn nnnnnnnnnn gttttagagc tagaaatagc aagttaaaat aaggctagtc  60
cg                                                                62

SEQ ID NO: 152         moltype = DNA  length = 73
FEATURE                Location/Qualifiers
misc_feature           1..73
                       note = Description of Artificial Sequence: Synthetic
                        oligonucleotide
source                 1..73
                       mol_type = other DNA
                       organism = synthetic construct
SEQUENCE: 152
tgaatggtcc caaaacggaa gggcctgagt ccgagcagaa gaagaagttt tagagctatg  60
ctgttttgaa tgg                                                    73

SEQ ID NO: 153         moltype = RNA  length = 99
FEATURE                Location/Qualifiers
misc_feature           1..99
                       note = Description of Artificial Sequence: Synthetic
                        oligonucleotide
misc_difference        1..20
                       note = a, c, u, g, unknown or other
source                 1..99
                       mol_type = other RNA
```

-continued

```
                     organism = synthetic construct
SEQUENCE: 153
nnnnnnnnnn nnnnnnnnnn gttttagagc tagaaatagc aagttaaaat aaggctagtc  60
cgttatcaac ttgaaaaagt ggcaccgagt cggtgcttt                        99

SEQ ID NO: 154         moltype = RNA  length = 127
FEATURE                Location/Qualifiers
misc_feature           1..127
                       note = Description of Artificial Sequence: Synthetic
                        polynucleotide
source                 1..127
                       mol_type = other RNA
                       organism = synthetic construct
SEQUENCE: 154
gtttttgtac tctcaagatt taagtaactg tacaacgtta cttaaatctt gcagaagcta  60
caaagataag gcttcatgcc gaaatcaaca ccctgtcatt ttatggcagg gtgttttcgt  120
tatttaa                                                           127

SEQ ID NO: 155         moltype = DNA  length = 56
FEATURE                Location/Qualifiers
misc_feature           1..56
                       note = Description of Artificial Sequence: Synthetic
                        oligonucleotide
misc_difference        1..20
                       note = a, c, t, g, unknown or other
source                 1..56
                       mol_type = other DNA
                       organism = synthetic construct
SEQUENCE: 155
nnnnnnnnnn nnnnnnnnnn gtttttgtac tctcaagatt taagtaactg tacaac      56

SEQ ID NO: 156         moltype = DNA  length = 91
FEATURE                Location/Qualifiers
misc_feature           1..91
                       note = Description of Artificial Sequence: Synthetic
                        oligonucleotide
source                 1..91
                       mol_type = other DNA
                       organism = synthetic construct
SEQUENCE: 156
gttacttaaa tcttgcagaa gctacaaaga taaggcttca tgccgaaatc aacaccctgt  60
cattttatgg cagggtgttt tcgttattta a                                91

SEQ ID NO: 157         moltype = DNA  length = 134
FEATURE                Location/Qualifiers
misc_feature           1..134
                       note = Description of Artificial Sequence: Synthetic
                        polynucleotide
misc_difference        1..20
                       note = a, c, t, g, unknown or other
source                 1..134
                       mol_type = other DNA
                       organism = synthetic construct
SEQUENCE: 157
nnnnnnnnnn nnnnnnnnnn gtttttgtac tctcaagatt taaggaaact aaatcttgca  60
gaagctacaa agataaggct tcatgccgaa atcaacaccc tgtcatttta tggcagggtg  120
ttttcgttat ttaa                                                   134

SEQ ID NO: 158         moltype = DNA  length = 131
FEATURE                Location/Qualifiers
misc_feature           1..131
                       note = Description of Artificial Sequence: Synthetic
                        polynucleotide
misc_difference        1..20
                       note = a, c, t, g, unknown or other
source                 1..131
                       mol_type = other DNA
                       organism = synthetic construct
SEQUENCE: 158
nnnnnnnnnn nnnnnnnnnn gtttttgtac tctcaagatt tagaaataaa tcttgcagaa  60
gctacaaaga taaggcttca tgccgaaatc aacaccctgt cattttatgg cagggtgttt  120
tcgttattta a                                                      131

SEQ ID NO: 159         moltype = DNA  length = 125
FEATURE                Location/Qualifiers
misc_feature           1..125
```

-continued

```
                     note = Description of Artificial Sequence: Synthetic
                      polynucleotide
misc_difference      1..20
                     note = a, c, t, g, unknown or other
source               1..125
                     mol_type = other DNA
                     organism = synthetic construct
SEQUENCE: 159
nnnnnnnnnn nnnnnnnnnn gtttttgtac tctcaagatg aaaatcttgc agaagctaca  60
aagataaggc ttcatgccga aatcaacacc ctgtcatttt atggcagggt gttttcgtta  120
tttaa                                                              125

SEQ ID NO: 160       moltype = DNA  length = 112
FEATURE              Location/Qualifiers
misc_feature         1..112
                     note = Description of Artificial Sequence: Synthetic
                      polynucleotide
misc_difference      1..20
                     note = a, c, t, g, unknown or other
source               1..112
                     mol_type = other DNA
                     organism = synthetic construct
SEQUENCE: 160
nnnnnnnnnn nnnnnnnnnn gtttttgtac tctgaaaaga agctacaaag ataaggcttc  60
atgccgaaat caacaccctg tcattttatg gcagggtgtt ttcgttattt aa          112

