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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
21 May 2015 (21.05.2015)

WIPO | PCT

(10) International Publication Number
WO 2015/071474 A2

(51) International Patent Classification:
C12N 15/11 (2006.01)

(21) International Application Number:
PCT/EP2014/074813

(22) International Filing Date:
17 November 2014 (17.11.2014)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/905,835 18 November 2013 (18.11.2013) US

(71) Applicant: **CRISPR THERAPEUTICS AG** [CH/CH]; Aeschenvorstadt 36, CH-4051 Basel (CH).

(72) Inventors: **CHARPENTIER, Emmanuelle**; Boecklerstrasse 18, 38102 Braunschweig (DE). **CHYLINSKI, Krzysztof**; Simmeringer Hauptstrasse 45/8, A-1110 Vienna (AT). **FONFARA, Ines**; Helmstedter Strasse 144, 38102 Braunschweig (DE).

(74) Agent: **PILKINGTON, Stephanie Joan**; Potter Clarkson LLP, The Belgrave Centre, Talbot Street, Nottingham NG1 5GG (GB).

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

[Continued on next page]

(54) Title: CRISPR-CAS SYSTEM MATERIALS AND METHODS

(57) Abstract: The invention relates to Type II CRISPR-Cas systems of Cas9 enzymes, guide RNAs and associated specific PAMs.

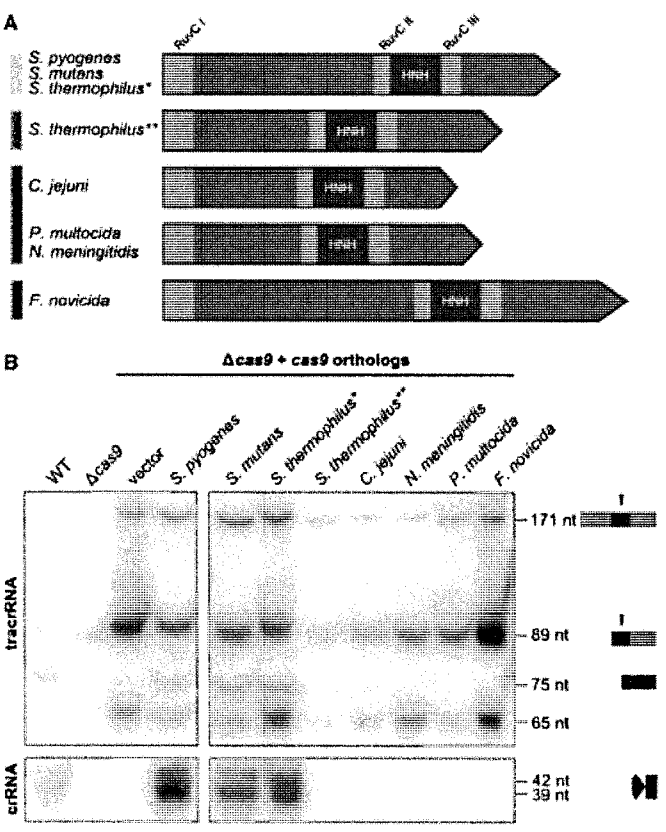


Figure 4

WO 2015/071474 A2

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

Published:

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

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

CRISPR-CAS SYSTEM MATERIALS AND METHODS

Field of the Invention

[0001] The invention relates to type II CRISPR-Cas systems of Cas9 enzymes, guide RNAs and associated specific PAMs. This application claims the benefit of the filing date of U.S. Provisional Patent Application No. 61/905,835 filed November 18, 2013, which is incorporated by reference herein in its entirety.

Incorporation by Reference of the Sequence Listing

[0002] This application contains, as a separate part of disclosure, a Sequence Listing in computer-readable form (filename: 48128_SeqListing.txt; 7,869,256 bytes – ASCII text file; created November 14, 2014) which is incorporated by reference herein in its entirety.

Background

[0003] Editing genomes using the RNA-guided DNA targeting principle of CRISPR-Cas (Clustered Regularly Interspaced Short Palindromic Repeats-CRISPR associated proteins) immunity has been exploited widely over the past few months (1-13). The main advantage provided by the bacterial type II CRISPR-Cas system lies in the minimal requirement for programmable DNA interference: an endonuclease, Cas9, guided by a customizable dual-RNA structure (14). As initially demonstrated in the original type II system of Streptococcus pyogenes, trans-activating CRISPR RNA (tracrRNA) (15,16) binds to the invariable repeats of precursor CRISPR RNA (pre-crRNA) forming a dual-RNA (14-17) that is essential for both RNA co-maturation by RNase III in the presence of Cas9 (15-17), and invading DNA cleavage by Cas9 (14,15,17-19). As demonstrated in Streptococcus, Cas9 guided by the duplex formed between mature activating tracrRNA and targeting crRNA (14-16) introduces site-specific double-stranded DNA (dsDNA) breaks in the invading cognate DNA (14,17-19). Cas9 is a multi-domain enzyme (14,20,21) that uses an HNH nuclease domain to cleave the target strand (defined as complementary to the spacer sequence of crRNA) and a RuvC-like domain to cleave the non-target strand (14,22,23), enabling the conversion of the dsDNA cleaving Cas9 into a nickase by selective motif inactivation (2,8,14,24,25). DNA cleavage specificity is determined by two parameters: the variable, spacer-derived sequence of crRNA targeting the protospacer sequence (a protospacer is defined as the sequence on the DNA target that is complementary to the spacer of crRNA) and a short sequence, the Protospacer Adjacent Motif (PAM), located immediately downstream of the protospacer on the non-target DNA strand (14,18,23,26-28).

[0004] Recent studies have demonstrated that RNA-guided Cas9 can be employed as an efficient genome editing tool in human cells (1,2,8,11), mice (9,10), zebrafish (6), drosophila (5), worms (4), plants (12,13), yeast (3) and bacteria (7). The system is versatile, enabling multiplex genome engineering by

programming Cas9 to edit several sites in a genome simultaneously by simply using multiple guide RNAs (2,7,8,10). The easy conversion of Cas9 into a nickase was shown to facilitate homology-directed repair in mammalian genomes with reduced mutagenic activity (2,8,24,25). In addition, the DNA-binding activity of a Cas9 catalytic inactive mutant has been exploited to engineer RNA-programmable transcriptional silencing and activating devices (29,30).

[0005] To date, RNA-guided Cas9 from *S. pyogenes, Streptococcus thermophilus,Neisseria meningitidis* and *Treponema denticola* have been described as tools for genome manipulation (1-13,24,25,31-34 and Esvelt et al. PMID: 24076762).

Summary

[0006] The present invention expands the RNA-programmable Cas9 toolbox to additional orthologous systems. The diversity and interchangeability of dual-RNA:Cas9 in eight representatives of phylogenetically defined type II CRISPR-Cas groups was examined herein. The results of this work not only introduce a wider range of Cas9 enzymes, guide RNA structures and associated specific PAMs but also enlighten the evolutionary aspects of type II CRISPR-Cas systems, including coevolution and horizontal transfer of the system components.

[0007] In an aspect, the present disclosure provides guide RNAs, both single-molecule and double-molecule guide RNAs, as well as methods for manipulating DNA in a cell using the guide RNAs and/or DNAs (including vectors) encoding the guide RNAs. Complexes comprising the guide RNAs and Cas9 endonucleases are also provided

[0008] In some embodiments, the single-molecule guide RNAs comprise a DNA-targeting segment and a protein-binding segment, wherein the protein-binding segment comprises a tracrRNA set out in Supplementary Table S5 or wherein the protein-binding segment comprises a tracrRNA at least 80% identical over at least 20 nucleotides to a tracrRNA set out in Supplementary Table S5. In some embodiments, the protein-binding segment comprises a CRISPR repeat set out in Supplementary Table S5 that is the CRISPR repeat cognate to the tracrRNA of the protein-binding segment. In some embodiments, the DNA-targeting segment comprises RNA complementary to a protospacer-like sequence in a target DNA 5' to a PAM sequence. In some embodiments, the tracrRNA and CRISPR repeat are respectively the C. jejuni tracrRNA and its cognate CRISPR repeat set out in Supplementary Table S5 and the PAM sequence is NNNNACA. In some embodiments, the tracrRNA and CRISPR repeat are respectively at least 80% identical to the C. jejuni tracrRNA and its cognate CRISPR repeat set out in Supplementary Table S5 and the PAM sequence is NNNNACA.

[0009] In some embodiments, the single-molecule guide RNA comprises a sequence that hybridizes to a protospacer-like sequence set out in one of SEQ ID NOs: 801-2701.

[0010] In another aspect, the disclosure provides a DNA encoding a single-molecule guide RNA of the invention.

[0011] In yet another aspect, the disclosure provides a vector comprising a DNA encoding a single-molecule guide RNA of the invention.

[0012] In still another aspect, the disclosure provides a cell comprising a DNA encoding a single-molecule guide RNA of the invention.

[0013] In an aspect, the disclosure provides a double-molecule guide RNA comprising: a targeter-RNA and an activator-RNA complementary thereto, wherein the activator-RNA comprises a tracrRNA set out in Supplementary Table S5 or wherein the activator-RNA comprises a tracrRNA at least 80% identical over at least 20 nucleotides to a tracrRNA set out in Supplementary Table S5. In some embodiments, the double-molecule guide RNA comprises a modified backbone, a non-natural internucleoside linkage, a nucleic acid mimetic, a modified sugar moiety, a base modification, a modification or sequence that provides for modified or regulated stability, a modification or sequence that provides for subcellular tracking, a modification or sequence that provides for tracking, or a modification or sequence that provides for a binding site for a protein or protein complex. In some embodiments, the targeter-RNA comprises a CRISPR repeat set out in Supplementary Table S5. In some embodiments, the targeter-RNA comprises a CRISPR repeat set out in Supplementary Table S5 that is the cognate CRISPR repeat of the tracrRNA of the activator-RNA. In some embodiments, the targeter-RNA further comprises RNA complementary to a protospacer-like sequence in a target DNA 5' to a PAM sequence. In some embodiments, the tracrRNA and CRISPR repeat are respectively the C. jejuni tracrRNA and its cognate CRISPR repeat set out in Supplementary Table S5 and the PAM sequence is NNNNACA. In some embodiments, the tracrRNA and CRISPR repeat are at least 80% identical to respectively the C. jejuni tracrRNA and its cognate CRISPR repeat set out in Supplementary Table S5 and the PAM sequence is NNNNACA.

[0014] In some embodiments, the double-molecule guide RNA comprises a sequence that hybridizes to a protospacer-like sequence set out in one of SEQ ID NOs: 801-2701.

[0015] In another aspect, the disclosure provides a DNA encoding a double-molecule guide RNA of the invention.

[0016] In yet another aspect, the disclosure provides a vector comprising a DNA encoding a double-molecule guide RNA of the invention.

[0017] In still another aspect, the disclosure provides a cell comprising a DNA encoding a double-molecule guide RNA of the invention.

[0018] In an aspect, the disclosure provides methods for manipulating DNA in a cell, comprising contacting the DNA with a Cas9 ortholog-guideRNA complex, wherein the complex comprises: (a) a C. jejuni Cas9 endonuclease heterologous to the cell or an endonuclease with an activity portion at least 90% identical to the activity portion of the C. jejuni Cas9 endonuclease, and a guide RNA targeting the complex to a protospacer-like sequence in the DNA 5' to the PAM sequence NNNNACA; (b) a P. multocida Cas9 endonuclease heterologous to the cell or an endonuclease with an activity portion at least

90% identical to the activity portion of the P. multocida Cas9 endonuclease, and a guide RNA targeting the complex to a protospacer-like sequence in the DNA 5' to the PAM sequence GNNNCNNA or NNNNC; (c) an F. novicida Cas9 endonuclease heterologous to the cell or an endonuclease with an activity portion at least 90% identical to the activity portion of the F. novicida Cas9 endonuclease, and a guide RNA targeting the complex to a protospacer-like sequence in the DNA 5' to the PAM sequence NG; (d) an S. thermophilus** Cas9 endonuclease heterologous to the cell or an endonuclease with an activity portion at least 90% identical to the activity portion of the S. thermophilus** Cas9 endonuclease, and a guide RNA targeting the complex to a protospacer-like sequence in the DNA 5' to the PAM sequence NNAAAAW; (e) an L. innocua Cas9 endonuclease heterologous to the cell or an endonuclease with an activity portion at least 90% identical to the activity portion of the L. innocua Cas9 endonuclease, and a guide RNA targeting the complex to a protospacer-like sequence in the DNA 5' to the PAM sequence NGG; or (f) an S. dysgalactiae Cas9 endonuclease heterologous to the cell or an endonuclease with an activity portion at least 90% identical to the activity portion of the S. dysgalactiae Cas9 endonuclease, and a guide RNA targeting the complex to a protospacer-like sequence in the DNA 5' to the PAM sequence NGG. In some embodiments, the guide is a single-molecule guide RNA. In some embodiments, the guide RNA is a double-molecule guide RNA. The complexes used in the methods are also provided.

[0019] In some embodiments of the methods, the protospacer-like sequence targeted is in a CCR5, CXCR4, KRT5, KRT14, PLEC or COL7A1 gene. In some embodiments, the protospacer-like sequence is in a chronic granulomatous disease (CGD)-related gene CYBA, CYBB, NCF1, NCF2 or NCF4. In some embodiments, the protospacer-like sequence targeted is in a gene encoding B-cell lymphoma/leukemia IIA (BCL11A) protein, an erythroid enhancer of BCL11A or a BCL11A binding site. In some embodiments, the protospacer-like sequence targeted is up to 1000 nucleotides upstream of the above mentioned genes. In some embodiments of the methods, the guide RNA comprises a sequence complementary to a protospacer-like sequence set out in one of SEQ ID NOs: 801-2701.

[0020] In an aspect, the disclosure provides a recombinant vector encoding: (a) a guide RNA, wherein the guide RNA comprises a DNA-targeting segment complementary to a protospacer-like sequence in the DNA 5' to the PAM sequence NNNNACA; and (b) a C. jejuni Cas9 endonuclease (for example, set out in SEQ ID NO: 50) or an endonuclease with an activity portion at least 90% identical to the activity portion of the C. jejuni Cas9 endonuclease. In some embodiments, the DNA-targeting segment complementary to the protospacer-like sequence is RNA complementary to the target sequences set out in one of SEQ ID NOs: 801-973, 1079-1222, 1313-1348, 1372-1415, 1444-1900, 2163-2482 or 2667-2686. Methods of using the vectors to manipulate DNA in a cell are also provided.

[0021] In another aspect, the disclosure provides a recombinant vector encoding: (a) a guide RNA, wherein the guide RNA comprises a DNA-targeting segment complementary to a protospacer-like sequence in the DNA 5' to the PAM sequence GNNNCNNA or NNNNC; and (b) a P. multocida Cas9 endonuclease (for example, set out in SEQ ID NO: 1) or an endonuclease with an activity portion at least 90% identical to the activity portion of the P. multocida Cas9 endonuclease. In some embodiments, the

DNA-targeting segment complementary to the protospacer-like sequence is RNA complementary to the target sequences set out in one of SEQ ID NOs:974-1078, 1223-1312, 1349-1371, 1416-1443, 1901-2162, 2483-2666 or 2687-2701. Methods of using the vectors to manipulate DNA in a cell are also provided.

[0022] In yet another aspect, the disclosure provides a recombinant vector encoding: (a) a guide RNA, wherein the guide RNA comprises a DNA-targeting segment complementary to a protospacer-like sequence in the DNA 5' to the PAM sequence NG; and (b) a F. novicida Cas9 endonuclease (fore example, set out in SEQ ID NO: 43) or an endonuclease with an activity portion at least 90% identical to the activity portion of the F. novicida Cas9 endonuclease. Methods of using the vectors to manipulate DNA in a cell are also provided.

[0023] In still another aspect, the disclosure provides a recombinant vector encoding: (a) a guide RNA, wherein the guide RNA comprises a DNA-targeting segment complementary to a protospacer-like sequence in the DNA 5' to the PAM sequence NNAAAAW; and (b) a S. thermophilus** Cas9 endonuclease or an endonuclease with an activity portion at least 90% identical to the activity portion of the S. thermophilus** Cas9 endonuclease. Methods of using the vectors to manipulate DNA in a cell are also provided.

[0024] In yet another aspect, the disclosure provides a recombinant vector encoding:(a) a guide RNA, wherein the guide RNA comprises a DNA-targeting segment complementary to a protospacer-like sequence in the DNA 5' to the PAM sequence NGG; and (b) a L. innocua Cas9 endonuclease (for example, set out in SEQ ID NO: 3) or an endonuclease with an activity portion at least 90% identical to the activity portion of the L. innocua Cas9 endonuclease. Methods of using the vectors to manipulate DNA in a cell are also provided.

[0025] In still another aspect, the disclosure provides a recombinant vector encoding: (a) a guide RNA, wherein the guide RNA comprises a DNA-targeting segment complementary to a protospacer-like sequence in the DNA 5' to the PAM sequence NGG; and (b) a S. dysgalactiae Cas9 endonuclease (for example, set out in SEQ ID NO: 105) or an endonuclease with an activity portion at least 90% identical to the activity portion of the S. dysgalactiae Cas9 endonuclease.

[0026] In some embodiments of the vectors, the guide RNA comprises a sequence complementary to a protospacer-like sequence set out in one of SEQ ID NOs: 801-2701.

[0027] In a related aspect, the disclosure provides a method comprising (a) identifying at least 7-20 bases of mammalian genomic DNA adjacent to any of the preceding protospacer-like sequences, and (b) manipulating the mammalian genomic DNA sequence by contacting a mammalian cell with, or administering to a mammal, (i) a DNA-targeting segment complementary to the DNA sequence identified in step (a) and (ii) a protein-binding segment, or nucleic acid(s) encoding (i) and (ii), and (iii) a cas9 endonuclease or a nucleic acid encoding said cas9 endonuclease; and (c) detecting cleavage of the mammalian genomic DNA.

[0028] In an aspect, the disclosure provides a modified Cas9 endonuclease, modified from any of the Cas9 orthologs disclosed herein, comprising one or more mutations corresponding to S. pyogenes Cas9 mutation E762A, HH983AA or D986A. In some embodiments, the modified Cas 9 endonuclease further comprises one or more mutations corresponding to S. pyogenes Cas9 mutation D10A, H840A, G12A, G17A, N854A, N863A, N982A or A984A.