SEQ ID NO: 161       moltype = DNA  length = 107
FEATURE              Location/Qualifiers
misc_feature         1..107
                     note = Description of Artificial Sequence: Synthetic
                      polynucleotide
misc_difference      1..20
                     note = a, c, t, g, unknown or other
source               1..107
                     mol_type = other DNA
                     organism = synthetic construct
SEQUENCE: 161
nnnnnnnnnn nnnnnnnnnn gtttttgtac tgaaaagcta caaagataag gcttcatgcc  60
gaaatcaaca ccctgtcatt ttatggcagg gtgttttcgt tatttaa                107

SEQ ID NO: 162       moltype = DNA  length = 108
FEATURE              Location/Qualifiers
misc_feature         1..108
                     note = Description of Artificial Sequence: Synthetic
                      polynucleotide
misc_difference      1..20
                     note = a, c, t, g, unknown or other
source               1..108
                     mol_type = other DNA
                     organism = synthetic construct
SEQUENCE: 162
nnnnnnnnnn nnnnnnnnnn gtttttgtac tctcaagatt tagaaataaa tcttgcagaa  60
gctacaaaga taaggcttca tgccgaaatc aacaccctgt cattttat               108

SEQ ID NO: 163       moltype = DNA  length = 86
FEATURE              Location/Qualifiers
misc_feature         1..86
                     note = Description of Artificial Sequence: Synthetic
                      oligonucleotide
misc_difference      1..20
                     note = a, c, t, g, unknown or other
source               1..86
                     mol_type = other DNA
                     organism = synthetic construct
SEQUENCE: 163
nnnnnnnnnn nnnnnnnnnn gtttttgtac tctcaagatt tagaaataaa tcttgcagaa  60
gctacaaaga taaggcttca tgccga                                       86

SEQ ID NO: 164       moltype = DNA  length = 79
FEATURE              Location/Qualifiers
misc_feature         1..79
                     note = Description of Artificial Sequence: Synthetic
                      oligonucleotide
misc_difference      1..20
                     note = a, c, t, g, unknown or other
source               1..79
```

-continued

```
                     mol_type = other DNA
                     organism = synthetic construct
SEQUENCE: 164
nnnnnnnnnn nnnnnnnnnn gtttttgtac tctcaagatt tagaaataaa tcttgcagaa  60
gctacaaaga taaggcttc                                              79

SEQ ID NO: 165       moltype = DNA  length = 73
FEATURE              Location/Qualifiers
misc_feature         1..73
                     note = Description of Artificial Sequence: Synthetic
                      oligonucleotide
misc_difference      1..20
                     note = a, c, t, g, unknown or other
source               1..73
                     mol_type = other DNA
                     organism = synthetic construct
SEQUENCE: 165
nnnnnnnnnn nnnnnnnnnn gtttttgtac tctcaagatt tagaaataaa tcttgcagaa  60
gctacaaaga taa                                                    73

SEQ ID NO: 166       moltype = RNA  length = 125
FEATURE              Location/Qualifiers
misc_feature         1..125
                     note = Description of Artificial Sequence: Synthetic
                      polynucleotide
source               1..125
                     mol_type = other RNA
                     organism = synthetic construct
SEQUENCE: 166
gttttagtcc ctttttaaat ttctttatgg taaaattata atctcataag aaatttaaaa  60
agggactaaa ataaagagtt tgcgggactc tgcggggtta caatccccta aaaccgcttt  120
taaaa                                                             125

SEQ ID NO: 167       moltype = RNA  length = 91
FEATURE              Location/Qualifiers
misc_feature         1..91
                     note = Description of Artificial Sequence: Synthetic
                      oligonucleotide
source               1..91
                     mol_type = other RNA
                     organism = synthetic construct
SEQUENCE: 167
gttacttaaa tcttgcagaa gctacaaaga taaggcttca tgccgaaatc aacaccctgt  60
cattttatgg cagggtgttt tcgttattta a                                91

SEQ ID NO: 168       moltype = RNA  length = 56
FEATURE              Location/Qualifiers
misc_feature         1..56
                     note = Description of Artificial Sequence: Synthetic
                      oligonucleotide
source               1..56
                     mol_type = other RNA
                     organism = synthetic construct
SEQUENCE: 168
gggactcaac caagtcattc gtttttgtac tctcaagatt taagtaactg tacaac      56

SEQ ID NO: 169       moltype = RNA  length = 147
FEATURE              Location/Qualifiers
misc_feature         1..147
                     note = Description of Artificial Sequence: Synthetic
                      polynucleotide
source               1..147
                     mol_type = other RNA
                     organism = synthetic construct
SEQUENCE: 169
gggactcaac caagtcattc gtttttgtac tctcaagatt taagtaactg tacaacgtta  60
cttaaatctt gcagaagcta caaagataag gcttcatgcc gaaatcaaca ccctgtcatt  120
ttatggcagg gtgttttcgt tatttaa                                     147