[0029] In an aspect, the disclosure provides a method for manipulating DNA in a cell, comprising contacting the DNA with a Cas9 ortholog-guide RNA complex, wherein the complex comprises: (a) a Cas9 endonuclease heterologous to the cell and (b) a cognate guide RNA of the Cas9 endonuclease comprising a tracrRNA set out in Supplementary Table S5 or a guide RNA comprising a tracrRNA at least 80% identical to a cognate tracrRNA set out in Supplementary Table S5 over at least 20 nucleotides. In some embodiments, the guide is a single-molecule guide RNA. In some embodiments, the guide RNA is a double-molecule guide RNA. In some embodiments of the methods, the guide RNA comprises a sequence complementary to a protospacer-like sequence set out in one of SEQ ID NOs: 801-2701. Complexes used in the methods are also provided.

[0030] In an aspect, the disclosure provides a method for manipulating DNA in a cell, comprising contacting the DNA with a Cas9 ortholog-guide RNA complex, wherein the complex comprises: (a) a cognate guide RNA for a first Cas9 endonuclease from a cluster in Supplementary Table S2 and (b) a second Cas9 endonuclease from the same cluster that is exchangeable with preserved high cleavage efficiency with the first endonuclease and shares at least 80% identity with the first endonuclease over 80% of their length. In some embodiments, the guide is a single-molecule guide RNA. In some embodiments, the guide RNA is a double-molecule guide RNA. In some embodiments, the first Cas9 endonuclease is from S. pyogenes and the second Cas9 endonuclease is from S. mutans. In some embodiments, the first Cas9 endonuclease is from S. thermophilus* and the second Cas9 endonuclease is from S. mutans. In some embodiments, the first Cas9 endonuclease is from N. meningitidis and the second Cas9 endonuclease is from P. multocida. Complexes used in the methods are also provided.

[0031] In an aspect, the disclosure provides a method for manipulating DNA in a cell, comprising contacting the DNA with a Cas9 ortholog-guide RNA complex, wherein the complex comprises: (a) a cognate guide RNA of a first Cas9 endonuclease from a cluster in Supplementary Table S6 and (b) an Cas9 endonuclease from the same cluster in Supplementary Table S6 that is exchangeable with the same or lowered cleavage efficiency with the first endonuclease and shares at least 50% amino acid sequence identity with the first endonuclease over 70% of their length. In some embodiments, the guide is a single-molecule guide RNA. In some embodiments, the guide RNA is a double-molecule guide RNA. In some embodiments, the first Cas9 endonuclease is from C. Jejuni and the second Cas9 endonuclease is from P. multocida. In some embodiments, the first Cas9 endonuclease is from N. meningitidis and the second Cas9 endonuclease is from P. multocida. Complexes used in the methods are also provided.

[0032] In an aspect, the disclosure provides a method for manipulating DNA in a cell, comprising contacting the DNA with two or more Cas9-guide RNA complexes, wherein each Cas9-guideRNA complex comprises: (a) a Cas9 endonuclease from a different cluster in Supplementary Table S6 exhibiting less than 50% amino acid sequence identity with the other endonucleases of the method over 70% of their length, and (b) a guide RNA specifically complexed with each Cas9 endonuclease. In some embodiments, the guide is a single-molecule guide RNA. In some embodiments, the guide RNA is a double-molecule guide RNA.In some embodiments, the Cas9 endonucleases are from F. novicida and S. pyogenes. In some embodiments, the Cas9 endonucleases are from N. meningitidis and S. mutans. In some embodiments, the Cas9 endonucleases are the S. thermophilus* Cas9 and the S. thermophilus** Cas9. Complexes used in the methods are also provided.

[0033] In some embodiments of the manipulation methods, the DNA targeted in the cell is a CCR5, CXCR4, KRT5, KRT14, PLEC or COL7A1 gene. In some embodiments, the DNA targeted in the cell is a chronic granulomatous disease (CGD)-related gene CYBA, CYBB, NCF1, NCF2 or NCF4. In some embodiments, the protospacer-like sequence targeted is in a gene encoding B-cell lymphoma/leukemia IIA (BCL11A) protein, an erythroid enhancer of BCL11A or a BCL11A binding site. In some embodiments, the protospacer-like sequence targeted is up to 1000 nucleotides upstream of the above mentioned genes. In some embodiments of the methods, the guide RNA comprises a sequence complementary to a protospacer-like sequence set out in one of SEQ ID NOs: 801-2701.

[0034] It is contemplated that any of the methods provided herein may ex vivo or in vivo.

Brief Description of the Drawings

[0035] **Figure 1.** Phylogeny of representative Cas9 orthologs and schematic representation of selected bacterial type II CRISPR-Cas systems. (**A**) Phylogenetic tree of Cas9 reconstructed from selected, informative positions of representative Cas9 orthologs multiple sequence alignment is shown (see Supplementary Figure S2 and Supplementary Table S2). The Cas9 orthologs of the subtypes classified as II-A, II-B and II-C are highlighted with shaded boxes. The colored branches group distinct proteins of closely related loci with similar locus architecture (15). Each protein is represented by the GenInfo (GI) identifier followed by the bacterial strain name. The bootstrap values are given for each node (see Materials and Methods). Note that the monophyletic clusters of subtypes II-A and II-B are supported by high bootstrap values. The scale bar for the branch length is given as the estimated number of amino acid substitution per site. (**B**) Genetic loci of type II (Nmeni/CASS4) CRISPR-Cas in *Streptococcus pyogenes* SF370, *Streptococcus mutans* UA159, *Streptococcus thermophilus* LMD-9 *(CRISPR3), **(CRISPR1), *Campylobacter jejuni* NCTC 11168, *Neisseria meningitidis* Z2491, *Pasteurella multocida* Pm70 and *Francisella novicida* U112. Red arrow, transcription direction of tracrRNA; blue arrows, *cas* genes; black rectangles, CRISPR repeats; green diamonds, spacers; thick black line, leader sequence; black arrow, putative pre-crRNA promoter; HP, Hypothetical Protein. The colored bars

represented on the left correspond to Cas9 tree branches colors. The transcription direction and putative leader position of *C. jejuni* and *N. meningitidis* pre-crRNAs were derived from previously published RNA sequencing data (15). The CRISPR-Cas locus architecture of *P. multocida* was predicted based on its close similarity to that of *N. meningitidis* and further confirmed by bioinformatics prediction of tracrRNA based on a strongly predicted promoter and a transcriptional terminator as described in (15). Type II CRISPR-Cas loci can differ in the *cas* gene composition, mostly with *cas9*, *cas1* and *cas2* being the minimal set of genes (type II-C, blue), sometimes accompanied with a fourth gene *csn2a/b* (type II-A, yellow and orange) or *cas4* (type II-B, green). The CRISPR array can be transcribed in the same (type II-A, yellow and orange) or in the opposite (types II-B and C, blue and green) direction of the *cas* operon. The location of tracrRNA and the direction of its transcription differ within the groups (compare type II-A of *S. thermophilus*** with type II-A from the other species indicated here (yellow) and compare type II-C of *C. jejuni* with type II-C of *N. meningitidis* and *P. multocida* (blue)).

[0036] **Figure 2.** RNase III is a general executioner of tracrRNA:pre-crRNA processing in type II CRISPR-Cas. Northern blot analysis of total RNA from *S. pyogenes* WT, Δ*rnc* and Δ*rnc* complemented with *rnc* orthologs or mutants (truncated *rnc* and inactivated (dead) (D51A) *rnc*) probed for tracrRNA (top) and crRNA repeat (bottom). RNA sizes in nt and schematic representations of tracrRNA (red-black) and crRNA (green-black) are indicated on the right (16). The vertical black arrows indicate the processing sites. tracrRNA-171 nt and tracrRNA-89 nt forms correspond to primary tracrRNA transcripts. The presence of tracrRNA-75 nt and crRNA 39-42 nt forms indicates tracrRNA and pre-crRNA co-processing. *S. pyogenes* tracrRNA and pre-crRNA are co-processed by all analyzed RNase III orthologs. The truncated version and catalytic inactive mutant of *S. pyogenes* RNase III are both deficient in tracrRNA:pre-crRNA processing.

[0037] **Figure 3.** Conserved motifs of Cas9 are required for DNA interference but not for dual-RNA processing by RNase III. (**A**) Schematic representation of *S. pyogenes* Cas9. The conserved HNH and splitted RuvC motifs and analyzed amino acids are indicated. (**B**) Northern blot analysis of total RNA from *S. pyogenes* WT, Δ*cas9* and Δ*cas9* complemented with pEC342 or pEC342 containing *cas9* WT or mutant genes, probed for tracrRNA and crRNA repeat. Maturation of tracrRNA and pre-crRNA generating tracrRNA-75 nt and crRNA-39-42 nt forms is observed in all Δ*cas9* strains complemented with the *cas9* mutants. (**C**) *In vivo* protospacer targeting. Transformation assays of *S. pyogenes* WT and Δ*cas9* with pEC85 (vector), pEC85Ω*cas9* (*cas9*), pEC85Ω*speM* (*speM*), and pEC85ΩtracrRNA-171 nt plasmids containing *speM* and *cas9* mutants. The CFUs (colony forming units) per µg of plasmid DNA were determined in at least three independent experiments. The results +/- SD of technical triplicates of one representative experiment are shown. Cas9 N854A is the only mutant that did not tolerate the protospacer plasmid as observed for WT Cas9, indicating that this residue is not involved in DNA interference. (**D**) *In vitro* plasmid cleavage. Agarose gel electrophoresis of plasmid DNA (5 nM) containing *speM* protospacer (pEC287) incubated with 25 nM Cas9 WT or mutants in the presence of equimolar

amounts of dual-RNA-*speM* (see Materials and Methods). Cas9 WT and N854A generated linear cleavage products while the other Cas9 mutants created only nicked products. M, 1 kb DNA ladder (Fermentas); oc: open circular, li: linear; sc: supercoiled.

[0038] **Figure 4.** Cas9 from closely related CRISPR-Cas systems can substitute the role of *S. pyogenes* Cas9 in RNA processing by RNase III. (**A**) Schematic representation of Cas9 from selected bacterial species. The protein sizes and distances between conserved motifs (RuvC and HNH) are drawn in scale. See Supplementary Figure S1. (**B**) Northern blot analysis of total RNA extracted from *S. pyogenes* WT, Δ*cas9* and Δ*cas9* complemented with pEC342 (backbone vector containing tracrRNA-171 nt and the *cas* operon promoter from *S. pyogenes*) or pEC342-based plasmids containing *cas9* orthologous genes, probed for tracrRNA and crRNA repeat. Mature forms of *S. pyogenes* tracrRNA and pre-crRNA are observed only in the presence of *S. pyogenes* Cas9 WT or closely related Cas9 orthologs from *S. mutans* and *S. thermophilus**.

[0039] **Figure 5.** Cas9 orthologs cleave DNA in the presence of their cognate dual-RNA and specific PAM *in vitro.* (**A**) Logo plot of protospacer adjacent sequences derived from BLAST analysis of spacer sequences for selected bacterial species. The logo plot gives graphical representation of most abundant nucleotides downstream of the protospacer sequence. The numbers in brackets correspond to the number of analyzed protospacers. (**B**) DNA substrates designed for specific PAM verification. Based on the logo plot for each species, plasmid DNA substrates were designed to contain the *speM* protospacer and the indicated sequence downstream, either comprising (PAM+) or not (PAM-) the proposed PAM. The predicted PAMs were verified by cleavage assays narrowing down the necessary nucleotides for activity (data not shown); therefore the sequence used differs slightly from the logoplot shown in (**A**). The high abundance of other nucleotides not being part of the PAM can be explained by redundancy of the coding sequences containing the protospacers, and by the limited number of found protospacer targets. The last column shows the PAM sequence for each species, which was already published (no symbol) or derived from this work (#). (**C**) *In vitro* plasmid cleavage assays by dual-RNA:Cas9 orthologs on plasmid DNA with the 10 bp protospacer adjacent sequence (summarized in (**B**)). Each Cas9 ortholog in complex with its cognate dual-RNA cleaves plasmids containing the corresponding species-specific PAM (PAM+). No cleavage is observed with plasmids that did not contain the specific PAM (PAM-). li: linear cleavage product, sc: supercoiled plasmid DNA.

[0040] **Figure 6.** Cas9 and dual-RNA co-evolved. (**A**) *In vitro* plasmid cleavage assays using *S. pyogenes* Cas9 in complex with orthologous dual-RNA (upper panel) and orthologous Cas9 enzymes in complex with *S. pyogenes* dual-RNA (lower panel). Plasmid DNA containing protospacer *speM* and *S. pyogenes* PAM (NGG) was incubated with different dual-RNAs in complex with *S. pyogenes* Cas9. tracrRNA and crRNA-repeat sequences of the dual-RNAs are from the indicated bacterial species, with crRNA spacer targeting *speM*. In the lower panel, plasmid DNA containing *speM* protospacer and the specific PAM was incubated with Cas9 orthologs in complex with *S. pyogenes* dual-RNA. *S. pyogenes*

Cas9 can cleave plasmid DNA only in the presence of dual-RNA from *S. pyogenes*, *S. mutans* and *S. thermophilus** (yellow). Dual-RNA from *S. pyogenes* can mediate DNA cleavage only with Cas9 from *S. pyogenes*, *S. mutans* and *S. thermophilus** (yellow). li: linear cleavage product; sc: supercoiled plasmid DNA. (**B**) Summary of Cas9 and dual-RNA orthologs exchangeability. Specific PAM sequences were used according to Figure 5. The color code reflects the type II CRISPR-Cas subgroups (Figure 1). +++: 100 - 75% cleavage activity; ++: 75 - 50% cleavage activity; +: 50 - 25% cleavage activity; -: 25 - 0% cleavage activity observed under the conditions tested. Cas9 and dual-RNA duplexes from the same type II group can be interchanged and still mediate plasmid cleavage providing that the PAM sequence is specific for Cas9. See also Supplementary Figure S10.

[0041] **Supplementary Figure S1**. Biochemical characteristics and SDS-PAGE analysis of Cas9 proteins purified in this study. (A) Overview of characteristics of Cas9 orthologous proteins [a]Note that the biochemical characteristics of *S. pyogenes* Cas9 WT and mutants are identical; [b]GenInfo (GI) Identifier; [c]ε, Extinction coefficient. (**B**) SDS PAGE analysis of purified mutants of Cas9 from *S. pyogenes*. (**C**) SDS PAGE analysis of purified Cas9 orthologs. M: PageRulerTM Unstained Protein Ladder (Thermo Scientific).

[0042] **Supplementary Figure S2.** Multiple sequence alignment of representative Cas9 sequences (see Supplementary Table S2 and Material and Methods). The rows described as Jnet with following GI identifier of a selected Cas9 sequence provide the predicted secondary structure of Cas9 within the corresponding subgroups (sequences indicated below each Jnet). Conserved motifs are marked below the alignment and the mutated amino acid residues are highlighted. Asterisks indicate informative positions chosen for the Cas9 tree reconstruction.

[0043] **Supplementary Figure S3.** Multiple sequence alignment of representative Cas1 sequences (see Supplementary Table S2 and Materials and Methods). Informative positions chosen for the Cas1 tree reconstruction are marked with asterisks at the bottom of the alignment.

[0044] **Supplementary Figure S4.** Phylogenetic analysis of representative Cas9 and Cas1 sequences. Phylogenetic trees of Cas1 (left) and Cas9 (right) reconstructed from selected, informative positions of Cas1 and Cas9 multiple sequence alignments are shown (see Figure 1 and Supplementary Figures S2 and S3). The Cas1 tree is rooted to the outgroup of selected Cas1 orthologs of type I CRISPR-Cas systems. The Cas1 and Cas9 orthologs of the types classified as II-A, II-B and II-C are highlighted with shaded boxes. The same branch colors were used for each bacterial strain on both trees. Each protein is represented by the GenInfo (GI) identifier followed by the bacterial strain name. The bootstrap values are given for each node (see Materials and Methods). The scale bars for the branch length are given as the estimated number of amino acid substitution per site. Note the similarity of the trees topology and monophyletic clusters of subtypes II-A and II-B on both trees supported by high bootstrap values.

[0045] **Supplementary Figure S5.** RNase III is a general executioner of tracrRNA:pre-crRNA processing in type II CRISPR-Cas. Northern blot analysis of total RNA from *S. pyogenes* WT, Δ*rnc* and Δ*rnc* complemented with *rnc* orthologs or *rnc* mutants probed with (**A**) tracrRNA and (**B**) crRNA repeat (Supplementary Table S1). The dashed-line boxes represented below the Northern blots in (**B**) show the area of the blots with enhanced exposure. All RNAse III orthologs can co-process *S. pyogenes* tracrRNA and pre-crRNA. No mature forms of tracrRNA and crRNAs could be observed in Δ*rnc* complemented with the truncated version or catalytically inactive (dead) mutant of RNase IIII.

[0046] **Supplementary Figure S6.** Multiple sequence alignment of bacterial endoribonucleases III used in the study. Domains indicated below the alignment are according to the domains identified in RNase III from *E. coli* (58, 59). The conserved catalytic aspartate residue mutated in the catalytically inactive "*rnc* dead" mutant and the last amino acid of the truncated *rnc* mutant are indicated above the alignment with an asterisk and an arrow, respectively.

[0047] **Supplementary Figure S7.** Conserved catalytic amino acid residues of Cas9 are not involved in dual-RNA processing by RNase III. Northern blot analysis of total RNA extracted from *S. pyogenes* WT, Δ*cas9* and Δ*cas9* complemented with pEC342 (backbone vector containing tracrRNA-171 nt and the native *cas* operon promoter from *S. pyogenes*) or pEC342-derived plasmids encoding Cas9 WT or mutants, hybridized with (**A**) tracrRNA or (**B**) crRNA repeat probe (Supplementary Table S1). tracrRNA:crRNA co-processing is observed in all strains encoding Cas9 point mutants. Note that in a previous study, we observed low abundance of tracrRNA in the *cas9* deletion mutant (16). For this reason, plasmids used in *cas9* complementation studies were designed to encode tracrRNA in addition to *cas9*.

[0048] **Supplementary Figure S8.** Cas9 and tracrRNA:crRNA co-evolved. Northern blot analysis of total RNA extracted from *S. pyogenes* WT, *Δcas9* and *Δcas9* complemented with pEC342 or pEC342-derived plasmids encoding Cas9 WT or mutants - hybridized with (**A**) tracrRNA or **(B)** crRNA repeat probe (Supplementary Table S1). Only *S. pyogenes* Cas9 WT and closely related Cas9 orthologs from *S. mutans* and *S. thermophilus** (CRISPR3) can contribute to coprocessing of *S. pyogenes* tracrRNA:pre-crRNA.

[0049] **Supplementary Figure S9.** Cas9 orthologs cleave plasmid DNA in the presence of their cognate dual-RNA and specific PAM. Agarose gel electrophoresis analysis of dual-RNA:Cas9 titration (0-100 nM dual-RNA-Cas9 complex) on plasmid DNA (5 nM) containing *speM* protospacer and adjacent WT PAM (PAM+), imperfect PAM (PAM±) or no PAM (PAM–). For *S. pyogenes*, *S. mutans*, *S. thermophilus**, *S. thermophilus*** and *N. meningitidis,* the PAM sequence has already been published (27,28,53,54). For the other bacterial species, PAMs were predicted based on the downstream sequence of protospacer identified in the investigated or related strains (see Supplementary Table S2 and Materials and Methods). The 10 bp sequence located directly downstream of the crRNA-targeted *speM* protospacer is shown. The

nucleotide(s) predicted to belong to the PAM sequence are shaded in grey. li: linear cleavage product, sc: supercoiled plasmid DNA, M: 1 kb DNA ladder.

[0050] **Supplementary Figure S10.** Summary of *in vitro* plasmid cleavage assays of Cas9 orthologs in combination with dual-RNAs. Agarose gel electrophoresis of cleavage assays. (**A**) *S. mutans* Cas9 (50 nM), (**B**) *S. thermophilus** Cas9 (25 nM), (**C**) *S. thermophilus*** Cas9 (100 nM), (D) *C. jejuni* Cas9 (100 nM), (**E**) *N. meningitidis* Cas9 (100 nM), (**F**) *P. multocida* Cas9 (25 nM), (**G**) *F. novicida* Cas9 (100 nM) in complex with equimolar concentrations of each of the dual-RNA orthologs were incubated with plasmid DNA (5 nM) containing *speM* protospacer sequence and the PAM sequence specific to the Cas9 ortholog analyzed. li: linear cleavage product, sc: supercoiled plasmid DNA, M: 1 kb DNA ladder.

[0051] **Supplementary Figure S11.** Cas9 tree topology suggests both horizontal and vertical transfer of type II CRISPR-Cas systems. See Figure 1, Supplementary Figure S4 and Supplementary Table S4. The codes for taxonomy (phyla in color) and habitat (symbols) of the bacterial strains harbouring representative Cas9 orthologs are indicated (right panel). The clusters grouping evolutionary distant bacteria (1 and 3) but isolated mainly from similar sources (human for cluster 1 and mostly environmental samples for cluster 3) suggest horizontal transfer of type II systems. Clusters 2, 4 and 5 group closely related bacteria isolated from diverse habitats indicating vertical transfer of the systems.

[0052] **Supplementary Figure S12.** tracrRNA:crRNA repeat duplexes form similar secondary structures in loci with closely related Cas9 orthologs. Antirepeat sequence of processed tracrRNA (red) and repeat-derived sequence of mature crRNA (grey) were co-folded for each type II CRISPR-Cas locus studied (see Materials and Methods). Color bars indicated on the left group dual-RNAs from loci with closely related Cas9 (see Figure 1 and Supplementary Figure S4). RNA duplexes belonging to the same groups display structural similarities, suggesting a role of the structure in dual-RNA recognition by Cas9.

Detailed Description

[0053] Terminology

[0054] All technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs, unless the technical or scientific term is defined differently herein.

[0055] The terms "polynucleotide" and "nucleic acid," used interchangeably herein, refer to a polymeric form of nucleotides of any length, either ribonucleotides or deoxyribonucleotides. Thus, this term includes, but is not limited to, single-, double-, or multi-stranded DNA or RNA, genomic DNA, cDNA, DNA-RNA hybrids, or a polymer comprising purine and pyrimidine bases or other natural, chemically or biochemically modified, non-natural, or derivatized nucleotide bases. "Oligonucleotide" generally refers to

polynucleotides of between about 5 and about 100 nucleotides of single- or double-stranded DNA. However, for the purposes of this disclosure, there is no upper limit to the length of an oligonucleotide. Oligonucleotides are also known as "oligomers" or "oligos" and may be isolated from genes, or chemically synthesized by methods known in the art. The terms "polynucleotide" and "nucleic acid" should be understood to include, as applicable to the embodiments being described, single-stranded (such as sense or antisense) and double-stranded polynucleotides.

[0056] "Genomic DNA" refers to the DNA of a genome of an organism including, but not limited to, the DNA of the genome of a bacterium, fungus, archea, plant or animal.

[0057] "Manipulating" DNA encompasses binding, nicking one strand, or cleaving (i.e., cutting) both strands of the DNA, or encompasses modifying the DNA or a polypeptide associated with the DNA (e.g., the modifications of paragraphs [00161] or [00162]) . Manipulating DNA can silence, activate, or modulate (either increase or decrease) the expression of an RNA or polypeptide encoded by the DNA.

[0058] A "stem-loop structure" refers to a nucleic acid having a secondary structure that includes a region of nucleotides which are known or predicted to form a double strand (stem portion) that is linked on one side by a region of predominantly single-stranded nucleotides (loop portion). The terms "hairpin" and "fold-back" structures are also used herein to refer to stem-loop structures. Such structures are well known in the art and these terms are used consistently with their known meanings in the art. As is known in the art, a stem-loop structure does not require exact base-pairing. Thus, the stem may include one or more base mismatches. Alternatively, the base-pairing may be exact, i.e. not include any mismatches.

[0059] By "hybridizable" or "complementary" or "substantially complementary" it is meant that a nucleic acid (e.g. RNA) comprises a sequence of nucleotides that enables it to non-covalently bind, i.e. form Watson-Crick base pairs and/or G/U base pairs, "anneal", or "hybridize," to another nucleic acid in a sequence-specific, antiparallel, manner (i.e., a nucleic acid specifically binds to a complementary nucleic acid) under the appropriate in vitro and/or in vivo conditions of temperature and solution ionic strength. As is known in the art, standard Watson-Crick base-pairing includes: adenine (A) pairing with thymidine (T), adenine (A) pairing with uracil (U), and guanine (G) pairing with cytosine (C) [DNA, RNA]. In addition, it is also known in the art that for hybridization between two RNA molecules (e.g., dsRNA), guanine (G) base pairs with uracil (U). For example, G/U base-pairing is partially responsible for the degeneracy (i.e., redundancy) of the genetic code in the context of tRNA anti-codon base-pairing with codons in mRNA. In the context of this disclosure, a guanine (G) of a protein-binding segment (dsRNA duplex) of a guide RNA molecule is considered complementary to a uracil (U), and vice versa. As such, when a G/U base-pair can be made at a given nucleotide position a protein-binding segment (dsRNA duplex) of a guide RNA molecule, the position is not considered to be non-complementary, but is instead considered to be complementary.

[0060] Hybridization and washing conditions are well known and exemplified in Sambrook, J., Fritsch, E. F. and Maniatis, T. Molecular Cloning: A Laboratory Manual, Second Edition, Cold Spring Harbor Laboratory Press, Cold Spring Harbor (1989), particularly Chapter 11 and Table 11.1 therein; and Sambrook, J. and Russell, W., Molecular Cloning: A Laboratory Manual, Third Edition, Cold Spring Harbor Laboratory Press, Cold Spring Harbor (2001). The conditions of temperature and ionic strength determine the "stringency" of the hybridization.

[0061] Hybridization requires that the two nucleic acids contain complementary sequences, although mismatches between bases are possible. The conditions appropriate for hybridization between two nucleic acids depend on the length of the nucleic acids and the degree of complementation, variables well known in the art. The greater the degree of complementation between two nucleotide sequences, the greater the value of the melting temperature (Tm) for hybrids of nucleic acids having those sequences. For hybridizations between nucleic acids with short stretches of complementarity (e.g. complementarity over 35 or less, 30 or less, 25 or less, 22 or less, 20 or less, or 18 or less nucleotides) the position of mismatches becomes important (see Sambrook et al., supra, 11.7-11.8). Typically, the length for a hybridizable nucleic acid is at least about 10 nucleotides. Illustrative minimum lengths for a hybridizable nucleic acid are: at least about 15 nucleotides; at least about 20 nucleotides; at least about 22 nucleotides; at least about 25 nucleotides; and at least about 30 nucleotides). Furthermore, the skilled artisan will recognize that the temperature and wash solution salt concentration may be adjusted as necessary according to factors such as length of the region of complementation and the degree of complementation.

[0062] It is understood in the art that the sequence of polynucleotide need not be 100% complementary to that of its target nucleic acid to be specifically hybridizable or hybridizable. Moreover, a polynucleotide may hybridize over one or more segments such that intervening or adjacent segments are not involved in the hybridization event (e.g., a loop structure or hairpin structure). A polynucleotide can comprise at least 70%, at least 80%, at least 90%, at least 95%, at least 99%, or 100% sequence complementarity to a target region within the target nucleic acid sequence to which they are targeted. For example, an antisense nucleic acid in which 18 of 20 nucleotides of the antisense compound are complementary to a target region, and would therefore specifically hybridize, would represent 90 percent complementarity. In this example, the remaining noncomplementary nucleotides may be clustered or interspersed with complementary nucleotides and need not be contiguous to each other or to complementary nucleotides. Percent complementarity between particular stretches of nucleic acid sequences within nucleic acids can be determined routinely using BLAST programs (basic local alignment search tools) and PowerBLAST programs known in the art (Altschul et al., J. Mol. Biol., 1990, 215, 403-410; Zhang and Madden, Genome Res., 1997, 7, 649-656) or by using the Gap program (Wisconsin Sequence Analysis Package, Version 8 for Unix, Genetics Computer Group, University

Research Park, Madison Wis.), using default settings, which uses the algorithm of Smith and Waterman (Adv. Appl. Math., 1981, 2, 482-489).

[0063] The terms "peptide," "polypeptide," and "protein" are used interchangeably herein, and refer to a polymeric form of amino acids of any length, which can include coded and non-coded amino acids, chemically or biochemically modified or derivatized amino acids, and polypeptides having modified peptide backbones.

[0064] "Binding" as used herein (e.g. with reference to an RNA-binding domain of a polypeptide) refers to a non-covalent interaction between macromolecules (e.g., between a protein and a nucleic acid). While in a state of non-covalent interaction, the macromolecules are said to be "associated" or "interacting" or "binding" (e.g., when a molecule X is said to interact with a molecule Y, it is meant the molecule X binds to molecule Y in a non-covalent manner). Not all components of a binding interaction need be sequence-specific (e.g., contacts with phosphate residues in a DNA backbone), but some portions of a binding interaction may be sequence-specific. Binding interactions are generally characterized by a dissociation constant (K_d) of less than 10^{-6} M, less than 10^{-7} M, less than 10^{-8} M, less than 10^{-9} M, less than 10^{-10} M, less than 10^{-11} M, less than 10^{-12} M, less than 10^{-13} M, less than 10^{-14} M, or less than 10^{-15} M. "Affinity" refers to the strength of binding, increased binding affinity being correlated with a lower K_d. By "binding domain" it is meant a protein domain that is able to bind non-covalently to another molecule. A binding domain can bind to, for example, a DNA molecule (a DNA-binding protein), an RNA molecule (an RNA-binding protein) and/or a protein molecule (a protein-binding protein). In the case of a protein domain-binding protein, it can bind to itself (to form homodimers, homotrimers, etc.) and/or it can bind to one or more molecules of a different protein or proteins.

[0065] The term "conservative amino acid substitution" refers to the interchangeability in proteins of amino acid residues having similar side chains. For example, a group of amino acids having aliphatic side chains consists of glycine, alanine, valine, leucine, and isoleucine; a group of amino acids having aliphatic-hydroxyl side chains consists of serine and threonine; a group of amino acids having amide containing side chains consisting of asparagine and glutamine; a group of amino acids having aromatic side chains consists of phenylalanine, tyrosine, and tryptophan; a group of amino acids having basic side chains consists of lysine, arginine, and histidine; a group of amino acids having acidic side chains consists of glutamate and aspartate; and a group of amino acids having sulfur containing side chains consists of cysteine and methionine. Exemplary conservative amino acid substitution groups are: valine-leucine-isoleucine, phenylalanine-tyrosine, lysine-arginine, alanine-valine, and asparagine-glutamine.

[0066] A polynucleotide or polypeptide has a certain percent "sequence identity" to another polynucleotide or polypeptide, meaning that, when aligned, that percentage of bases or amino acids are the same, and in the same relative position, when comparing the two sequences. Sequence identity can be determined in a number of different manners. To determine sequence identity, sequences can be aligned

using various methods and computer programs (e.g., BLAST, T-COFFEE, MUSCLE, MAFFT, etc.), available over the world wide web at sites including ncbi.nlm.nili.gov/BLAST, ebi.ac.uk/Tools/msa/tcoffee/, ebi.ac.uk/Tools/msa/muscle/, mafft.cbrc.jp/alignment/software/. See, e.g., Altschul et al. (1990), J. Mol. Bioi. 215:403-10. Sequence alignments standard in the art are used according to the invention to determine amino acid residues in a Cas9 ortholog that "correspond to" amino acid residues in another Cas9 ortholog. The amino acid residues of Cas9 orthologs that correspond to amino acid residues of other Cas9 orthologs appear at the same position in alignments of the sequences.

[0067] A DNA sequence that "encodes" a particular RNA is a DNA nucleic acid sequence that is transcribed into RNA. A DNA polynucleotide may encode an RNA (mRNA) that is translated into protein, or a DNA polynucleotide may encode an RNA that is not translated into protein (e.g. tRNA, rRNA, or a guide RNA; also called "non-coding" RNA or "ncRNA"). A "protein coding sequence" or a sequence that encodes a particular protein or polypeptide, is a nucleic acid sequence that is transcribed into mRNA (in the case of DNA) and is translated (in the case of mRNA) into a polypeptide in vitro or in vivo when placed under the control of appropriate regulatory sequences. The boundaries of the coding sequence are determined by a start codon at the 5' terminus (N-terminus) and a translation stop nonsense codon at the 3' terminus (C-terminus). A coding sequence can include, but is not limited to, cDNA from prokaryotic or eukaryotic mRNA, genomic DNA sequences from prokaryotic or eukaryotic DNA, and synthetic nucleic acids. A transcription termination sequence will usually be located 3' to the coding sequence.

[0068] As used herein, a "promoter sequence" is a DNA regulatory region capable of binding RNA polymerase and initiating transcription of a downstream (3' direction) coding or non-coding sequence. For purposes of defining the present invention, the promoter sequence is bounded at its 3' terminus by the transcription initiation site and extends upstream (5' direction) to include the minimum number of bases or elements necessary to initiate transcription at levels detectable above background. Within the promoter sequence will be found a transcription initiation site, as well as protein binding domains responsible for the binding of RNA polymerase. Eukaryotic promoters will often, but not always, contain "TATA" boxes and "CAT" boxes. Various promoters, including inducible promoters, may be used to drive the various vectors of the present invention.

[0069] A promoter can be a constitutively active promoter (i.e., a promoter that is constitutively in an active/"ON" state), it may be an inducible promoter (i.e., a promoter whose state, active/"ON" or inactive/"OFF", is controlled by an external stimulus, e.g., the presence of a particular temperature, compound, or protein.), it may be a spatially restricted promoter (i.e., transcriptional control element, enhancer, etc.)(e.g., tissue specific promoter, cell type specific promoter, etc.), and it may be a temporally restricted promoter (i.e., the promoter is in the "ON" state or "OFF" state during specific stages of embryonic development or during specific stages of a biological process, e.g., hair follicle cycle in mice).

[0070] Suitable promoters can be derived from viruses and can therefore be referred to as viral promoters, or they can be derived from any organism, including prokaryotic or eukaryotic organisms. Suitable promoters can be used to drive expression by any RNA polymerase (e.g., pol I, pol II, pol III). Exemplary promoters include, but are not limited to the SV40 early promoter, mouse mammary tumor virus long terminal repeat (LTR) promoter; adenovirus major late promoter (Ad MLP); a herpes simplex virus (HSV) promoter, a cytomegalovirus (CMV) promoter such as the CMV immediate early promoter region (CMVIE), a rous sarcoma virus (RSV) promoter, a human U6 small nuclear promoter (U6) (Miyagishi et al. , Nature Biotechnology 20, 497 - 500 (2002)), an enhanced U6 promoter (e.g., Xia et al., Nucleic Acids Res. 2003 Sep 1;31(17)), a human H1 promoter (H1), and the like.

[0071] Examples of inducible promoters include, but are not limited toT7 RNA polymerase promoter, T3 RNA polymerase promoter, Isopropyl-beta-D-thiogalactopyranoside (IPTG)-regulated promoter, lactose induced promoter, heat shock promoter, Tetracycline-regulated promoter, Steroid-regulated promoter, Metal-regulated promoter, estrogen receptor-regulated promoter, etc. Inducible promoters can therefore be regulated by molecules including, but not limited to, doxycycline; RNA polymerase, e.g., T7 RNA polymerase; an estrogen receptor; an estrogen receptor fusion; etc.

[0072] In some embodiments, the promoter is a spatially restricted promoter (i.e., cell type specific promoter, tissue specific promoter, etc.) such that in a multi-cellular organism, the promoter is active (i.e., "ON") in a subset of specific cells. Spatially restricted promoters may also be referred to as enhancers, transcriptional control elements, control sequences, etc. Any convenient spatially restricted promoter may be used and the choice of suitable promoter (e.g., a brain specific promoter, a promoter that drives expression in a subset of neurons, a promoter that drives expression in the germline, a promoter that drives expression in the lungs, a promoter that drives expression in muscles, a promoter that drives expression in islet cells of the pancreas, etc.) will depend on the organism. For example, various spatially restricted promoters are known for plants, flies, worms, mammals, mice, etc. Thus, a spatially restricted promoter can be used to regulate the expression of a nucleic acid encoding a site-directed modifying polypeptide in a wide variety of different tissues and cell types, depending on the organism. Some spatially restricted promoters are also temporally restricted such that the promoter is in the "ON" state or "OFF" state during specific stages of embryonic development or during specific stages of a biological process (e.g., hair follicle cycle in mice).

[0073] For illustration purposes, examples of spatially restricted promoters include, but are not limited to, neuron-specific promoters, adipocyte-specific promoters, cardiomyocyte-specific promoters, smooth muscle-specific promoters, photoreceptor-specific promoters, etc. Neuron-specific spatially restricted promoters include, but are not limited to, a neuron-specific enolase (NSE) promoter (see, e.g., EMBL HSENO2, X51956); an aromatic amino acid decarboxylase (AADC) promoter; a neurofilament promoter (see, e.g., GenBank HUMNFL, L04147); a synapsin promoter (see, e.g., GenBank HUMSYNIB, M55301); a thy-1 promoter (see, e.g., Chen et al. (1987) Cell 51:7-19; and Llewellyn, et al. (2010) Nat. Med.