SEQ ID NO: 170       moltype = RNA  length = 70
FEATURE              Location/Qualifiers
misc_feature         1..70
                     note = Description of Artificial Sequence: Synthetic
                      oligonucleotide
source               1..70
                     mol_type = other RNA
```

-continued

```
                     organism = synthetic construct
SEQUENCE: 170
cttgcagaag ctacaaagat aaggcttcat gccgaaatca acaccctgtc attttatggc  60
agggtgtttt                                                         70

SEQ ID NO: 171         moltype = RNA  length = 42
FEATURE                Location/Qualifiers
misc_feature           1..42
                       note = Description of Artificial Sequence: Synthetic
                        oligonucleotide
source                 1..42
                       mol_type = other RNA
                       organism = synthetic construct
SEQUENCE: 171
gggactcaac caagtcattc gtttttgtac tctcaagatt ta                      42

SEQ ID NO: 172         moltype = RNA  length = 112
FEATURE                Location/Qualifiers
misc_feature           1..112
                       note = Description of Artificial Sequence: Synthetic
                        polynucleotide
source                 1..112
                       mol_type = other RNA
                       organism = synthetic construct
SEQUENCE: 172
gggactcaac caagtcattc gtttttgtac tctcaagatt tacttgcaga agctacaaag  60
ataaggcttc atgccgaaat caacaccctg tcattttatg gcagggtgtt tt          112

SEQ ID NO: 173         moltype = RNA  length = 116
FEATURE                Location/Qualifiers
misc_feature           1..116
                       note = Description of Artificial Sequence: Synthetic
                        polynucleotide
source                 1..116
                       mol_type = other RNA
                       organism = synthetic construct
SEQUENCE: 173
gggactcaac caagtcattc gtttttgtac tctcaagatt tagaaacttg cagaagctac  60
aaagataagg cttcatgccg aaatcaacac cctgtcattt tatggcaggg tgtttt      116

SEQ ID NO: 174         moltype = RNA  length = 116
FEATURE                Location/Qualifiers
misc_feature           1..116
                       note = Description of Artificial Sequence: Synthetic
                        polynucleotide
source                 1..116
                       mol_type = other RNA
                       organism = synthetic construct
SEQUENCE: 174
gggactcaac caagtcattc gtttttgtac tctcaagatt tagaaacttg cagaagctac  60
aaagataagg cttcatgccg aaatcaacac cctgtcattt tatggcaggg tgtttt      116

SEQ ID NO: 175         moltype = RNA  length = 102
FEATURE                Location/Qualifiers
misc_feature           1..102
                       note = Description of Artificial Sequence: Synthetic
                        polynucleotide
source                 1..102
                       mol_type = other RNA
                       organism = synthetic construct
SEQUENCE: 175
gggactcaac caagtcattc gtttttgtag aaatacaaag ataaggcttc atgccgaaat  60
caacaccctg tcattttatg gcagggtgtt ttcgttattt aa                     102

SEQ ID NO: 176         moltype = RNA  length = 102
FEATURE                Location/Qualifiers
misc_feature           1..102
                       note = Description of Artificial Sequence: Synthetic
                        polynucleotide
source                 1..102
                       mol_type = other RNA
                       organism = synthetic construct
SEQUENCE: 176
gggactcaac caagtcattc gtttttgtag aaatacaaag ataaggcttc atgccgaaat  60
caacaccctg tcattttatg gcagggtgtt ttcgttattt aa                     102
```

-continued

```
SEQ ID NO: 177         moltype = RNA  length = 57
FEATURE                Location/Qualifiers
misc_feature           1..57
                       note = Description of Artificial Sequence: Synthetic
                        oligonucleotide
source                 1..57
                       mol_type = other RNA
                       organism = synthetic construct
SEQUENCE: 177
gggactcaac caagtcattc gtttttgtag aaatacaaag ataaggcttc atgccga     57

SEQ ID NO: 178         moltype = RNA  length = 57
FEATURE                Location/Qualifiers
misc_feature           1..57
                       note = Description of Artificial Sequence: Synthetic
                        oligonucleotide
source                 1..57
                       mol_type = other RNA
                       organism = synthetic construct
SEQUENCE: 178
gggactcaac caagtcattc gtttttgtag aaatacaaag ataaggcttc atgccga     57

SEQ ID NO: 179         moltype = DNA  length = 23
FEATURE                Location/Qualifiers
misc_feature           1..23
                       note = Description of Artificial Sequence: Synthetic
                        oligonucleotide
source                 1..23
                       mol_type = other DNA
                       organism = synthetic construct
SEQUENCE: 179
gtggtgtcac gctcgtcgtt tgg                                          23

SEQ ID NO: 180         moltype = DNA  length = 23
FEATURE                Location/Qualifiers
misc_feature           1..23
                       note = Description of Artificial Sequence: Synthetic
                        oligonucleotide
source                 1..23
                       mol_type = other DNA
                       organism = synthetic construct
SEQUENCE: 180
tccagtctat taattgttgc cgg                                          23