16(10):1161-1166); a serotonin receptor promoter (see, e.g., GenBank S62283); a tyrosine hydroxylase promoter (TH) (see, e.g., Oh et al. (2009) Gene Ther 16:437; Sasaoka et al. (1992) Mol. Brain Res. 16:274; Boundy et al. (1998) J. Neurosci. 18:9989; and Kaneda et al. (1991) Neuron 6:583-594); a GnRH promoter (see, e.g., Radovick et al. (1991) Proc. Natl. Acad. Sci. USA 88:3402-3406); an L7 promoter (see, e.g., Oberdick et al. (1990) Science 248:223-226); a DNMT promoter (see, e.g., Bartge et al. (1988) Proc. Natl. Acad. Sci. USA 85:3648-3652); an enkephalin promoter (see, e.g., Comb et al. (1988) EMBO J. 17:3793-3805); a myelin basic protein (MBP) promoter; a Ca2+-calmodulin-dependent protein kinase II-alpha (CamKIM) promoter (see, e.g., Mayford et al. (1996) Proc. Natl. Acad. Sci. USA 93:13250; and Casanova et al. (2001) Genesis 31:37); a CMV enhancer/platelet-derived growth factor-p promoter (see, e.g., Liu et al. (2004) Gene Therapy 11:52-60); and the like.

[0074] Adipocyte-specific spatially restricted promoters include, but are not limited to aP2 gene promoter/enhancer, e.g., a region from -5.4 kb to +21 bp of a human aP2 gene (see, e.g., Tozzo et al. (1997) Endocrinol. 138:1604; Ross et al. (1990) Proc. Natl. Acad. Sci. USA 87:9590; and Pavjani et al. (2005) Nat. Med. 11:797); a glucose transporter-4 (GLUT4) promoter (see, e.g., Knight et al. (2003) Proc. Natl. Acad. Sci. USA 100:14725); a fatty acid translocase (FAT/CD36) promoter (see, e.g., Kuriki et al. (2002) Biol. Pharm. Bull. 25:1476; and Sato et al. (2002) J. Biol. Chem. 277:15703); a stearoyl-CoA desaturase-1 (SCD1) promoter (Tabor et al. (1999) J. Biol. Chem. 274:20603); a leptin promoter (see, e.g., Mason et al. (1998) Endocrinol. 139:1013; and Chen et al. (1999) Biochem. Biophys. Res. Comm. 262:187); an adiponectin promoter (see, e.g., Kita et al. (2005) Biochem. Biophys. Res. Comm. 331:484; and Chakrabarti (2010) Endocrinol. 151:2408); an adipsin promoter (see, e.g., Platt et al. (1989) Proc. Natl. Acad. Sci. USA 86:7490); a resistin promoter (see, e.g., Seo et al. (2003) Molec. Endocrinol. 17:1522); and the like.

[0075] Cardiomyocyte-specific spatially restricted promoters include, but are not limited to control sequences derived from the following genes: myosin light chain-2, a-myosin heavy chain, AE3, cardiac troponin C, cardiac actin, and the like. Franz et al. (1997) Cardiovasc. Res. 35:560-566; Robbins et al. (1995) Ann. N.Y. Acad. Sci. 752:492-505; Linn et al. (1995) Circ. Res. 76:584591; Parmacek et al. (1994) Mol. Cell. Biol. 14:1870-1885; Hunter et al. (1993) Hypertension 22:608-617; and Sartorelli et al. (1992) Proc. Natl. Acad. Sci. USA 89:4047-4051.

[0076] Smooth muscle-specific spatially restricted promoters include, but are not limited to an SM22a promoter (see, e.g., Akyiirek et al. (2000) Mol. Med. 6:983; and U.S. Patent No. 7,169,874); a smoothelin promoter (see, e.g., WO 2001/018048); an a-smooth muscle actin promoter; and the like. For example, a 0.4 kb region of the SM22a promoter, within which lie two CArG elements, has been shown to mediate vascular smooth muscle cell-specific expression (see, e.g., Kim, et al. (1997) Mol. Cell. Biol. 17, 2266-2278; Li, et al., (1996) J. Cell Biol. 132, 849-859; and Moessler, et al. (1996) Development 122, 2415-2425).

[0077] Photoreceptor-specific spatially restricted promoters include, but are not limited to, a rhodopsin promoter; a rhodopsin kinase promoter (Young et al. (2003) Ophthalmol. Vis. Sci. 44:4076); a beta phosphodiesterase gene promoter (Nicoud et al. (2007) J. Gene Med. 9:1015); a retinitis pigmentosa gene promoter (Nicoud et al. (2007) supra); an interphotoreceptor retinoid-binding protein (IRBP) gene enhancer (Nicoud et al. (2007) supra); an IRBP gene promoter (Yokoyama et al. (1992) Exp Eye Res. 55:225); and the like.

[0078] The terms "DNA regulatory sequences," "control elements," and "regulatory elements," used interchangeably herein, refer to transcriptional and translational control sequences, such as promoters, enhancers, polyadenylation signals, terminators, protein degradation signals, and the like, that provide for and/or regulate transcription of a non-coding sequence (e.g., guide RNA) or a coding sequence (e.g., site-directed modifying polypeptide, or Cas9 polypeptide) and/or regulate translation of an encoded polypeptide.

[0079] IThe term "naturally-occurring" or "unmodified" as used herein as applied to a nucleic acid, a polypeptide, a cell, or an organism, refers to a nucleic acid, polypeptide, cell, or organism that is found in nature. For example, a polypeptide or polynucleotide sequence that is present in an organism (including viruses) that can be isolated from a source in nature and which has not been intentionally modified by a human in the laboratory is naturally occurring.

[0080] The term "chimeric" as used herein as applied to a nucleic acid or polypeptide refers to two components that are defined by structures derived from different sources. For example, where "chimeric" is used in the context of a chimeric polypeptide (e.g., a chimeric Cas9 protein), the chimeric polypeptide includes amino acid sequences that are derived from different polypeptides. A chimeric polypeptide may comprise either modified or naturally-occurring polypeptide sequences (e.g., a first amino acid sequence from a modified or unmodified Cas9 protein; and a second amino acid sequence other than the Cas9 protein). Similarly, "chimeric" in the context of a polynucleotide encoding a chimeric polypeptide includes nucleotide sequences derived from different coding regions (e.g., a first nucleotide sequence encoding a modified or unmodified Cas9 protein; and a second nucleotide sequence encoding a polypeptide other than a Cas9 protein).

[0081] The term "chimeric polypeptide" refers to a polypeptide which is not naturally occurring, e.g., is made by the artificial combination (i.e., "fusion") of two otherwise separated segments of amino sequence through human intervention. A polypeptide that comprises a chimeric amino acid sequence is a chimeric polypeptide. Some chimeric polypeptides can be referred to as "fusion variants."

[0082] "Heterologous," as used herein, means a nucleotide or peptide that is not found in the native nucleic acid or protein, respectively. For example, in a chimeric Cas9 protein, the RNA-binding domain of a naturally-occurring bacterial Cas9 polypeptide (or a variant thereof) may be fused to a heterologous

polypeptide sequence (i.e. a polypeptide sequence from a protein other than Cas9 or a polypeptide sequence from another organism). The heterologous polypeptide may exhibit an activity (e.g., enzymatic activity) that will also be exhibited by the chimeric Cas9 protein (e.g., methyltransferase activity, acetyltransferase activity, kinase activity, ubiquitinating activity, etc.). A heterologous nucleic acid may be linked to a naturally-occurring nucleic acid (or a variant thereof) (e.g., by genetic engineering) to generate a chimeric polynucleotide encoding a chimeric polypeptide. As another example, in a fusion variant Cas9 site-directed polypeptide, a variant Cas9 site-directed polypeptide may be fused to a heterologous polypeptide (i.e. a polypeptide other than Cas9), which exhibits an activity that will also be exhibited by the fusion variant Cas9 site-directed polypeptide. A heterologous nucleic acid may be linked to a variant Cas9 site-directed polypeptide (e.g., by genetic engineering) to generate a polynucleotide encoding a fusion variant Cas9 site-directed polypeptide. "Heterologous," as used herein, additionally means a nucleotide or polypeptide in a cell that is not its native cell.

[0083] The term “cognate” refers to two biomolecules that normally interact or co-exist in nature.

[0084] "Recombinant," as used herein, means that a particular nucleic acid (DNA or RNA) or vector is the product of various combinations of cloning, restriction, polymerase chain reaction (PCR) and/or ligation steps resulting in a construct having a structural coding or non-coding sequence distinguishable from endogenous nucleic acids found in natural systems. DNA sequences encoding polypeptides can be assembled from cDNA fragments or from a series of synthetic oligonucleotides, to provide a synthetic nucleic acid which is capable of being expressed from a recombinant transcriptional unit contained in a cell or in a cell-free transcription and translation system. Genomic DNA comprising the relevant sequences can also be used in the formation of a recombinant gene or transcriptional unit. Sequences of non-translated DNA may be present 5' or 3' from the open reading frame, where such sequences do not interfere with manipulation or expression of the coding regions, and may indeed act to modulate production of a desired product by various mechanisms (see "DNA regulatory sequences", below). Alternatively, DNA sequences encoding RNA (e.g., guide RNA) that is not translated may also be considered recombinant. Thus, e.g., the term "recombinant" nucleic acid refers to one which is not naturally occurring, e.g., is made by the artificial combination of two otherwise separated segments of sequence through human intervention. This artificial combination is often accomplished by either chemical synthesis means, or by the artificial manipulation of isolated segments of nucleic acids, e.g., by genetic engineering techniques. Such is usually done to replace a codon with a codon encoding the same amino acid, a conservative amino acid, or a non-conservative amino acid. Alternatively, it is performed to join together nucleic acid segments of desired functions to generate a desired combination of functions. This artificial combination is often accomplished by either chemical synthesis means, or by the artificial manipulation of isolated segments of nucleic acids, e.g., by genetic engineering techniques. When a recombinant polynucleotide encodes a polypeptide, the sequence of the encoded polypeptide can be naturally occurring ("wild type") or can be a variant (e.g., a mutant) of the naturally occurring sequence.

Thus, the term "recombinant" polypeptide does not necessarily refer to a polypeptide whose sequence does not naturally occur. Instead, a "recombinant" polypeptide is encoded by a recombinant DNA sequence, but the sequence of the polypeptide can be naturally occurring ("wild type") or non-naturally occurring (e.g., a variant, a mutant, etc.). Thus, a "recombinant" polypeptide is the result of human intervention, but may be a naturally occurring amino acid sequence.

[0085] A "vector" or "expression vector" is a replicon, such as plasmid, phage, virus, or cosmid, to which another DNA segment, i.e. an "insert", may be attached so as to bring about the replication of the attached segment in a cell.

[0086] An "expression cassette" comprises a DNA coding sequence operably linked to a promoter. "Operably linked" refers to a juxtaposition wherein the components so described are in a relationship permitting them to function in their intended manner. For instance, a promoter is operably linked to a coding sequence if the promoter affects its transcription or expression. The terms "recombinant expression vector," or "DNA construct" are used interchangeably herein to refer to a DNA molecule comprising a vector and at least one insert. Recombinant expression vectors are usually generated for the purpose of expressing and/or propagating the insert(s), or for the construction of other recombinant nucleotide sequences. The nucleic acid(s) may or may not be operably linked to a promoter sequence and may or may not be operably linked to DNA regulatory sequences.

[0087] A cell has been "genetically modified" or "transformed" or "transfected" by exogenous DNA, e.g. a recombinant expression vector, when such DNA has been introduced inside the cell. The presence of the exogenous DNA results in permanent or transient genetic change. The transforming DNA may or may not be integrated (covalently linked) into the genome of the cell.

[0088] In prokaryotes, yeast, and mammalian cells for example, the transforming DNA may be maintained on an episomal element such as a plasmid. With respect to eukaryotic cells, a stably transformed cell is one in which the transforming DNA has become integrated into a chromosome so that it is inherited by daughter cells through chromosome replication. This stability is demonstrated by the ability of the eukaryotic cell to establish cell lines or clones that comprise a population of daughter cells containing the transforming DNA. A "clone" is a population of cells derived from a single cell or common ancestor by mitosis. A "cell line" is a clone of a primary cell that is capable of stable growth *in vitro* for many generations.

[0089] Suitable methods of genetic modification (also referred to as "transformation") include e.g., viral or bacteriophage infection, transfection, conjugation, protoplast fusion, lipofection, electroporation, calcium phosphate precipitation, polyethyleneimine (PEI)-mediated transfection, DEAE-dextran mediated transfection, liposome-mediated transfection, particle gun technology, calcium phosphate precipitation, direct micro injection, nanoparticle-mediated nucleic acid delivery (see, e.g., Panyam et., al Adv Drug Deliv Rev. 2012 Sep 13. pii: S0169-409X(12)00283-9. doi: 10.1016/j.addr.2012.09.023), and the like.

[0090] The choice of method of genetic modification is generally dependent on the type of cell being transformed and the circumstances under which the transformation is taking place (e.g., in vitro, ex vivo, or in vivo). A general discussion of these methods can be found in Ausubel, et al., Short Protocols in Molecular Biology, 3rd ed., Wiley & Sons, 1995.

[0091] A "host cell," as used herein, denotes an in vivo or in vitro eukaryotic cell, a prokaryotic cell (e.g., bacterial or archaeal cell), or a cell from a multicellular organism (e.g., a cell line) cultured as a unicellular entity, which eukaryotic or prokaryotic cells can be, or have been, used as recipients for a nucleic acid, and include the progeny of the original cell which has been transformed by the nucleic acid. It is understood that the progeny of a single cell may not necessarily be completely identical in morphology or in genomic or total DNA complement as the original parent, due to natural, accidental, or deliberate mutation. A "recombinant host cell" (also referred to as a "genetically modified host cell") is a host cell into which has been introduced a heterologous nucleic acid, e.g., an expression vector. For example, a bacterial host cell is a genetically modified bacterial host cell by virtue of introduction into a suitable bacterial host cell of an exogenous nucleic acid (e.g., a plasmid or recombinant expression vector) and a eukaryotic host cell is a genetically modified eukaryotic host cell (e.g., a mammalian germ cell), by virtue of introduction into a suitable eukaryotic host cell of an exogenous nucleic acid.

[0092] A "target DNA" as used herein is a DNA polynucleotide that comprises a "target site" or "target sequence." The terms "target site," "target sequence," "target protospacer DNA, " or "protospacer-like sequence" are used interchangeably herein to refer to a nucleic acid sequence present in a target DNA to which a DNA-targeting segment of a guide RNA will bind, provided sufficient conditions for binding exist. For example, the target site (or target sequence) 5'-GAGCATATC-3' within a target DNA is targeted by (or is bound by, or hybridizes with, or is complementary to) the RNA sequence 5'-GAUAUGCUC-3'. Suitable DNA/RNA binding conditions include physiological conditions normally present in a cell. Other suitable DNA/RNA binding conditions (e.g., conditions in a cell-free system) are known in the art; see, e.g., Sambrook, supra. The strand of the target DNA that is complementary to and hybridizes with the guide RNA is referred to as the "complementary strand" and the strand of the target DNA that is complementary to the "complementary strand" (and is therefore not complementary to the guide RNA) is referred to as the "noncomplementary strand" or "non-complementary strand." By "site-directed modifying polypeptide" or "RNA-binding site-directed polypeptide" or "RNA-binding site-directed modifying polypeptide" or "site-directed polypeptide" it is meant a polypeptide that binds RNA and is targeted to a specific DNA sequence. A site-directed modifying polypeptide as described herein is targeted to a specific DNA sequence by the RNA molecule to which it is bound. The RNA molecule comprises a sequence that binds, hybridizes to, or is complementary to a target sequence within the target DNA, thus targeting the bound polypeptide to a specific location within the target DNA (the target sequence). Exemplary target sequences of the invention are set out in SEQ ID NOs: 801-2701. SEQ ID NOs: 801-973 are protospacer-like target sequences 5' to the PAM sequence NNNNACA in the human CCR5 gene. SEQ

ID NOs: 974-1078 are protospacer-like sequences 5' to the PAM sequence GNNNCNNA in the human CCR5 gene. SEQ ID NOs: 1079-1222 are protospacer-like target sequences 5' to the PAM sequence NNNNACA in the exons of the human CCR5 gene. SEQ ID NOs: 1223-1312 are protospacer-like sequences 5' to the PAM sequence GNNNCNNA in the exons of the human CCR5 gene. SEQ ID NOs: 1313-1348 are protospacer-like target sequences 5' to the PAM sequence NNNNACA around the 5' end of the human CCR5 gene. SEQ ID NOs: 1349-1371 are protospacer-like sequences 5' to the PAM sequence GNNNCNNA around the 5' end of the human CCR5 gene. SEQ ID NOs: 1372-1415 are protospacer-like target sequences 5' to the PAM sequence NNNNACA around the delta 32 locus in the human CCR5 gene. SEQ ID NOs: 1416-1443 are protospacer-like sequences 5' to the PAM sequence GNNNCNNA around the delta 32 locus in the human CCR5 gene. SEQ ID NOs: 1444-1900 are protospacer-like target sequences 5' to the PAM sequence NNNNACA in the human BCL11A gene. SEQ ID NOs: 1901-2162 are protospacer-like sequences 5' to the PAM sequence GNNNCNNA in the human BCL11A gene. SEQ ID NOs: 2163-2482 are protospacer-like target sequences 5' to the PAM sequence NNNNACA in the exons of the human BCL11A gene. SEQ ID NOs: 2483-2666 are protospacer-like sequences 5' to the PAM sequence GNNNCNNA in the exons of the human BCL11A gene. SEQ ID NOs: 2667-2686 are protospacer-like target sequences 5' to the PAM sequence NNNNACA around the 5' end of the human BCL11A gene. SEQ ID NOs: 2687-2701 are protospacer-like sequences 5' to the PAM sequence GNNNCNNA around the 5' end of the human BCL11A gene. Target sequences at least 80% identical to the sequences set out in SEQ ID NOs: 801-2701 are also contemplated.

[0093] By "cleavage" it is meant the breakage of the covalent backbone of a DNA molecule. Cleavage can be initiated by a variety of methods including, but not limited to, enzymatic or chemical hydrolysis of a phosphodiester bond. Both single-stranded cleavage and double-stranded cleavage are possible, and double-stranded cleavage can occur as a result of two distinct single-stranded cleavage events. DNA cleavage can result in the production of either blunt ends or staggered ends. In certain embodiments, a complex comprising a guide RNA and a site-directed modifying polypeptide is used for targeted double-stranded DNA cleavage.

[0094] "Nuclease" and "endonuclease" are used interchangeably herein to mean an enzyme which possesses endonucleolytic catalytic activity for DNA cleavage.

[0095] By "cleavage domain" or "active domain" or "nuclease domain" of a nuclease it is meant the polypeptide sequence or domain within the nuclease which possesses the catalytic activity for DNA cleavage. A cleavage domain can be contained in a single polypeptide chain or cleavage activity can result from the association of two (or more) polypeptides. A single nuclease domain may consist of more than one isolated stretch of amino acids within a given polypeptide.

[0096] By "site-directed polypeptide" or "RNA-binding site-directed polypeptide" or "RNA-binding site-directed polypeptide" it is meant a polypeptide that binds RNA and is targeted to a specific DNA

sequence. A site-directed polypeptide as described herein is targeted to a specific DNA sequence by the RNA molecule to which it is bound. The RNA molecule comprises a sequence that is complementary to a target sequence within the target DNA, thus targeting the bound polypeptide to a specific location within the target DNA (the target sequence).

[0097] The RNA molecule that binds to the site-directed modifying polypeptide and targets the polypeptide to a specific location within the target DNA is referred to herein as the "guide RNA" or "guide RNA polynucleotide" (also referred to herein as a "guide RNA" or "gRNA"). A guide RNA comprises two segments, a "DNA-targeting segment" and a "protein-binding segment." By "segment" it is meant a segment/section/region of a molecule, e.g., a contiguous stretch of nucleotides in an RNA. A segment can also mean a region/section of a complex such that a segment may comprise regions of more than one molecule. For example, in some cases the protein-binding segment (described below) of a guide RNA is one RNA molecule and the protein-binding segment therefore comprises a region of that RNA molecule. In other cases, the protein-binding segment (described below) of a guide RNA comprises two separate molecules that are hybridized along a region of complementarity. As an illustrative, non-limiting example, a protein-binding segment of a guide RNA that comprises two separate molecules can comprise (i) base pairs 40-75 of a first RNA molecule that is 100 base pairs in length; and (ii) base pairs 10-25 of a second RNA molecule that is 50 base pairs in length. The definition of "segment," unless otherwise specifically defined in a particular context, is not limited to a specific number of total base pairs, is not limited to any particular number of base pairs from a given RNA molecule, is not limited to a particular number of separate molecules within a complex, and may include regions of RNA molecules that are of any total length and may or may not include regions with complementarity to other molecules.

[0098] The DNA-targeting segment (or "DNA-targeting sequence") comprises a nucleotide sequence that is complementary to a specific sequence within a target DNA (the complementary strand of the target DNA) designated the “protospacer-like” sequence herein. The protein-binding segment (or "protein-binding sequence") interacts with a site-directed modifying polypeptide. When the site-directed modifying polypeptide is a Cas9 or Cas9 related polypeptide (described in more detail below), site-specific cleavage of the target DNA occurs at locations determined by both (i) base-pairing complementarity between the guide RNA and the target DNA; and (ii) a short motif (referred to as the protospacer adjacent motif (PAM)) in the target DNA.

[0099] The protein-binding segment of a guide RNA comprises, in part, two complementary stretches of nucleotides that hybridize to one another to form a double stranded RNA duplex (dsRNA duplex).

[00100] In some embodiments, a nucleic acid (e.g., a guide RNA, a nucleic acid comprising a nucleotide sequence encoding a guide RNA; a nucleic acid encoding a site-directed polypeptide; etc.) comprises a modification or sequence that provides for an additional desirable feature (e.g., modified or regulated stability; subcellular targeting; tracking, e.g., a fluorescent label; a binding site for a protein or

protein complex; etc.). Non-limiting examples include: a 5' cap (e.g., a 7-methylguanylate cap (m7G)); a 3' polyadenylated tail (i.e., a 3' poly(A) tail); a riboswitch sequence (e.g., to allow for regulated stability and/or regulated accessibility by proteins and/or protein complexes); a stability control sequence; a sequence that forms a dsRNA duplex (i.e., a hairpin)); a modification or sequence that targets the RNA to a subcellular location (e.g., nucleus, mitochondria, chloroplasts, and the like); a modification or sequence that provides for tracking (e.g., direct conjugation to a fluorescent molecule, conjugation to a moiety that facilitates fluorescent detection, a sequence that allows for fluorescent detection, etc.); a modification or sequence that provides a binding site for proteins (e.g., proteins that act on DNA, including transcriptional activators, transcriptional repressors, DNA methyltransferases, DNA demethylases, histone acetyltransferases, histone deacetylases, and the like); and combinations thereof.

[00101] In some embodiments, a guide RNA comprises an additional segment at either the 5' or 3' end that provides for any of the features described above. For example, a suitable third segment can comprise a 5' cap (e.g., a 7-methylguanylate cap (m7G)); a 3' polyadenylated tail (i.e., a 3' poly(A) tail); a riboswitch sequence (e.g., to allow for regulated stability and/or regulated accessibility by proteins and protein complexes); a stability control sequence; a sequence that forms a dsRNA duplex (i.e., a hairpin)); a sequence that targets the RNA to a subcellular location (e.g., nucleus, mitochondria, chloroplasts, and the like); a modification or sequence that provides for tracking (e.g., direct conjugation to a fluorescent molecule, conjugation to a moiety that facilitates fluorescent detection, a sequence that allows for fluorescent detection, etc.); a modification or sequence that provides a binding site for proteins (e.g., proteins that act on DNA, including transcriptional activators, transcriptional repressors, DNA methyltransferases, DNA demethylases, histone acetyltransferases, histone deacetylases, and the like); and combinations thereof.

[00102] A guide RNA and a site-directed modifying polypeptide (i.e., site-directed polypeptide) form a complex (i.e., bind via non-covalent interactions). The guide RNA provides target specificity to the complex by comprising a nucleotide sequence that is complementary to a sequence of a target DNA. The site-directed modifying polypeptide of the complex provides the site-specific activity. In other words, the site-directed modifying polypeptide is guided to a target DNA sequence (e.g. a target sequence in a chromosomal nucleic acid; a target sequence in an extrachromosomal nucleic acid, e.g. an episomal nucleic acid, a minicircle, etc.; a target sequence in a mitochondrial nucleic acid; a target sequence in a chloroplast nucleic acid; a target sequence in a plasmid; etc.) by virtue of its association with the protein-binding segment of the guide RNA.

[00103] In some embodiments, a guide RNA comprises two separate RNA molecules (RNA polynucleotides: an "activator-RNA" and a "targeter-RNA", see below) and is referred to herein as a "double-molecule guide RNA" or a "two-molecule guide RNA." In other embodiments, the guide RNA is a single RNA molecule (single RNA polynucleotide) and is referred to herein as a "single-molecule guide

RNA," a "single-guide RNA," or an "sgRNA." The term "guide RNA" or "gRNA" is inclusive, referring both to double-molecule guide RNAs and to single-molecule guide RNAs (i.e., sgRNAs).

[00104] A two-molecule guide RNA comprises two separate RNA molecules (a "targeter-RNA" and an "activator-RNA"). Each of the two RNA molecules of a two-molecule guide RNA comprises a stretch of nucleotides that are complementary to one another such that the complementary nucleotides of the two RNA molecules hybridize to form the double stranded RNA duplex of the protein-binding segment.

[00105] An exemplary two-molecule guide RNA comprises a crRNA-like ("CRISPR RNA" or "targeter-RNA") molecule (which includes a CRISPR repeat or CRISPR repeat-like sequence) and a corresponding tracrRNA-like ("trans-activating CRISPR RNA" or "activator-RNA" or "tracrRNA") molecule. A crRNA-like molecule (targeter-RNA) comprises both the DNA-targeting segment (single stranded) of the guide RNA and a stretch ("duplex-forming segment") of nucleotides that forms one half of the dsRNA duplex of the protein-binding segment of the guide RNA. A corresponding tracrRNA-like molecule (activator-RNA) comprises a stretch of nucleotides (duplex-forming segment) that forms the other half of the dsRNA duplex of the protein-binding segment of the guide RNA. In other words, a stretch of nucleotides of a crRNA-like molecule are complementary to and hybridize with a stretch of nucleotides of a tracrRNA-like molecule to form the dsRNA duplex of the protein-binding domain of the guide RNA. As such, each crRNA-like molecule can be said to have a corresponding tracrRNA-like molecule. The crRNA-like molecule additionally provides the single stranded DNA-targeting segment. Thus, a crRNA-like and a tracrRNA-like molecule (as a corresponding pair) hybridize to form a guide RNA. A double-molecule guide RNA can comprise any corresponding crRNA and tracrRNA pair.

[00106] A two-molecule guide RNA can be designed to allow for controlled (i.e., conditional) binding of a targeter-RNA with an activator-RNA. Because a two-molecule guide RNA is not functional unless both the activator-RNA and the targeter-RNA are bound in a functional complex with Cas9, a two-molecule guide RNA can be inducible (e.g., drug inducible) by rendering the binding between the activator-RNA and the targeter-RNA to be inducible. As one non-limiting example, RNA aptamers can be used to regulate (i.e., control) the binding of the activator-RNA with the targeter-RNA. Accordingly, the activator-RNA and/or the targeter-RNA can comprise an RNA aptamer sequence.

[00107] A single-molecule guide RNA comprises two stretches of nucleotides (a targeter-RNA and an activator-RNA) that are complementary to one another, are covalently linked (directly, or by intervening nucleotides), and hybridize to form the double stranded RNA duplex (dsRNA duplex) of the protein-binding segment, thus resulting in a stem-loop structure. The targeter-RNA and the activator-RNA can be covalently linked via the 3' end of the targeter-RNA and the 5' end of the activator-RNA. Alternatively, targeter-RNA and the activator-RNA can be covalently linked via the 5' end of the targeter-RNA and the 3' end of the activator-RNA.

[00108] An exemplary single-molecule guide RNA comprises two complementary stretches of nucleotides that hybridize to form a dsRNA duplex. In some embodiments, one of the two complementary stretches of nucleotides of the single-molecule guide RNA (or the DNA encoding the stretch) is at least about 60% identical to one of the activator-RNA (tracrRNA) sequences set forth in Supplementary Table S5 over a stretch of at least 8 contiguous nucleotides. For example, one of the two complementary stretches of nucleotides of the single-molecule guide RNA (or the DNA encoding the stretch) is at least about 65% identical, at least about 70% identical, at least about 75% identical, at least about 80% identical, at least about 85% identical, at least about 90% identical, at least about 95% identical, at least about 98% identical, at least about 99% identical or 100 % identical to one of the tracrRNA sequences set forth in Supplementary Table S5 over a stretch of at least 8 contiguous , at least 9 contiguous, at least 10 contiguous, at least 11 contiguous, at least 12 contiguous, at least 13 contiguous, at least 14 contiguous or at least 15 contiguous nucleotides. For example, the single-molecule guide RNA may comprise a nucleotide sequence that is at least 70% identical over at least 10 contiguous nucleotides, at least 80% identical over at least 10 contiguous nucleotides, at least 70% identical over at least 11 contiguous nucleotides, at least 80% identical over at least 11 contiguous nucleotides, at least 70% identical over at least 12 contiguous nucleotides, or at least 80% identical over at least 12 contiguous nucleotides of one of the tracrRNA sequences set forth in Supplementary Table S5. It is understood that where a series of percent identities and a series of lengths of nucleotides sequences are set out as options, each and every combination of a percent identity with a length (e.g. 8, 9, 10, 12, 13, 14, 15 nucleotides) of nucleotide sequence is contemplated.

[00109] In some embodiments, one of the two complementary stretches of nucleotides of the single-molecule guide RNA (or the DNA encoding the stretch) is at least about 60% identical to one of the targeter-RNA (crRNA/CRISPR repeat) sequences set forth in Supplementary Table S5 over a stretch of at least 8 contiguous nucleotides. For example, one of the two complementary stretches of nucleotides of the single-molecule guide RNA (or the DNA encoding the stretch) is at least about 65% identical, at least about 70% identical, at least about 75% identical, at least about 80% identical, at least about 85% identical, at least about 90% identical, at least about 95% identical, at least about 98% identical, at least about 99% identical or 100 % identical to one of the crRNA/CRISPR repeat sequences set forth in Supplementary Table S5 over a stretch of at least 8 contiguous, at least 9 contiguous, at least 10 contiguous, at least 11 contiguous, at least 12 contiguous, at least 13 contiguous, at least 14 contiguous or at least 15 contiguous nucleotides. For example, the single-molecule guide RNA may comprise a nucleotide sequence that is at least 70% identical over at least 10 contiguous nucleotides, at least 80% identical over at least 10 contiguous nucleotides, at least 70% identical over at least 11 contiguous nucleotides, at least 80% identical over at least 11 contiguous nucleotides, at least 70% identical over at least 12 contiguous nucleotides, or at least 80% identical over at least 12 contiguous nucleotides of one of the CRISPR repeat sequences set forth in Supplementary Table S5. It is understood that where a series of percent identities and a series of lengths of nucleotides sequences are set out as options, each

and every combination of a percent identity with a length (e.g. 8, 9, 10, 11, 12, 13, 14, 15 nucleotides) of nucleotide sequence is contemplated.

[00110] The term "activator-RNA" is used herein to mean a tracrRNA-like molecule of a double-molecule guide RNA. The term "targeter-RNA" is used herein to mean a crRNA-like molecule of a double-molecule guide RNA. The term "duplex-forming segment" is used herein to mean the stretch of nucleotides of an activator-RNA or a targeter-RNA that contributes to the formation of the dsRNA duplex by hybridizing to a stretch of nucleotides of a corresponding activator-RNA or targeter-RNA molecule. In other words, an activator-RNA comprises a duplex-forming segment that is complementary to the duplex-forming segment of the corresponding targeter-RNA. As such, an activator-RNA comprises a duplex-forming segment while a targeter-RNA comprises both a duplex-forming segment and the DNA-targeting segment of the guide RNA. Therefore, a double-molecule guide RNA can be comprised of any corresponding activator-RNA and targeter-RNA pair.

[00111] RNA aptamers are known in the art and are generally a synthetic version of a riboswitch. The terms "RNA aptamer" and "riboswitch" are used interchangeably herein to encompass both synthetic and natural nucleic acid sequences that provide for inducible regulation of the structure (and therefore the availability of specific sequences) of the RNA molecule of which they are part. RNA aptamers usually comprise a sequence that folds into a particular structure (e.g., a hairpin), which specifically binds a particular drug (e.g., a small molecule). Binding of the drug causes a structural change in the folding of the RNA, which changes a feature of the nucleic acid of which the aptamer is a part. As non-limiting examples: (i) an activator-RNA with an aptamer may not be able to bind to the cognate targeter-RNA unless the aptamer is bound by the appropriate drug; (ii) a targeter-RNA with an aptamer may not be able to bind to the cognate activator-RNA unless the aptamer is bound by the appropriate drug; and (iii) a targeter-RNA and an activator-RNA, each comprising a different aptamer that binds a different drug, may not be able to bind to each other unless both drugs are present. As illustrated by these examples, a two-molecule guide RNA can be designed to be inducible.

[00112] Examples of aptamers and riboswitches can be found, for example, in: Nakamura et al., Genes Cells. 2012 May;17(5):344-64; Vavalle et al., Future Cardiol. 2012 May;8(3):371-82; Citartan et al., Biosens Bioelectron. 2012 Apr 15;34(1):1-11; and Liberman et al., Wiley Interdiscip Rev RNA. 2012 May-Jun;3(3):369-84; all of which are herein incorporated by reference in their entirety.

[00113] The term "stem cell" is used herein to refer to a cell (e.g., plant stem cell, vertebrate stem cell) that has the ability both to self-renew and to generate a differentiated cell type (see Morrison et al. (1997) Cell 88:287-298). In the context of cell ontogeny, the adjective "differentiated", or "differentiating" is a relative term. A "differentiated cell" is a cell that has progressed further down the developmental pathway than the cell it is being compared with. Thus, pluripotent stem cells (described below) can differentiate into lineage-restricted progenitor cells (e.g.,mesodermal stem cells), which in turn can differentiate into

cells that are further restricted (e.g., neuron progenitors), which can differentiate into end-stage cells (i.e., terminally differentiated cells, e.g., neurons, cardiomyocytes, etc.), which play a characteristic role in a certain tissue type, and may or may not retain the capacity to proliferate further. Stem cells may be characterized by both the presence of specific markers (e.g., proteins, RNAs, etc.) and the absence of specific markers. Stem cells may also be identified by functional assays both in vitro and in vivo, particularly assays relating to the ability of stem cells to give rise to multiple differentiated progeny.

[00114] Stem cells of interest include pluripotent stem cells (PSCs). The term "pluripotent stem cell" or "PSC" is used herein to mean a stem cell capable of producing all cell types of the organism. Therefore, a PSC can give rise to cells of all germ layers of the organism (e.g., the endoderm, mesoderm, and ectoderm of a vertebrate). Pluripotent cells are capable of forming teratomas and of contributing to ectoderm, mesoderm, or endoderm tissues in a living organism. Pluripotent stem cells of plants are capable of giving rise to all cell types of the plant (e.g., cells of the root, stem, leaves, etc.).

[00115] PSCs of animals can be derived in a number of different ways. For example, embryonic stem cells (ESCs) are derived from the inner cell mass of an embryo (Thomson et. al, Science. 1998 Nov 6;282(5391):1145-7) whereas induced pluripotent stem cells (iPSCs) are derived from somatic cells (Takahashi et. al, Cell. 2007 Nov 30;131(5):861-72; Takahashi et. al, Nat Protoc. 2007;2(12):3081-9; Yu et. al, Science. 2007 Dec 21;318(5858):1917-20. Epub 2007 Nov 20). Because the term PSC refers to pluripotent stem cells regardless of their derivation, the term PSC encompasses the terms ESC and iPSC, as well as the term embryonic germ stem cells (EGSC), which are another example of a PSC. PSCs may be in the form of an established cell line, they may be obtained directly from primary embryonic tissue, or they may be derived from a somatic cell. PSCs can be target cells of the methods described herein.

[00116] By "embryonic stem cell" (ESC) is meant a PSC that was isolated from an embryo, typically from the inner cell mass of the blastocyst. ESC lines are listed in the NIH Human Embryonic Stem Cell Registry, e.g. hESBGN-01, hESBGN-02, hESBGN-03, hESBGN-04 (BresaGen, Inc.); HES-1, HES-2, HES-3, HES-4, HES-5, HES-6 (ES Cell International); Miz-hES1 (MizMedi Hospital-Seoul National University); HSF-1, HSF-6 (University of California at San Francisco); and H1, H7, H9, H13, H14 (Wisconsin Alumni Research Foundation (WiCell Research Institute)). Stem cells of interest also include embryonic stem cells from other primates, such as Rhesus stem cells and marmoset stem cells. The stem cells may be obtained from any mammalian species, e.g. human, equine, bovine, porcine, canine, feline, rodent, e.g. mice, rats, hamster, primate, etc. (Thomson et al. (1998) Science 282:1145; Thomson et al. (1995) Proc. Natl. Acad. Sci USA 92:7844; Thomson et al. (1996) Biol. Reprod. 55:254; Shamblott et al., Proc. Natl. Acad. Sci. USA 95:13726, 1998). In culture, ESCs typically grow as flat colonies with large nucleo-cytoplasmic ratios, defined borders and prominent nucleoli. In addition, ESCs express SSEA-3, SSEA-4, TRA-1-60, TRA-1-81, and Alkaline Phosphatase, but not SSEA-1. Examples of methods of generating and characterizing ESCs may be found in, for example, US Patent No.