SEQ ID NO: 181         moltype = DNA  length = 64
FEATURE                Location/Qualifiers
source                 1..64
                       mol_type = unassigned DNA
                       organism = Homo sapiens
SEQUENCE: 181
caagaggctt gagtaggaga ggagtgccgc cgaggcgggg cggggcgggg cgtggagctg  60
ggct                                                               64

SEQ ID NO: 182         moltype = RNA  length = 99
FEATURE                Location/Qualifiers
misc_feature           1..99
                       note = Description of Artificial Sequence: Synthetic
                        oligonucleotide
misc_difference        1..20
                       note = a, c, u, g, unknown or other
source                 1..99
                       mol_type = other RNA
                       organism = synthetic construct
SEQUENCE: 182
nnnnnnnnnn nnnnnnnnnn gtattagagc tagaaatagc aagttaatat aaggctagtc  60
cgttatcaac ttgaaaaagt ggcaccgagt cggtgcttt                         99

SEQ ID NO: 183         moltype = RNA  length = 119
FEATURE                Location/Qualifiers
misc_feature           1..119
                       note = Description of Artificial Sequence: Synthetic
                        polynucleotide
misc_difference        1..20
                       note = a, c, u, g, unknown or other
source                 1..119
                       mol_type = other RNA
```

-continued

```
                     organism = synthetic construct
SEQUENCE: 183
nnnnnnnnnn nnnnnnnnnn gttttagagc tatgctgttt tggaaacaaa acagcatagc  60
aagttaaaat aaggctagtc cgttatcaac ttgaaaaagt ggcaccgagt cggtgcttt   119

SEQ ID NO: 184         moltype = RNA  length = 119
FEATURE                Location/Qualifiers
misc_feature           1..119
                       note = Description of Artificial Sequence: Synthetic
                        polynucleotide
misc_difference        1..20
                       note = a, c, u, g, unknown or other
source                 1..119
                       mol_type = other RNA
                       organism = synthetic construct
SEQUENCE: 184
nnnnnnnnnn nnnnnnnnnn gtattagagc tatgctgtat tggaaacaat acagcatagc  60
aagttaatat aaggctagtc cgttatcaac ttgaaaaagt ggcaccgagt cggtgcttt   119

SEQ ID NO: 185         moltype = DNA  length = 12
FEATURE                Location/Qualifiers
source                 1..12
                       mol_type = unassigned DNA
                       organism = Homo sapiens
SEQUENCE: 185
tagcgggtaa gc                                                     12

SEQ ID NO: 186         moltype = DNA  length = 12
FEATURE                Location/Qualifiers
source                 1..12
                       mol_type = unassigned DNA
                       organism = Homo sapiens
SEQUENCE: 186
tcggtgacat gt                                                     12

SEQ ID NO: 187         moltype = DNA  length = 12
FEATURE                Location/Qualifiers
source                 1..12
                       mol_type = unassigned DNA
                       organism = Homo sapiens
SEQUENCE: 187
actccccgta gg                                                     12

SEQ ID NO: 188         moltype = DNA  length = 12
FEATURE                Location/Qualifiers
source                 1..12
                       mol_type = unassigned DNA
                       organism = Homo sapiens
SEQUENCE: 188
actgcgtgtt aa                                                     12

SEQ ID NO: 189         moltype = DNA  length = 12
FEATURE                Location/Qualifiers
source                 1..12
                       mol_type = unassigned DNA
                       organism = Homo sapiens
SEQUENCE: 189
acgtcgcctg at                                                     12

SEQ ID NO: 190         moltype = DNA  length = 12
FEATURE                Location/Qualifiers
source                 1..12
                       mol_type = unassigned DNA
                       organism = Homo sapiens
SEQUENCE: 190
taggtcgacc ag                                                     12

SEQ ID NO: 191         moltype = DNA  length = 12
FEATURE                Location/Qualifiers
source                 1..12
                       mol_type = unassigned DNA
                       organism = Homo sapiens
SEQUENCE: 191
ggcgttaatg at                                                     12

SEQ ID NO: 192         moltype = DNA  length = 12
```

-continued

```
FEATURE             Location/Qualifiers
source              1..12
                    mol_type = unassigned DNA
                    organism = Homo sapiens
SEQUENCE: 192
tgtcgcatgt ta                                                    12

SEQ ID NO: 193      moltype = DNA  length = 12
FEATURE             Location/Qualifiers
source              1..12
                    mol_type = unassigned DNA
                    organism = Homo sapiens
SEQUENCE: 193
atggaaacgc at                                                    12

SEQ ID NO: 194      moltype = DNA  length = 12
FEATURE             Location/Qualifiers
source              1..12
                    mol_type = unassigned DNA
                    organism = Homo sapiens
SEQUENCE: 194
gccgaattcc tc                                                    12

SEQ ID NO: 195      moltype = DNA  length = 12
FEATURE             Location/Qualifiers
source              1..12
                    mol_type = unassigned DNA
                    organism = Homo sapiens
SEQUENCE: 195
gcatggtacg ga                                                    12

SEQ ID NO: 196      moltype = DNA  length = 12
FEATURE             Location/Qualifiers
source              1..12
                    mol_type = unassigned DNA
                    organism = Homo sapiens
SEQUENCE: 196
cggtactctt ac                                                    12