7,029,913, US Patent No. 5,843,780, and US Patent No. 6,200,806, the disclosures of which are incorporated herein by reference. Methods for proliferating hESCs in the undifferentiated form are described in WO 99/20741, WO 01/51616, and WO 03/020920. By "embryonic germ stem cell" (EGSC) or "embryonic germ cell" or "EG cell" is meant a PSC that is derived from germ cells and/or germ cell progenitors, e.g. primordial germ cells, i.e. those that would become sperm and eggs. Embryonic germ cells (EG cells) are thought to have properties similar to embryonic stem cells as described above. Examples of methods of generating and characterizing EG cells may be found in, for example, US Patent No. 7,153,684; Matsui, Y., et al., (1992) Cell 70:841; Shamblott, M., et al. (2001) Proc. Natl. Acad. Sci. USA 98: 113; Shamblott, M., et al. (1998) Proc. Natl. Acad. Sci. USA, 95:13726; and Koshimizu, U., et al. (1996) Development, 122:1235, the disclosures of which are incorporated herein by reference.

[00117] By "induced pluripotent stem cell" or "iPSC" it is meant a PSC that is derived from a cell that is not a PSC (i.e., from a cell this is differentiated relative to a PSC). iPSCs can be derived from multiple different cell types, including terminally differentiated cells. iPSCs have an ES cell-like morphology, growing as flat colonies with large nucleo-cytoplasmic ratios, defined borders and prominent nuclei. In addition, iPSCs express one or more key pluripotency markers known by one of ordinary skill in the art, including but not limited to Alkaline Phosphatase, SSEA3, SSEA4, Sox2, Oct3/4, Nanog, TRA160, TRA181, TDGF 1, Dnmt3b, FoxD3, GDF3, Cyp26al, TERT, and zfp42. Examples of methods of generating and characterizing iPSCs may be found in, for example, U.S. Patent Publication Nos. US20090047263, US20090068742, US20090191159, US20090227032, US20090246875, and US20090304646, the disclosures of which are incorporated herein by reference. Generally, to generate iPSCs, somatic cells are provided with reprogramming factors (e.g. Oct4, SOX2, KLF4, MYC, Nanog, Lin28, etc.) known in the art to reprogram the somatic cells to become pluripotent stem cells.

[00118] By "somatic cell" it is meant any cell in an organism that, in the absence of experimental manipulation, does not ordinarily give rise to all types of cells in an organism. In other words, somatic cells are cells that have differentiated sufficiently that they will not naturally generate cells of all three germ layers of the body, i.e. ectoderm, mesoderm and endoderm. For example, somatic cells would include both neurons and neural progenitors, the latter of which may be able to naturally give rise to all or some cell types of the central nervous system but cannot give rise to cells of the mesoderm or endoderm lineages.

[00119] By "mitotic cell" it is meant a cell undergoing mitosis. Mitosis is the process by which a eukaryotic cell separates the chromosomes in its nucleus into two identical sets in two separate nuclei. It is generally followed immediately by cytokinesis, which divides the nuclei, cytoplasm, organelles and cell membrane into two cells containing roughly equal shares of these cellular components.

[00120] By "post-mitotic cell" it is meant a cell that has exited from mitosis, i.e., it is "quiescent", i.e. it is no longer undergoing divisions. This quiescent state may be temporary, i.e. reversible, or it may be permanent.

[00121] By "meiotic cell" it is meant a cell that is undergoing meiosis. Meiosis is the process by which a cell divides its nuclear material for the purpose of producing gametes or spores. Unlike mitosis, in meiosis, the chromosomes undergo a recombination step which shuffles genetic material between chromosomes. Additionally, the outcome of meiosis is four (genetically unique) haploid cells, as compared with the two (genetically identical) diploid cells produced from mitosis.

[00122] By "recombination" it is meant a process of exchange of genetic information between two polynucleotides. As used herein, "homology-directed repair (HDR)" refers to the specialized form DNA repair that takes place, for example, during repair of double-strand breaks in cells. This process requires nucleotide sequence homology, uses a "donor" molecule to template repair of a "target" molecule (i.e., the one that experienced the double-strand break), and leads to the transfer of genetic information from the donor to the target. Homology-directed repair may result in an alteration of the sequence of the target molecule (e.g., insertion, deletion, mutation), if the donor polynucleotide differs from the target molecule and part or all of the sequence of the donor polynucleotide is incorporated into the target DNA. In some embodiments, the donor polynucleotide, a portion of the donor polynucleotide, a copy of the donor polynucleotide, or a portion of a copy of the donor polynucleotide integrates into the target DNA.

[00123] By "non-homologous end joining (NHEJ)" it is meant the repair of double-strand breaks in DNA by direct ligation of the break ends to one another without the need for a homologous template (in contrast to homology-directed repair, which requires a homologous sequence to guide repair). NHEJ often results in the loss (deletion) of nucleotide sequence near the site of the double-strand break.

[00124] The terms "treatment", "treating" and the like are used herein to generally mean obtaining a desired pharmacologic and/or physiologic effect. The effect may be prophylactic in terms of completely or partially preventing a disease or symptom thereof and/or may be therapeutic in terms of a partial or complete cure for a disease and/or adverse effect attributable to the disease. "Treatment" as used herein covers any treatment of a disease or symptom in a mammal, and includes: (a) preventing the disease or symptom from occurring in a subject which may be predisposed to acquiring the disease or symptom but has not yet been diagnosed as having it; (b) inhibiting the disease or symptom, i.e., arresting its development; or (c) relieving the disease, i.e., causing regression of the disease. The therapeutic agent may be administered before, during or after the onset of disease or injury. The treatment of ongoing disease, where the treatment stabilizes or reduces the undesirable clinical symptoms of the patient, is of particular interest. Such treatment is desirably performed prior to complete loss of function in the affected tissues. The therapy will desirably be administered during the symptomatic stage of the disease, and in some cases after the symptomatic stage of the disease.

[00125] The terms "individual," "subject," "host," and "patient," are used interchangeably herein and refer to any mammalian subject for whom diagnosis, treatment, or therapy is desired, particularly humans.

[00126] General methods in molecular and cellular biochemistry can be found in such standard textbooks as Molecular Cloning: A Laboratory Manual, 3rd Ed. (Sambrook et al., HaRBor Laboratory Press 2001); Short Protocols in Molecular Biology, 4th Ed. (Ausubel et al. eds., John Wiley & Sons 1999); Protein Methods (Bollag et al., John Wiley & Sons 1996); Nonviral Vectors for Gene Therapy (Wagner et al. eds., Academic Press 1999); Viral Vectors (Kaplift & Loewy eds., Academic Press 1995); Immunology Methods Manual (I. Lefkovits ed., Academic Press 1997); and Cell and Tissue Culture: Laboratory Procedures in Biotechnology (Doyle & Griffiths, John Wiley & Sons 1998), the disclosures of which are incorporated herein by reference.

[00127] Where a range of values is provided, it is understood that each intervening value, to the tenth of the unit of the lower limit unless the context clearly dictates otherwise, between the upper and lower limit of that range and any other stated or intervening value in that stated range, is encompassed within the invention. The upper and lower limits of these smaller ranges may independently be included in the smaller ranges, and are also encompassed within the invention, subject to any specifically excluded limit in the stated range. Where the stated range includes one or both of the limits, ranges excluding either or both of those included limits are also included in the invention.

[00128] The phrase “consisting essentially of” is meant herein to exclude anything that is not the specified active component or components of a system, or that is not the specified active portion or portions of a molecule.

[00129] Certain ranges are presented herein with numerical values being preceded by the term "about." The term "about" is used herein to provide literal support for the exact number that it precedes, as well as a number that is near to or approximately the number that the term precedes. In determining whether a number is near to or approximately a specifically recited number, the near or approximating unrecited number may be a number which, in the context in which it is presented, provides the substantial equivalent of the specifically recited number.

[00130] It is appreciated that certain features of the invention, which are, for clarity, described in the context of separate embodiments, may also be provided in combination in a single embodiment. Conversely, various features of the invention, which are, for brevity, described in the context of a single embodiment, may also be provided separately or in any suitable sub-combination. All combinations of the embodiments pertaining to the invention are specifically embraced by the present invention and are disclosed herein just as if each and every combination was individually and explicitly disclosed. In addition, all sub-combinations of the various embodiments and elements thereof are also specifically embraced by

the present invention and are disclosed herein just as if each and every such sub-combination was individually and explicitly disclosed herein.

[00131] Aspects of the Disclosure - Part I

[00132] Nucleic Acids

[00133] Guide RNA

[00134] The present disclosure provides a guide RNA that directs the activities of an associated polypeptide (e.g., a site-directed modifying polypeptide) to a specific target sequence within a target DNA. A guide RNA comprises: a first segment (also referred to herein as a "DNA-targeting segment" or a "DNA-targeting sequence") and a second segment (also referred to herein as a "protein-binding segment" or a "protein-binding sequence").

[00135] DNA-targeting segment of a guide RNA

[00136] The DNA-targeting segment of a guide RNA comprises a nucleotide sequence that is complementary to a sequence in a target DNA. In other words, the DNA-targeting segment of a guide RNA interacts with a target DNA in a sequence-specific manner via hybridization (i.e., base pairing). As such, the nucleotide sequence of the DNA-targeting segment may vary and determines the location within the target DNA that the guide RNA and the target DNA will interact. The DNA-targeting segment of a guide RNA can be modified (e.g., by genetic engineering) to hybridize to any desired sequence within a target DNA.

[00137] The DNA-targeting segment can have a length of from about 12 nucleotides to about 100 nucleotides. For example, the DNA-targeting segment can have a length of from about 12 nucleotides (nt) to about 80 nt, from about 12 nt to about 50 nt, from about12 nt to about 40 nt, from about 12 nt to about 30 nt, from about 12 nt to about 25 nt, from about 12 nt to about 20 nt, or from about 12 nt to about 19 nt. For example, the DNA-targeting segment can have a length of from about 19 nt to about 20 nt, from about 19 nt to about 25 nt, from about 19 nt to about 30 nt, from about 19 nt to about 35 nt, from about 19 nt to about 40 nt, from about 19 nt to about 45 nt, from about 19 nt to about 50 nt, from about 19 nt to about 60 nt, from about 19 nt to about 70 nt, from about 19 nt to about 80 nt, from about 19 nt to about 90 nt, from about 19 nt to about 100 nt, from about 20 nt to about 25 nt, from about 20 nt to about 30 nt, from about 20 nt to about 35 nt, from about 20 nt to about 40 nt, from about 20 nt to about 45 nt, from about 20 nt to about 50 nt, from about 20 nt to about 60 nt, from about 20 nt to about 70 nt, from about 20 nt to about 80 nt, from about 20 nt to about 90 nt, or from about 20 nt to about 100 nt. The nucleotide sequence (the DNA-targeting sequence) of the DNA-targeting segment that is complementary to a nucleotide sequence (target sequence) of the target DNA can have a length at least about 12 nt. For example, the DNA-targeting sequence of the DNA-targeting segment that is complementary to a target sequence of the target DNA can have a length at least about 12 nt, at least about 15 nt, at least about 18 nt, at least about 19 nt,

at least about 20 nt, at least about 25 nt, at least about 30 nt, at least about 35 nt or at least about 40 nt. For example, the DNA-targeting sequence of the DNA-targeting segment that is complementary to a target sequence of the target DNA can have a length of from about 12 nucleotides (nt) to about 80 nt, from about 12 nt to about 50 nt, from about 12 nt to about 45 nt, from about 12 nt to about 40 nt, from about 12 nt to about 35 nt, from about 12 nt to about 30 nt, from about 12 nt to about 25 nt, from about 12 nt to about 20 nt, from about 12 nt to about 19 nt, from about 19 nt to about 20 nt, from about 19 nt to about 25 nt, from about 19 nt to about 30 nt, from about 19 nt to about 35 nt, from about 19 nt to about 40 nt, from about 19 nt to about 45 nt, from about 19 nt to about 50 nt, from about 19 nt to about 60 nt, from about 20 nt to about 25 nt, from about 20 nt to about 30 nt, from about 20 nt to about 35 nt, from about 20 nt to about 40 nt, from about 20 nt to about 45 nt, from about 20 nt to about 50 nt, or from about 20 nt to about 60 nt. The nucleotide sequence (the DNA-targeting sequence) of the DNA-targeting segment that is complementary to a nucleotide sequence (target sequence) of the target DNA can have a length at least about 12 nt.

[00138] In some cases, the DNA-targeting sequence of the DNA-targeting segment that is complementary to a target sequence of the target DNA is 20 nucleotides in length. In some cases, the DNA-targeting sequence of the DNA-targeting segment that is complementary to a target sequence of the target DNA is 16 nucleotides, 17 nucleotides, 18 nucleotides or 19 nucleotides in length.

[00139] The percent complementarity between the DNA-targeting sequence of the DNA-targeting segment and the target sequence of the target DNA can be at least 60% (e.g., at least 65%, at least 70%, at least 75%, at least 80%, at least 85%, at least 90%, at least 95%, at least 97%, at least 98%, at least 99%, or 100%). For example, the DNA-targeting sequence may be at least about 80% identical to about 10 contiguous nucleotides, or at least about 80% identical to about 11 contiguous nucleotides, or at least about 80% identical to about 12 contiguous nucleotides, or at least about 80% identical to about 13 contiguous nucleotides, or at least about 80% identical to about 14 contiguous nucleotides, or at least about 80% identical to about 15 contiguous nucleotides, or at least about 80% identical to about 16 contiguous nucleotides, or at least about 80% identical to about 17 contiguous nucleotides of the target sequence. In some cases, the percent complementarity between the DNA-targeting sequence of the DNA-targeting segment and the target sequence of the target DNA is 100% over the seven contiguous 5'-most nucleotides of the target sequence of the complementary strand of the target DNA. In some cases, the percent complementarity between the DNA-targeting sequence of the DNA-targeting segment and the target sequence of the target DNA is at least 60% over about 20 contiguous nucleotides. In some cases, the percent complementarity between the DNA-targeting sequence of the DNA-targeting segment and the target sequence of the target DNA is 100% over the fourteen contiguous 5'-most nucleotides of the target sequence of the complementary strand of the target DNA and as low as 0% over the remainder. In such a case, the DNA-targeting sequence can be considered to be 14 nucleotides in length. In some cases, the percent complementarity between the DNA-targeting sequence of the DNA-targeting segment and the

target sequence of the target DNA is 100% over the seven contiguous 5'-most nucleotides of the target sequence of the complementary strand of the target DNA and as low as 0% over the remainder. In such a case, the DNA-targeting sequence can be considered to be 7 nucleotides in length.

[00140] Protein-binding segment of a guide RNA

[00141] The protein-binding segment of a guide RNA interacts with a site-directed modifying polypeptide. The guide RNA guides the bound polypeptide to a specific nucleotide sequence within target DNA via the above mentioned DNA-targeting segment. The protein-binding segment of a guide RNA comprises two stretches of nucleotides that are complementary to one another. The complementary nucleotides of the protein-binding segment hybridize to form a double stranded RNA duplex (dsRNA).

[00142] A double-molecule guide RNA comprises two separate RNA molecules. Each of the two RNA molecules of a double-molecule guide RNA comprises a stretch of nucleotides that are complementary to one another such that the complementary nucleotides of the two RNA molecules hybridize to form the double-stranded RNA duplex of the protein-binding segment.

[00143] In some embodiments, the duplex-forming segment of the activator-RNA is at least about 60% identical to one of the activator-RNA (tracrRNA) molecules set forth in Supplementary Table S5, or a complement thereof, over a stretch of at least 8 contiguous nucleotides. For example, the duplex-forming segment of the activator-RNA (or the DNA encoding the duplex-forming segment of the activator-RNA) is at least about 60% identical, at least about 65% identical, at least about 70% identical, at least about 75% identical, at least about 80% identical, at least about 85% identical, at least about 90% identical, at least about 95% identical, at least about 98% identical, at least about 99% identical, or 100 % identical, to one of the tracrRNA sequences set forth in Supplementary Table S5, or a complement thereof, over a stretch of at least 8 contiguous , at least 9 contiguous, at least 10 contiguous, at least 11 contiguous, at least 12 contiguous, at least 13 contiguous, at least 14 contiguous or at least 15 contiguous nucleotides. For example, the activator-RNA may comprise a nucleotide sequence that is at least 70% identical over at least 10 contiguous nucleotides, at least 80% identical over at least 10 contiguous nucleotides, at least 70% identical over at least 11 contiguous nucleotides, at least 80% identical over at least 11 contiguous nucleotides, at least 70% identical over at least 12 contiguous nucleotides, or at least 80% identical over at least 12 contiguous nucleotides of one of the tracrRNA sequences set forth in Supplementary Table S5.

[00144] In some embodiments, the duplex-forming segment of the targeter-RNA is at least about 60% identical to one of the targeter-RNA (crRNA/CRISPR repeat) seqeunces set forth in Supplementary Table S5, or a complement thereof, over a stretch of at least 8 contiguous nucleotides. For example, the duplex-forming segment of the targeter-RNA (or the DNA encoding the duplex-forming segment of the targeter-RNA) is at least about 65% identical, at least about 70% identical, at least about 75% identical, at least about 80% identical, at least about 85% identical, at least about 90% identical, at least about 95%

identical, at least about 98% identical, at least about 99% identical or 100 % identical to one of the crRNA/CRISPR repeat sequences set forth in Supplementary Table S5, or a complement thereof, over a stretch of at least 8 contiguous , at least 9 contiguous, at least 10 contiguous, at least 11 contiguous, at least 12 contiguous, at least 13 contiguous, at least 14 contiguous or at least 15 contiguous nucleotides. For example, the targeter-RNA may comprise a nucleotide sequence that is at least 70% identical over at least 10 contiguous nucleotides, at least 80% identical over at least 10 contiguous nucleotides, at least 70% identical over at least 11 contiguous nucleotides, at least 80% identical over at least 11 contiguous nucleotides, at least 70% identical over at least 12 contiguous nucleotides, at least 80% identical over at least 12 contiguous nucleotides, at least 80% identical over at least 13 contiguous nucleotides, at least 80% identical over at least 14 contiguous nucleotides, at least 80% identical over at least 15 contiguous nucleotides, at least 80% identical over at least 16 contiguous nucleotides, or at least 80% identical over at least 17 contiguous nucleotides, to one of the CRISPR repeat sequences set forth in Supplementary Table S5.

[00145] A two-molecule guide RNA can be designed to allow for controlled (i.e., conditional) binding of a targeter-RNA with an activator-RNA. Because a two-molecule guide RNA is not functional unless both the activator-RNA and the targeter-RNA are bound in a functional complex with Cas9, a two-molecule guide RNA can be inducible (e.g., drug inducible) by rendering the binding between the activator-RNA and the targeter-RNA to be inducible. As one non-limiting example, RNA aptamers can be used to regulate (i.e., control) the binding of the activator-RNA with the targeter-RNA. Accordingly, the activator-RNA and/or the targeter-RNA can comprise an RNA aptamer sequence.

[00146] RNA aptamers are known in the art and are generally a synthetic version of a riboswitch. The terms "RNA aptamer" and "riboswitch" are used interchangeably herein to encompass both synthetic and natural nucleic acid sequences that provide for inducible regulation of the structure (and therefore the availability of specific sequences) of the RNA molecule of which they are part. RNA aptamers usually comprise a sequence that folds into a particular structure (e.g., a hairpin), which specifically binds a particular drug (e.g., a small molecule). Binding of the drug causes a structural change in the folding of the RNA, which changes a feature of the nucleic acid of which the aptamer is a part. As non-limiting examples: (i) an activator-RNA with an aptamer may not be able to bind to the cognate targeter-RNA unless the aptamer is bound by the appropriate drug; (ii) a targeter-RNA with an aptamer may not be able to bind to the cognate activator-RNA unless the aptamer is bound by the appropriate drug; and (iii) a targeter-RNA and an activator-RNA, each comprising a different aptamer that binds a different drug, may not be able to bind to each other unless both drugs are present. As illustrated by these examples, a two-molecule guide RNA can be designed to be inducible.

[00147] Examples of aptamers and riboswitches can be found, for example, in: Nakamura et al., Genes Cells. 2012 May;17(5):344-64; Vavalle et al., Future Cardiol. 2012 May;8(3):371-82; Citartan et al.,

Biosens Bioelectron. 2012 Apr 15;34(1):1-11; and Liberman et al., Wiley Interdiscip Rev RNA. 2012 May-Jun;3(3):369-84; all of which are herein incorporated by reference in their entirety.

[00148] Non-limiting examples of nucleotide sequences that can be included in a two-molecule guide RNA include either of the sequences set forth in Supplmentary Table S5, or complements thereof pairing with any sequences set forth in Supplementary Table S5,or complements thereof that can hybridize to form a protein binding segment.

[00149] A single-molecule guide RNA comprises two stretches of nucleotides (a targeter-RNA and an activator-RNA) that are complementary to one another, are covalently linked (directly, or by intervening nucleotides referred to as "linkers" or "linker nucleotides"), and hybridize to form the double stranded RNA duplex (dsRNA duplex) of the protein-binding segment, thus resulting in a stem-loop structure. The targeter-RNA and the activator-RNA can be covalently linked via the 3' end of the targeter-RNA and the 5' end of the activator-RNA. Alternatively, targeter-RNA and the activator-RNA can be covalently linked via the 5' end of the targeter-RNA and the 3' end of the activator-RNA.

[00150] The linker of a single-molecule guide RNA can have a length of from about 3 nucleotides to about 100 nucleotides. For example, the linker can have a length of from about 3 nucleotides (nt) to about 90 nt, from about 3 nucleotides (nt) to about 80 nt, from about 3 nucleotides (nt) to about 70 nt, from about 3 nucleotides (nt) to about 60 nt, from about 3 nucleotides (nt) to about 50 nt, from about 3 nucleotides (nt) to about 40 nt, from about 3 nucleotides (nt) to about 30 nt, from about 3 nucleotides (nt) to about 20 nt or from about 3 nucleotides (nt) to about 10 nt. For example, the linker can have a length of from about 3 nt to about 5 nt, from about 5 nt to about 10 nt, from about 10 nt to about 15 nt, from about 15 nt to about 20 nt, from about 20 nt to about 25 nt, from about 25 nt to about 30 nt, from about 30 nt to about 35 nt, from about 35 nt to about 40 nt, from about 40 nt to about 50 nt, from about 50 nt to about 60 nt, from about 60 nt to about 70 nt, from about 70 nt to about 80 nt, from about 80 nt to about 90 nt, or from about 90 nt to about 100 nt. In some embodiments, the linker of a single-molecule guide RNA is 4 nt.

[00151] An exemplary single-molecule guide RNA comprises two complementary stretches of nucleotides that hybridize to form a dsRNA duplex. In some embodiments, one of the two complementary stretches of nucleotides of the single-molecule guide RNA (or the DNA encoding the stretch) is at least about 60% identical to one of the activator-RNA (tracrRNA) molecules set forth in Supplementary Table S5, or a complement thereof, over a stretch of at least 8 contiguous nucleotides. For example, one of the two complementary stretches of nucleotides of the single-molecule guide RNA (or the DNA encoding the stretch) is at least about 65% identical, at least about 70% identical, at least about 75% identical, at least about 80% identical, at least about 85% identical, at least about 90% identical, at least about 95% identical, at least about 98% identical, at least about 99% identical or 100 % identical to one of the tracrRNA sequences set forth in Supplementary Table S5, or a complement thereof, over a stretch of at least 8 contiguous , at least 9 contiguous, at least 10 contiguous, at least 11 contiguous, at least 12

contiguous, at least 13 contiguous, at least 14 contiguous or at least 15 contiguous nucleotides. For example, the single-molecule guide RNA may comprise a nucleotide sequence that is at least 70% identical over at least 10 contiguous nucleotides, at least 80% identical over at least 10 contiguous nucleotides, at least 70% identical over at least 11 contiguous nucleotides, at least 80% identical over at least 11 contiguous nucleotides, at least 70% identical over at least 12 contiguous nucleotides, or at least 80% identical over at least 12 contiguous nucleotides of one of the tracrRNA sequences set forth in Supplementary Table S5.

[00152] In some embodiments, one of the two complementary stretches of nucleotides of the single-molecule guide RNA (or the DNA encoding the stretch) is at least about 60% identical to one of the targeter-RNA (crRNA/CRISPR repeat) sequences set forth in Supplementary Table S5, or a complement thereof, over a stretch of at least 8 contiguous nucleotides. For example, one of the two complementary stretches of nucleotides of the single-molecule guide RNA (or the DNA encoding the stretch) is at least about 65% identical, at least about 70% identical, at least about 75% identical, at least about 80% identical, at least about 85% identical, at least about 90% identical, at least about 95% identical, at least about 98% identical, at least about 99% identical or 100 % identical to one of the crRNA/CRISPR repeat sequences set forth in Supplementary Table S5, or a complement thereof, over a stretch of at least 8 contiguous , at least 9 contiguous, at least 10 contiguous, at least 11 contiguous, at least 12 contiguous, at least 13 contiguous, at least 14 contiguous or at least 15 contiguous nucleotides. For example, the single-molecule guide RNA may comprise a nucleotide sequence that is at least 70% identical over at least 10 contiguous nucleotides, at least 80% identical over at least 10 contiguous nucleotides, at least 70% identical over at least 11 contiguous nucleotides, at least 80% identical over at least 11 contiguous nucleotides, at least 70% identical over at least 12 contiguous nucleotides, or at least 80% identical over at least 12 contiguous nucleotides, or at least about 80% identical to about 13 contiguous nucleotides, or at least about 80% identical to about 14 contiguous nucleotides, or at least about 80% identical to about 15 contiguous nucleotides, or at least about 80% identical to about 16 contiguous nucleotides, or at least about 80% identical to about 17 contiguous nucleotides of one of the CRISPR repeat sequences set forth in Supplementary Table S5.

[00153] Appropriate naturally occurring cognate pairs of crRNAs and tracrRNAs can be routinely determined by taking into account the species name and base-pairing (for the dsRNA duplex of the protein-binding domain) when determining appropriate cognate pairs. Non-cognate pairs are also contemplated for use in the invention. In some embodiments of non-cognate pairs, each RNA is from a Cas9 cluster herein wherein the Cas9 endonucleases share 80% identity over 80% of their amino acid sequences.

[00154] Artificial sequences that share very little identity (roughly 50% identity, or alternatively about 70% identity over about 50% of the full length protein) with naturally occurring a tracrRNAs and crRNAs can function with Cas9 to cleave target DNA as long as the structure of the protein-binding domain of the

guide RNA is conserved. Thus, RNA folding structure of a naturally occuring protein-binding domain of a DNA-trageting RNA can be taken into account in order to design artificial protein-binding domains (either two-molecule or single-molecule versions). As structures can readily be produced by one of ordinary skill in the art for any naturally occurring crRNA:tracrRNA pair from any, an artificial DNA-targeting-RNA can be designed to mimic the natural structure for a given species when using the Cas9 (or a related Cas9) from that species. Thus, a suitable guide RNA can be an artificially designed RNA (non-naturally occurring) comprising a protein-binding domain that was designed to mimic the structure of a protein-binding domain of a naturally occurring guide RNA.

[00155] The protein-binding segment can have a length of from about 10 nucleotides to about 100 nucleotides. For example, the protein-binding segment can have a length of from about 15 nucleotides (nt) to about 80 nt, from about 15 nt to about 50 nt, from about 15 nt to about 40 nt, from about 15 nt to about 30 nt or from about 15 nt to about 25 nt.

[00156] Also with regard to both a single-molecule guide RNA and to a double-molecule guide RNA, the dsRNA duplex of the protein-binding segment can have a length from about 6 base pairs (bp) to about 50bp. For example, the dsRNA duplex of the protein-binding segment can have a length from about 6 bp to about 40 bp, from about 6 bp to about 30bp, from about 6 bp to about 25 bp, from about 6 bp to about 20 bp, from about 6 bp to about 15 bp, from about 8 bp to about 40 bp, from about 8 bp to about 30bp, from about 8 bp to about 25 bp, from about 8 bp to about 20 bp or from about 8 bp to about 15 bp. For example, the dsRNA duplex of the protein-binding segment can have a length from about from about 8 bp to about 10 bp, from about 10 bp to about 15 bp, from about 15 bp to about 18 bp, from about 18 bp to about 20 bp, from about 20 bp to about 25 bp, from about 25 bp to about 30 bp, from about 30 bp to about 35 bp, from about 35 bp to about 40 bp, or from about 40 bp to about 50 bp. In some embodiments, the dsRNA duplex of the protein-binding segment has a length of 36 base pairs. The percent complementarity between the nucleotide sequences that hybridize to form the dsRNA duplex of the protein-binding segment can be at least about 60%. For example, the percent complementarity between the nucleotide sequences that hybridize to form the dsRNA duplex of the protein-binding segment can be at least about 65%, at least about 70%, at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 95%, at least about 98%, or at least about 99% . In some cases, the percent complementarity between the nucleotide sequences that hybridize to form the dsRNA duplex of the protein-binding segment is 100%.

[00157] Site-directed modifying polypeptide

[00158] A guide RNA and a site-directed modifying polypeptide form a complex. The guide RNA provides target specificity to the complex by comprising a nucleotide sequence that is complementary to a sequence of a target DNA (as noted above). The site-directed modifying polypeptide is guided to a DNA sequence (e.g. a chromosomal sequence or an extrachromosomal sequence, e.g. an episomal

sequence, a minicircle sequence, a mitochondrial sequence, a chloroplast sequence, etc.) by virtue of its association with at least the protein-binding segment of the guide RNA (described above).

[00159] A site-directed modifying polypeptide modifies target DNA (e.g., cleavage or methylation of target DNA) and/or a polypeptide associated with target DNA (e.g., methylation or acetylation of a histone tail). A site-directed modifying polypeptide is also referred to herein as a "site-directed polypeptide" or an "RNA binding site-directed modifying polypeptide." In some cases, the site-directed modifying polypeptide is a naturally-occurring modifying polypeptide. In other cases, the site-directed modifying polypeptide is not a naturally-occurring polypeptide (e.g., a chimeric polypeptide as discussed below or a naturally-occurring polypeptide that is modified, e.g., mutation, deletion, insertion).

[00160] Naturally-occurring site-directed modifying polypeptides bind a guide RNA, are thereby directed to a specific sequence within a target DNA, and cleave the target DNA to generate a double strand break. The amino acid sequences of exemplary naturally-occurring Cas9 site-directed modifying polypeptide orthologs are set out in SEQ ID NOs: 1-800. The amino acid sequence of the S. pyrogenes Cas9 endonuclease is set out in SEQ ID NO: 8. A site-directed modifying polypeptide comprises two portions, an RNA-binding portion and an activity portion. In some embodiments, a site-directed modifying polypeptide comprises: (i) an RNA-binding portion that interacts with a guide RNA, wherein the guide RNA comprises a nucleotide sequence that is complementary to a sequence in a target DNA; and (ii) an activity portion that exhibits site-directed enzymatic activity (e.g., activity for DNA methylation, activity for DNA cleavage, activity for histone acetylation, activity for histone methylation, etc.), wherein the site of enzymatic activity is determined by the guide RNA.

[00161] In other embodiments, a site-directed modifying polypeptide comprises: (i) an RNA-binding portion that interacts with a guide RNA, wherein the guide RNA comprises a nucleotide sequence that is complementary to a sequence in a target DNA; and (ii) an activity portion that modulates transcription within the target DNA (e.g., to increase or decrease transcription), wherein the site of modulated transcription within the target DNA is determined by the guide RNA.

[00162] In some cases, a site-directed modifying polypeptide has enzymatic activity that modifies target DNA (e.g., nuclease activity, methyltransferase activity, demethylase activity, DNA repair activity, DNA damage activity, deamination activity, dismutase activity, alkylation activity, depurination activity, oxidation activity, pyrimidine dimer forming activity, integrase activity, transposase activity, recombinase activity, polymerase activity, ligase activity, helicase activity, photolyase activity or glycosylase activity).

[00163] In other cases, a site-directed modifying polypeptide has enzymatic activity that modifies a polypeptide (e.g., a histone) associated with target DNA (e.g., methyltransferase activity, demethylase activity, acetyltransferase activity, deacetylase activity, kinase activity, phosphatase activity, ubiquitin ligase activity, deubiquitinating activity, adenylation activity, deadenylation activity, SUMOylating activity,

deSUMOylating activity, ribosylation activity, deribosylation activity, myristoylation activity or demyristoylation activity).

[00164] Exemplary site-directed modifying polypeptides

[00165] In some cases, the site-directed modifying polypeptide comprises an amino acid sequence having at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 95%, at least about 99%, or 100%, amino acid sequence identity to amino acids 7-166 and/or 731-1003 of SEQ ID NO: 8, or to the corresponding portions in any of the amino acid sequences set forth as SEQ ID NOs: 1-800.

[00166] Nucleic acid modifications

[00167] In some embodiments, a nucleic acid (e.g., a guide RNA) comprises one or more modifications, e.g., a base modification, a backbone modification, etc, to provide the nucleic acid with a new or enhanced feature (e.g., improved stability). As is known in the art, a nucleoside is a base-sugar combination. The base portion of the nucleoside is normally a heterocyclic base. The two most common classes of such heterocyclic bases are the purines and the pyrimidines. Nucleotides are nucleosides that further include a phosphate group covalently linked to the sugar portion of the nucleoside. For those nucleosides that include a pentofuranosyl sugar, the phosphate group can be linked to the 2', the 3', or the 5' hydroxyl moiety of the sugar. In forming oligonucleotides, the phosphate groups covalently link adjacent nucleosides to one another to form a linear polymeric compound. In turn, the respective ends of this linear polymeric compound can be further joined to form a circular compound, however, linear compounds are generally suitable. In addition, linear compounds may have internal nucleotide base complementarity and may therefore fold in a manner as to produce a fully or partially double-stranded compound. Within oligonucleotides, the phosphate groups are commonly referred to as forming the internucleoside backbone of the oligonucleotide. The normal linkage or backbone of RNA and DNA is a 3' to 5' phosphodiester linkage.

[00168] Modified backbones and modified internucleoside linkages

[00169] Examples of suitable nucleic acids containing modifications include nucleic acids containing modified backbones or non-natural internucleoside linkages. Nucleic acids having modified backbones include those that retain a phosphorus atom in the backbone and those that do not have a phosphorus atom in the backbone.

[00170] Suitable modified oligonucleotide backbones containing a phosphorus atom therein include, for example, phosphorothioates, chiral phosphorothioates, phosphorodithioates, phosphotriesters, aminoalkylphosphotriesters, methyl and other alkyl phosphonates including 3'-alkylene phosphonates, 5'-alkylene phosphonates and chiral phosphonates, phosphinates, phosphoramidates including 3^1-amino phosphoramidate and aminoalkylphosphoramidates, phosphorodiamidates, thionophosphoramidates,

thionoalkylphosphonates, thionoalkylphosphotriesters, selenophosphates and boranophosphates having normal 3'-5' linkages, 2'-5' linked analogs of these, and those having inverted polarity wherein one or more internucleotide linkages is a 3' to 3', 5' to 5' or 2' to 2' linkage. Suitable oligonucleotides having inverted polarity comprise a single 3' to 3' linkage at the 3'-most internucleotide linkage i.e. a single inverted nucleoside residue which may be a basic (the nucleobase is missing or has a hydroxyl group in place thereof). Various salts (such as, for example, potassium or sodium), mixed salts and free acid forms are also included.

[00171] In some embodiments, a nucleic acid comprises one or more phosphorothioate and/or heteroatom internucleoside linkages, in particular $-CH_2-NH-O-CH_2-$, $-CH_2-N(CH_3)-O-CH_2-$ (known as a methylene (methylimino) or MMI backbone), $-CH_2-O-N(CH_3)-CH_2-$, $-CH_2-N(CH_3)-N(CH_3)-CH_2-$ and $-O-N(CH_3)-CH_2-CH_2-$ (wherein the native phosphodiester internucleotide linkage is represented as $-O-P(=O)(OH)-O-CH_2-$). MMI type internucleoside linkages are disclosed in the above referenced U.S. Pat. No. 5,489,677. Suitable amide internucleoside linkages are disclosed in t U.S. Pat. No. 5,602,240.