SEQ ID NO: 197      moltype = DNA  length = 12
FEATURE             Location/Qualifiers
source              1..12
                    mol_type = unassigned DNA
                    organism = Homo sapiens
SEQUENCE: 197
gcctgtgccg ta                                                    12

SEQ ID NO: 198      moltype = DNA  length = 12
FEATURE             Location/Qualifiers
source              1..12
                    mol_type = unassigned DNA
                    organism = Homo sapiens
SEQUENCE: 198
tacggtaagt cg                                                    12

SEQ ID NO: 199      moltype = DNA  length = 12
FEATURE             Location/Qualifiers
source              1..12
                    mol_type = unassigned DNA
                    organism = Homo sapiens
SEQUENCE: 199
cacgaaatta cc                                                    12

SEQ ID NO: 200      moltype = DNA  length = 12
FEATURE             Location/Qualifiers
source              1..12
                    mol_type = unassigned DNA
                    organism = Homo sapiens
SEQUENCE: 200
aaccaagata cg                                                    12

SEQ ID NO: 201      moltype = DNA  length = 12
FEATURE             Location/Qualifiers
source              1..12
                    mol_type = unassigned DNA
                    organism = Homo sapiens
```

-continued

```
SEQUENCE: 201
gagtcgatac gc                                                    12

SEQ ID NO: 202          moltype = DNA  length = 12
FEATURE                 Location/Qualifiers
source                  1..12
                        mol_type = unassigned DNA
                        organism = Homo sapiens
SEQUENCE: 202
gtctcacgat cg                                                    12

SEQ ID NO: 203          moltype = DNA  length = 12
FEATURE                 Location/Qualifiers
source                  1..12
                        mol_type = unassigned DNA
                        organism = Homo sapiens
SEQUENCE: 203
tcgtcgggtg ca                                                    12

SEQ ID NO: 204          moltype = DNA  length = 12
FEATURE                 Location/Qualifiers
source                  1..12
                        mol_type = unassigned DNA
                        organism = Homo sapiens
SEQUENCE: 204
actccgtagt ga                                                    12

SEQ ID NO: 205          moltype = DNA  length = 12
FEATURE                 Location/Qualifiers
source                  1..12
                        mol_type = unassigned DNA
                        organism = Homo sapiens
SEQUENCE: 205
caggacgtcc gt                                                    12

SEQ ID NO: 206          moltype = DNA  length = 12
FEATURE                 Location/Qualifiers
source                  1..12
                        mol_type = unassigned DNA
                        organism = Homo sapiens
SEQUENCE: 206
tcgtatccct ac                                                    12

SEQ ID NO: 207          moltype = DNA  length = 12
FEATURE                 Location/Qualifiers
source                  1..12
                        mol_type = unassigned DNA
                        organism = Homo sapiens
SEQUENCE: 207
tttcaaggcc gg                                                    12

SEQ ID NO: 208          moltype = DNA  length = 12
FEATURE                 Location/Qualifiers
source                  1..12
                        mol_type = unassigned DNA
                        organism = Homo sapiens
SEQUENCE: 208
cgccggtgga at                                                    12

SEQ ID NO: 209          moltype = DNA  length = 12
FEATURE                 Location/Qualifiers
source                  1..12
                        mol_type = unassigned DNA
                        organism = Homo sapiens
SEQUENCE: 209
gaacccgtcc ta                                                    12

SEQ ID NO: 210          moltype = DNA  length = 12
FEATURE                 Location/Qualifiers
source                  1..12
                        mol_type = unassigned DNA
                        organism = Homo sapiens
SEQUENCE: 210
gattcatcag cg                                                    12

SEQ ID NO: 211          moltype = DNA  length = 12
```

-continued

```
FEATURE             Location/Qualifiers
source              1..12
                    mol_type = unassigned DNA
                    organism = Homo sapiens
SEQUENCE: 211
acaccggtct tc                                                    12

SEQ ID NO: 212      moltype = DNA  length = 12
FEATURE             Location/Qualifiers
source              1..12
                    mol_type = unassigned DNA
                    organism = Homo sapiens
SEQUENCE: 212
atcgtgccct aa                                                    12

SEQ ID NO: 213      moltype = DNA  length = 12
FEATURE             Location/Qualifiers
source              1..12
                    mol_type = unassigned DNA
                    organism = Homo sapiens
SEQUENCE: 213
gcgtcaatgt tc                                                    12

SEQ ID NO: 214      moltype = DNA  length = 12
FEATURE             Location/Qualifiers
source              1..12
                    mol_type = unassigned DNA
                    organism = Homo sapiens
SEQUENCE: 214
ctccgtatct cg                                                    12

SEQ ID NO: 215      moltype = DNA  length = 12
FEATURE             Location/Qualifiers
source              1..12
                    mol_type = unassigned DNA
                    organism = Homo sapiens
SEQUENCE: 215
ccgattcctt cg                                                    12

SEQ ID NO: 216      moltype = DNA  length = 12
FEATURE             Location/Qualifiers
source              1..12
                    mol_type = unassigned DNA
                    organism = Homo sapiens
SEQUENCE: 216
tgcgcctcca gt                                                    12

SEQ ID NO: 217      moltype = DNA  length = 12
FEATURE             Location/Qualifiers
source              1..12
                    mol_type = unassigned DNA
                    organism = Homo sapiens
SEQUENCE: 217
taacgtcgga gc                                                    12

SEQ ID NO: 218      moltype = DNA  length = 12
FEATURE             Location/Qualifiers
source              1..12
                    mol_type = unassigned DNA
                    organism = Homo sapiens
SEQUENCE: 218
aaggtcgccc at                                                    12

SEQ ID NO: 219      moltype = DNA  length = 12
FEATURE             Location/Qualifiers
source              1..12
                    mol_type = unassigned DNA
                    organism = Homo sapiens
SEQUENCE: 219
gtcggggact at                                                    12

SEQ ID NO: 220      moltype = DNA  length = 12
FEATURE             Location/Qualifiers
source              1..12
                    mol_type = unassigned DNA
                    organism = Homo sapiens
```

-continued

SEQUENCE: 220
ttcgagcgat tt 12

SEQ ID NO: 221 moltype = DNA length = 12
FEATURE Location/Qualifiers
source 1..12
mol_type = unassigned DNA
organism = Homo sapiens
SEQUENCE: 221
tgagtcgtcg ag 12

SEQ ID NO: 222 moltype = DNA length = 12
FEATURE Location/Qualifiers
source 1..12
mol_type = unassigned DNA
organism = Homo sapiens
SEQUENCE: 222
tttacgcaga gg 12

SEQ ID NO: 223 moltype = DNA length = 12
FEATURE Location/Qualifiers
source 1..12
mol_type = unassigned DNA
organism = Homo sapiens
SEQUENCE: 223
aggaagtatc gc 12

SEQ ID NO: 224 moltype = DNA length = 12
FEATURE Location/Qualifiers
source 1..12
mol_type = unassigned DNA
organism = Homo sapiens
SEQUENCE: 224
actcgatacc at 12

SEQ ID NO: 225 moltype = DNA length = 12
FEATURE Location/Qualifiers
source 1..12
mol_type = unassigned DNA
organism = Homo sapiens
SEQUENCE: 225
cgctacatag ca 12

SEQ ID NO: 226 moltype = DNA length = 12
FEATURE Location/Qualifiers
source 1..12
mol_type = unassigned DNA
organism = Homo sapiens
SEQUENCE: 226
ttcataaccg gc 12

SEQ ID NO: 227 moltype = DNA length = 12
FEATURE Location/Qualifiers
source 1..12
mol_type = unassigned DNA
organism = Homo sapiens
SEQUENCE: 227
ccaaacggtt aa 12

SEQ ID NO: 228 moltype = DNA length = 12
FEATURE Location/Qualifiers
source 1..12
mol_type = unassigned DNA
organism = Homo sapiens
SEQUENCE: 228
cgattccttc gt 12

SEQ ID NO: 229 moltype = DNA length = 12
FEATURE Location/Qualifiers
source 1..12
mol_type = unassigned DNA
organism = Homo sapiens
SEQUENCE: 229
cgtcatgaat aa 12

SEQ ID NO: 230 moltype = DNA length = 12

-continued

```
FEATURE             Location/Qualifiers
source              1..12
                    mol_type = unassigned DNA
                    organism = Homo sapiens
SEQUENCE: 230
agtggcgatg ac                                                    12

SEQ ID NO: 231      moltype = DNA  length = 12
FEATURE             Location/Qualifiers
source              1..12
                    mol_type = unassigned DNA
                    organism = Homo sapiens
SEQUENCE: 231
cccctacggc ac                                                    12

SEQ ID NO: 232      moltype = DNA  length = 12
FEATURE             Location/Qualifiers
source              1..12
                    mol_type = unassigned DNA
                    organism = Homo sapiens
SEQUENCE: 232
gccaacccgc ac                                                    12

SEQ ID NO: 233      moltype = DNA  length = 12
FEATURE             Location/Qualifiers
source              1..12
                    mol_type = unassigned DNA
                    organism = Homo sapiens
SEQUENCE: 233
tgggacaccg gt                                                    12

SEQ ID NO: 234      moltype = DNA  length = 12
FEATURE             Location/Qualifiers
source              1..12
                    mol_type = unassigned DNA
                    organism = Homo sapiens
SEQUENCE: 234
ttgactgcgg cg                                                    12

SEQ ID NO: 235      moltype = DNA  length = 12
FEATURE             Location/Qualifiers
source              1..12
                    mol_type = unassigned DNA
                    organism = Homo sapiens
SEQUENCE: 235
actatgcgta gg                                                    12

SEQ ID NO: 236      moltype = DNA  length = 12
FEATURE             Location/Qualifiers
source              1..12
                    mol_type = unassigned DNA
                    organism = Homo sapiens
SEQUENCE: 236
tcacccaaag cg                                                    12

SEQ ID NO: 237      moltype = DNA  length = 12
FEATURE             Location/Qualifiers
source              1..12
                    mol_type = unassigned DNA
                    organism = Homo sapiens
SEQUENCE: 237
gcaggacgtc cg                                                    12

SEQ ID NO: 238      moltype = DNA  length = 12
FEATURE             Location/Qualifiers
source              1..12
                    mol_type = unassigned DNA
                    organism = Homo sapiens
SEQUENCE: 238
acaccgaaaa cg                                                    12

SEQ ID NO: 239      moltype = DNA  length = 12
FEATURE             Location/Qualifiers
source              1..12
                    mol_type = unassigned DNA
                    organism = Homo sapiens
```

-continued

SEQUENCE: 239
cggtgtattg ag 12

SEQ ID NO: 240 moltype = DNA length = 12
FEATURE Location/Qualifiers
source 1..12
mol_type = unassigned DNA
organism = Homo sapiens
SEQUENCE: 240
cacgaggtat gc 12

SEQ ID NO: 241 moltype = DNA length = 12
FEATURE Location/Qualifiers
source 1..12
mol_type = unassigned DNA
organism = Homo sapiens
SEQUENCE: 241
taaagcgacc cg 12

SEQ ID NO: 242 moltype = DNA length = 12
FEATURE Location/Qualifiers
source 1..12
mol_type = unassigned DNA
organism = Homo sapiens
SEQUENCE: 242
cttagtcggc ca 12

SEQ ID NO: 243 moltype = DNA length = 12
FEATURE Location/Qualifiers
source 1..12
mol_type = unassigned DNA
organism = Homo sapiens
SEQUENCE: 243
cgaaaacgtg gc 12

SEQ ID NO: 244 moltype = DNA length = 12
FEATURE Location/Qualifiers
source 1..12
mol_type = unassigned DNA
organism = Homo sapiens
SEQUENCE: 244
cgtgccctga ac 12

SEQ ID NO: 245 moltype = DNA length = 12
FEATURE Location/Qualifiers
source 1..12
mol_type = unassigned DNA
organism = Homo sapiens
SEQUENCE: 245
tttaccatcg aa 12

SEQ ID NO: 246 moltype = DNA length = 12
FEATURE Location/Qualifiers
source 1..12
mol_type = unassigned DNA
organism = Homo sapiens
SEQUENCE: 246
cgtagccatg tt 12

SEQ ID NO: 247 moltype = DNA length = 12
FEATURE Location/Qualifiers
source 1..12
mol_type = unassigned DNA
organism = Homo sapiens
SEQUENCE: 247
cccaaacggt ta 12

SEQ ID NO: 248 moltype = DNA length = 12
FEATURE Location/Qualifiers
source 1..12
mol_type = unassigned DNA
organism = Homo sapiens
SEQUENCE: 248
gcgttatcag aa 12

SEQ ID NO: 249 moltype = DNA length = 12

-continued

```
FEATURE             Location/Qualifiers
source              1..12
                    mol_type = unassigned DNA
                    organism = Homo sapiens
SEQUENCE: 249
tcgatggtaa ac                                                    12

SEQ ID NO: 250      moltype = DNA  length = 12
FEATURE             Location/Qualifiers
source              1..12
                    mol_type = unassigned DNA
                    organism = Homo sapiens
SEQUENCE: 250
cgacttttg ca                                                     12

SEQ ID NO: 251      moltype = DNA  length = 12
FEATURE             Location/Qualifiers
source              1..12
                    mol_type = unassigned DNA
                    organism = Homo sapiens
SEQUENCE: 251
tcgacgactc ac                                                    12

SEQ ID NO: 252      moltype = DNA  length = 12
FEATURE             Location/Qualifiers
source              1..12
                    mol_type = unassigned DNA
                    organism = Homo sapiens
SEQUENCE: 252
acgcgtcaga ta                                                    12

SEQ ID NO: 253      moltype = DNA  length = 12
FEATURE             Location/Qualifiers
source              1..12
                    mol_type = unassigned DNA
                    organism = Homo sapiens
SEQUENCE: 253
cgtacggcac ag                                                    12

SEQ ID NO: 254      moltype = DNA  length = 12
FEATURE             Location/Qualifiers
source              1..12
                    mol_type = unassigned DNA
                    organism = Homo sapiens
SEQUENCE: 254
ctatgccgtg ca                                                    12

SEQ ID NO: 255      moltype = DNA  length = 12
FEATURE             Location/Qualifiers
source              1..12
                    mol_type = unassigned DNA
                    organism = Homo sapiens
SEQUENCE: 255
cgcgtcagat at                                                    12

SEQ ID NO: 256      moltype = DNA  length = 12
FEATURE             Location/Qualifiers
source              1..12
                    mol_type = unassigned DNA
                    organism = Homo sapiens
SEQUENCE: 256
aagatcggta gc                                                    12

SEQ ID NO: 257      moltype = DNA  length = 12
FEATURE             Location/Qualifiers
source              1..12
                    mol_type = unassigned DNA
                    organism = Homo sapiens
SEQUENCE: 257
cttcgcaagg ag                                                    12

SEQ ID NO: 258      moltype = DNA  length = 12
FEATURE             Location/Qualifiers
source              1..12
                    mol_type = unassigned DNA
                    organism = Homo sapiens
```

-continued

```
SEQUENCE: 258
gtcgtggact ac                                                    12

SEQ ID NO: 259         moltype = DNA  length = 12
FEATURE                Location/Qualifiers
source                 1..12
                       mol_type = unassigned DNA
                       organism = Homo sapiens
SEQUENCE: 259
ggtcgtcatc aa                                                    12

SEQ ID NO: 260         moltype = DNA  length = 12
FEATURE                Location/Qualifiers
source                 1..12
                       mol_type = unassigned DNA
                       organism = Homo sapiens
SEQUENCE: 260
gttaacagcg tg                                                    12

SEQ ID NO: 261         moltype = DNA  length = 12
FEATURE                Location/Qualifiers
source                 1..12
                       mol_type = unassigned DNA
                       organism = Homo sapiens
SEQUENCE: 261
tagctaaccg tt                                                    12

SEQ ID NO: 262         moltype = DNA  length = 12
FEATURE                Location/Qualifiers
source                 1..12
                       mol_type = unassigned DNA
                       organism = Homo sapiens
SEQUENCE: 262
agtaaaggcg ct                                                    12

SEQ ID NO: 263         moltype = DNA  length = 12
FEATURE                Location/Qualifiers
source                 1..12
                       mol_type = unassigned DNA
                       organism = Homo sapiens
SEQUENCE: 263
ggtaatttcg tg                                                    12

SEQ ID NO: 264         moltype = RNA  length = 147
FEATURE                Location/Qualifiers
misc_feature           1..147
                       note = Description of Artificial Sequence: Synthetic
                        polynucleotide
misc_difference        1..20
                       note = a, c, u, g, unknown or other
source                 1..147
                       mol_type = other RNA
                       organism = synthetic construct
SEQUENCE: 264
nnnnnnnnnn nnnnnnnnnn gttttagtac tctgtaattt taggtatgag gtagacgaaa  60
attgtactta tacctaaaat tacagaatct actaaaacaa ggcaaaatgc cgtgtttatc  120
tcgtcaactt gttggcgaga ttttttt                                    147

SEQ ID NO: 265         moltype =   length =
SEQUENCE: 265
000

SEQ ID NO: 266         moltype =   length =
SEQUENCE: 266
000

SEQ ID NO: 267         moltype =   length =
SEQUENCE: 267
000

SEQ ID NO: 268         moltype =   length =
SEQUENCE: 268
000

SEQ ID NO: 269         moltype =   length =
SEQUENCE: 269
```

-continued

000

SEQ ID NO: 270 moltype = length =
SEQUENCE: 270
000

SEQ ID NO: 271 moltype = length =
SEQUENCE: 271
000

SEQ ID NO: 272 moltype = length =
SEQUENCE: 272
000

1. (canceled)

2. An engineered CRISPR-Cas system chimeric RNA comprising NNNNNNNNNNNNNNNNNNNNGUUUUAGAGCUAGAAAUAGCAAGUUAAAAUAAGG CUAGUCCGUUAUCA.

3. The engineered CRISPR-Cas system chimeric RNA of claim **2**, wherein NNNNNNNNNNNNNNNNNNNN is a guide sequence capable of hybridizing to a target sequence in a eukaryotic cell adjacent to a protospacer adjacent motif (PAM).

4. The engineered CRISPR-Cas system chimeric RNA of claim **3**, wherein the PAM is NGG.

5. The engineered CRISPR-Cas system chimeric RNA of claim **2**, further comprising a poly-U sequence.

6. The engineered CRISPR-Cas system chimeric RNA of claim **2**, wherein the RNA sequence is encoded by SEQ ID NO:26.

7. The engineered CRISPR-Cas system chimeric RNA of claim **2**, comprising one or more modified nucleotides.

8. The engineered CRISPR-Cas system chimeric RNA of claim **2**, comprising one or more methylated nucleotides or nucleotide analogs.

9. An engineered CRISPR-Cas system chimeric RNA comprising, from 5' to 3':

(a) a guide sequence capable of hybridizing to a target sequence in a eukaryotic cell adjacent to a protospacer adjacent motif (PAM),

(b) a tracr-mate sequence, and

(c) a tracr sequence comprising at least 40 nucleotides in length,

wherein the tracr-mate sequence is capable of hybridizing to the tracr sequence, and wherein the chimeric RNA is capable of forming a CRISPR complex with *S. pyogenes* Cas9 and directs sequence-specific binding of the CRISPR complex to the target sequence adjacent to the PAM in the eukaryotic cell.

10. The engineered CRISPR-Cas system chimeric RNA of claim **9**, wherein tracr sequence comprises at least 50 nucleotides in length.

11. The engineered CRISPR-Cas system chimeric RNA of claim **9**, wherein guide sequence comprises 15-25 nucleotides in length.

12. The engineered CRISPR-Cas system chimeric RNA of claim **9**, wherein the PAM is NGG.

13. The engineered CRISPR-Cas system chimeric RNA of claim **9**, further comprising a poly-U sequence.

14. The engineered CRISPR-Cas system chimeric RNA of claim **9**, comprising one or more modified nucleotides.

15. The engineered CRISPR-Cas system chimeric RNA of claim **9**, comprising one or more methylated nucleotides or nucleotide analogs.

* * * * *