[00172] Also suitable are nucleic acids having morpholino backbone structures as described in, e.g., U.S. Pat. No. 5,034,506. For example, in some embodiments, a nucleic acid comprises a 6-membered morpholino ring in place of a ribose ring. In some of these embodiments, a phosphorodiamidate or other non-phosphodiester internucleoside linkage replaces a phosphodiester linkage.

[00173] Suitable modified polynucleotide backbones that do not include a phosphorus atom therein have backbones that are formed by short chain alkyl or cycloalkyl internucleoside linkages, mixed heteroatom and alkyl or cycloalkyl internucleoside linkages, or one or more short chain heteroatomic or heterocyclic internucleoside linkages. These include those having morpholino linkages (formed in part from the sugar portion of a nucleoside); siloxane backbones; sulfide, sulfoxide and sulfone backbones; formacetyl and thioformacetyl backbones; methylene formacetyl and thioformacetyl backbones; riboacetyl backbones; alkene containing backbones; sulfamate backbones; methyleneimino and methylenehydrazino backbones; sulfonate and sulfonamide backbones; amide backbones; and others having mixed N, 0, S and CH_2 component parts.

[00174] Mimetics

[00175] A nucleic acid can be a nucleic acid mimetic. The term "mimetic" as it is applied to polynucleotides is intended to include polynucleotides wherein only the furanose ring or both the furanose ring and the internucleotide linkage are replaced with non-furanose groups, replacement of only the furanose ring is also referred to in the art as being a sugar surrogate. The heterocyclic base moiety or a modified heterocyclic base moiety is maintained for hybridization with an appropriate target nucleic acid. One such nucleic acid, a polynucleotide mimetic that has been shown to have excellent hybridization properties, is referred to as a peptide nucleic acid (PNA). In PNA, the sugar-backbone of a polynucleotide is

replaced with an amide containing backbone, in particular an aminoethylglycine backbone. The nucleotides are retained and are bound directly or indirectly to aza nitrogen atoms of the amide portion of the backbone.

[00176] One polynucleotide mimetic that has been reported to have excellent hybridization properties is a peptide nucleic acid (PNA). The backbone in PNA compounds is two or more linked aminoethylglycine units which gives PNA an amide containing backbone. The heterocyclic base moieties are bound directly or indirectly to aza nitrogen atoms of the amide portion of the backbone. Representative U.S. patents that describe the preparation of PNA compounds include, but are not limited to: U.S. Pat. Nos. 5,539,082; 5,714,331; and 5,719,262.

[00177] Another class of polynucleotide mimetic that has been studied is based on linked morpholino units (morpholino nucleic acid) having heterocyclic bases attached to the morpholino ring. A number of linking groups have been reported that link the morpholino monomeric units in a morpholino nucleic acid. One class of linking groups has been selected to give a non-ionic oligomeric compound. The non-ionic morpholino-based oligomeric compounds are less likely to have undesired interactions with cellular proteins. Morpholino-based polynucleotides are nonionic mimics of oligonucleotides which are less likely to form undesired interactions with cellular proteins (Dwaine A. Braasch and David R. Corey, Biochemistry, 2002, 41(14), 45034510). Morpholino-based polynucleotides are disclosed in U.S. Pat. No. 5,034,506. A variety of compounds within the morpholino class of polynucleotides have been prepared, having a variety of different linking groups joining the monomeric subunits.

[00178] A further class of polynucleotide mimetic is referred to as cyclohexenyl nucleic acids (CeNA). The furanose ring normally present in a DNA/RNA molecule is replaced with a cyclohexenyl ring. CeNA DMT protected phosphoramidite monomers have been prepared and used for oligomeric compound synthesis following classical phosphoramidite chemistry. Fully modified CeNA oligomeric compounds and oligonucleotides having specific positions modified with CeNA have been prepared and studied (see Wang et al., J. Am. Chem. Soc., 2000, 122, 85958602). In general the incorporation of CeNA monomers into a DNA chain increases its stability of a DNA/RNA hybrid. CeNA oligoadenylates formed complexes with RNA and DNA complements with similar stability to the native complexes. The study of incorporating CeNA structures into natural nucleic acid structures was shown by NMR and circular dichroism to proceed with easy conformational adaptation.

[00179] A further modification includes Locked Nucleic Acids (LNAs) in which the 2'-hydroxyl group is linked to the 4' carbon atom of the sugar ring thereby forming a 2'-C,4'-C-oxymethylene linkage thereby forming a bicyclic sugar moiety. The linkage can be a methylene (-CH_2-), group bridging the 2' oxygen atom and the 4' carbon atom wherein n is 1 or 2 (Singh et al., Chem. Commun., 1998, 4, 455-456). LNA and LNA analogs display very high duplex thermal stabilities with complementary DNA and RNA (Tm=+3 to +10° C), stability towards 3'-exonucleolytic degradation and good solubility properties. Potent and

nontoxic antisense oligonucleotides containing LNAs have been described (Wahlestedt et al., Proc. Natl. Acad. Sci. U.S.A., 2000, 97, 5633-5638).

[00180] The synthesis and preparation of the LNA monomers adenine, cytosine, guanine, 5-methyl-cytosine, thymine and uracil, along with their oligomerization, and nucleic acid recognition properties have been described (Koshkin et al., Tetrahedron, 1998, 54, 3607-3630). LNAs and preparation thereof are also described in WO 98/39352 and WO 99/14226.

[00181] Modified sugar moieties

[00182] A nucleic acid can also include one or more substituted sugar moieties. Suitable polynucleotides comprise a sugar substituent group selected from: OH; F; O-, S-, or N-alkyl; O-, S-, or N-alkenyl; O-, S- or N-alkynyl; or O-alkyl-O-alkyl, wherein the alkyl, alkenyl and alkynyl may be substituted or unsubstituted C.sub.1 to C_{10} alkyl or C_2 to C_{10} alkenyl and alkynyl. Particularly suitable are $O((CH_2)_nO)_mCH_3$, $O(CH_2)_nOCH_3$, $O(CH_2)_nNH_2$, $O(CH_2)CH_3$, $O(CH_2)_nONH_2$, and $O(CH_2)_nON((CH_2)_nCH_3)_2$, where n and m are from 1 to about 10. Other suitable polynucleotides comprise a sugar substituent group selected from: C_1 to C_{10} lower alkyl, substituted lower alkyl, alkenyl, alkynyl, alkaryl, aralkyl, O-alkaryl or O-aralkyl, SH, SCH_3, OCN, Cl, Br, CN, CF_3, OCF_3, $SOCH_3$, SO_2CH_3, ONO_2, NO_2, N_3, NH_2, heterocycloalkyl, heterocycloalkaryl, aminoalkylamino, polyalkylamino, substituted silyl, an RNA cleaving group, a reporter group, an intercalator, a group for improving the pharmacokinetic properties of an oligonucleotide, or a group for improving the pharmacodynamic properties of an oligonucleotide, and other substituents having similar properties. A suitable modification includes 2'-methoxyethoxy 2'-O-CH_2 CH_2OCH_3, also known as -2'-O-(2-methoxyethyl) or 2'-MOE) (Martin et al., Hely. Chim. Acta, 1995, 78, 486-504) i.e., an alkoxyalkoxy group. A further suitable modification includes 2'-dimethylaminooxyethoxy, i.e., a $O(CH_2)_2ON(CH_3)_2$ group, also known as 2'-DMAOE, as described in examples hereinbelow, and 2'-dimethylaminoethoxyethoxy (also known in the art as 2'-O-dimethyl-amino-ethoxy-ethyl or 2'-DMAEOE), i.e., 2'-O-CH_2-O-CH_2-$N(CH_3)_2$.

[00183] Other suitable sugar substituent groups include methoxy (-O-CH_3), aminopropoxy (--O--CH_2CH_2 CH_2NH_2), allyl (-CH_2-CH=CH_2), -O-allyl (--O--CH_2—CH=CH_2) and fluoro (F). 2'-sugar substituent groups may be in the arabino (up) position or ribo (down) position. A suitable 2'-arabino modification is 2'-F. Similar modifications may also be made at other positions on the oligomeric compound, particularly the 3' position of the sugar on the 3' terminal nucleoside or in 2'-5' linked oligonucleotides and the 5' position of 5' terminal nucleotide. Oligomeric compounds may also have sugar mimetics such as cyclobutyl moieties in place of the pentofuranosyl sugar.

[00184] Base modifications and substitutions

[00185] A nucleic acid may also include nucleobase (often referred to in the art simply as "base") modifications or substitutions. As used herein, "unmodified" or "natural" nucleobases include the purine

bases adenine (A) and guanine (G), and the pyrimidine bases thymine (T), cytosine (C) and uracil (U). Modified nucleobases include other synthetic and natural nucleobases such as 5-methylcytosine (5-me-C), 5-hydroxymethyl cytosine, xanthine, hypoxanthine, 2-aminoadenine, 6-methyl and other alkyl derivatives of adenine and guanine, 2-propyl and other alkyl derivatives of adenine and guanine, 2-thiouracil, 2-thiothymine and 2-thiocytosine, 5-halouracil and cytosine, 5-propynyl (-C=C-CH_3) uracil and cytosine and other alkynyl derivatives of pyrimidine bases, 6-azo uracil, cytosine and thymine, 5-uracil (pseudouracil), 4-thiouracil, 8-halo, 8-amino, 8-thiol, 8-thioalkyl, 8-hydroxyl and other 8-substituted adenines and guanines, 5-halo particularly 5-bromo, 5-trifluoromethyl and other 5-substituted uracils and cytosines, 7-methylguanine and 7-methyladenine, 2-F-adenine, 2-amino-adenine, 8-azaguanine and 8-azaadenine, 7-deazaguanine and 7-deazaadenine and 3-deazaguanine and 3-deazaadenine. Further modified nucleobases include tricyclic pyrimidines such as phenoxazine cytidine(1H-pyrimido(5,4-b)(1,4)benzoxazin-2(3H)-one), phenothiazine cytidine (1H-pyrimido(5,4-b)(1,4)benzothiazin-2(3H)-one), G-clamps such as a substituted phenoxazine cytidine (e.g. 9-(2-aminoethoxy)-H-pyrimido(5,4-(b)(1,4)benzoxazin-2(3H)-one), carbazole cytidine (2H-pyrimido(4,5-b)indol-2-one), pyridoindole cytidine (H-pyrido(3',2':4,5)pyrrolo(2,3-d)pyrimidin-2-one).

[00186] Heterocyclic base moieties may also include those in which the purine or pyrimidine base is replaced with other heterocycles, for example 7-deaza-adenine, 7-deazaguanosine, 2-aminopyridine and 2-pyridone. Further nucleobases include those disclosed in U.S. Pat. No. 3,687,808, those disclosed in The Concise Encyclopedia Of Polymer Science And Engineering, pages 858-859, Kroschwitz, J. I., ed. John Wiley & Sons, 1990, those disclosed by Englisch et al., Angewandte Chemie, International Edition, 1991, 30, 613, and those disclosed by Sanghvi, Y. S., Chapter 15, Antisense Research and Applications, pages 289-302, Crooke, S. T. and Lebleu, B., ed., CRC Press, 1993. Certain of these nucleobases are useful for increasing the binding affinity of an oligomeric compound. These include 5-substituted pyrimidines, 6-azapyrimidines and N-2, N-6 and O-6 substituted purines, including 2-aminopropyladenine, 5-propynyluracil and 5-propynylcytosine. 5-methylcytosine substitutions have been shown to increase nucleic acid duplex stability by 0.6-1.2° C. (Sanghvi et al., eds., Antisense Research and Applications, CRC Press, Boca Raton, 1993, pp. 276-278) and are suitable base substitutions, e.g., when combined with 2'-O-methoxyethyl sugar modifications.

[00187] "Complementary" refers to the capacity for pairing, through base stacking and specific hydrogen bonding, between two sequences comprising naturally or non-naturally occurring (e.g., modified as described above) bases (nucleosides) or analogs thereof. For example, if a base at one position of a nucleic acid is capable of hydrogen bonding with a base at the corresponding position of a target, then the bases are considered to be complementary to each other at that position. Nucleic acids can comprise universal bases, or inert abasic spacers that provide no positive or negative contribution to hydrogen bonding. Base pairings may include both canonical Watson-Crick base pairing and non-Watson-Crick base pairing (e.g., Wobble base pairing and Hoogsteen base pairing). It is understood that for

complementary base pairings, adenosine-type bases (A) are complementary to thymidine-type bases (T) or uracil-type bases (U), that cytosine-type bases (C) are complementary to guanosine-type bases (G), and that universal bases such as such as 3-nitropyrrole or 5-nitroindole can hybridize to and are considered complementary to any A, C, U, or T. Nichols et al., Nature, 1994;369:492-493 and Loakes et al., Nucleic Acids Res., 1994;22:4039-4043. Inosine (I) has also been considered in the art to be a universal base and is considered complementary to any A, C, U, or T. See Watkins and SantaLucia, Nucl. Acids Research, 2005; 33 (19): 6258-6267.

[00188] Conjugates

[00189] Another possible modification of a nucleic acid involves chemically linking to the polynucleotide one or more moieties or conjugates which enhance the activity, cellular distribution or cellular uptake of the oligonucleotide. These moieties or conjugates can include conjugate groups covalently bound to functional groups such as primary or secondary hydroxyl groups. Conjugate groups include, but are not limited to, intercalators, reporter molecules, polyamines, polyamides, polyethylene glycols, polyethers, groups that enhance the pharmacodynamic properties of oligomers, and groups that enhance the pharmacokinetic properties of oligomers. Suitable conjugate groups include, but are not limited to, cholesterols, lipids, phospholipids, biotin, phenazine, folate, phenanthridine, anthraquinone, acridine, fluoresceins, rhodamines, coumarins, and dyes. Groups that enhance the pharmacodynamic properties include groups that improve uptake, enhance resistance to degradation, and/or strengthen sequence-specific hybridization with the target nucleic acid. Groups that enhance the pharmacokinetic properties include groups that improve uptake, distribution, metabolism or excretion of a nucleic acid.

[00190] Conjugate moieties include but are not limited to lipid moieties such as a cholesterol moiety (Letsinger et al., Proc. Natl. Acad. Sci. USA, 1989, 86, 6553-6556), cholic acid (Manoharan et al., Bioorg. Med. Chem. Let., 1994, 4, 1053-1060), a thioether, e.g., hexyl-S-tritylthiol (Manoharan et al., Ann. N.Y. Acad. Sci., 1992, 660, 306-309; Manoharan et al., Bioorg. Med. Chem. Let., 1993, 3, 2765-2770), a thiocholesterol (Oberhauser et al., Nucl. Acids Res., 1992, 20, 533-538), an aliphatic chain, e.g., dodecandiol or undecyl residues (Saison-Behmoaras et al., EMBO J., 1991, 10, 1111-1118; Kabanov et al., FEBS Lett., 1990, 259, 327-330; Svinarchuk et al., Biochimie, 1993, 75, 49-54), a phospholipid, e.g., di-hexadecyl-rac-glycerol or triethylammonium 1,2-di-O-hexadecyl-rac-glycero-3-H-phosphonate (Manoharan et al., Tetrahedron Lett., 1995, 36, 3651-3654; Shea et al., Nucl. Acids Res., 1990, 18, 3777-3783), a polyamine or a polyethylene glycol chain (Manoharan et al., Nucleosides & Nucleotides, 1995, 14, 969-973), or adamantane acetic acid (Manoharan et al., Tetrahedron Lett., 1995, 36, 36513654), a palmityl moiety (Mishra et al., Biochim. Biophys. Acta, 1995, 1264, 229-237), or an octadecylamine or hexylamino-carbonyl-oxycholesterol moiety (Crooke et al., J. Pharmacol. Exp. Ther., 1996, 277, 923-937.\

[00191] A conjugate may include a "Protein Transduction Domain" or PTD (also known as a CPP - cell penetrating peptide), which may refer to a polypeptide, polynucleotide, carbohydrate, or organic or inorganic compound that facilitates traversing a lipid bilayer, micelle, cell membrane, organelle membrane, or vesicle

membrane. A PTD attached to another molecule, which can range from a small polar molecule to a large macromolecule and/or a nanoparticle, facilitates the molecule traversing a membrane, for example going from extracellular space to intracellular space, or cytosol to within an organelle. In some embodiments, a PTD is covalently linked to the amino terminus of an exogenous polypeptide (e.g., a site-directed modifying polypeptide). In some embodiments, a PTD is covalently linked to the carboxyl terminus of an exogenous polypeptide (e.g., a site-directed modifying polypeptide). In some embodiments, a PTD is covalently linked to a nucleic acid (e.g., a guide RNA, a polynucleotide encoding a guide RNA, a polynucleotide encoding a site-directed modifying polypeptide, etc.). Exemplary PTDs include but are not limited to a minimal undecapeptide protein transduction domain (corresponding to residues 47-57 of HIV-1 TAT comprising YGRKKRRQRRR; a polyarginine sequence comprising a number of arginines sufficient to direct entry into a cell (e.g., 3, 4, 5, 6, 7, 8, 9, 10, or 10-50 arginines); a VP22 domain (Zender et al. (2002) *Cancer Gene Ther.* 9(6):489-96); an Drosophila Antennapedia protein transduction domain (Noguchi et al. (2003) *Diabetes* 52(7):1732-1737); a truncated human calcitonin peptide (Trehin et al. (2004) *Pharm. Research* 21:1248-1256); polylysine (Wender et al. (2000) *Proc. Natl. Acad. Sci. USA* 97:13003-13008); RRQRRTSKLMKR; Transport= GWTLNSAGYLLGKINLKALAALAKKIL; KALAWEAKLAKALAKALAKHLAKALAKALKCEA; and RQIKIWFQNRRMKWKK. Exemplary PTDs include but are not limited to, YGRKKRRQRRR; RKKRRQRRR; an arginine homopolymer of from 3 arginine residues to 50 arginine residues; Exemplary PTD domain amino acid sequences include, but are not limited to, any of the following: YGRKKRRQRRR; RKKRRQRR; YARAAARQARA; THRLPRRRRRR; and GGRRARRRRRR. In some embodiments, the PTD is an activatable CPP (ACPP) (Aguilera et al. (2009) *Integr Biol (Camb)* June; 1(5-6): 371-381). ACPPs comprise a polycationic CPP (e.g., Arg9 or "R9") connected via a cleavable linker to a matching polyanion (e.g., Glu9 or "E9"), which reduces the net charge to nearly zero and thereby inhibits adhesion and uptake into cells. Upon cleavage of the linker, the polyanion is released, locally unmasking the polyarginine and its inherent adhesiveness, thus "activating" the ACPP to traverse the membrane.

[00192] Exemplary guide RNAs

[00193] In some embodiments, a guide RNA comprises two separate RNA polynucleotide molecules. The first of the two separate RNA polynucleotide molecules (the activator-RNA) comprises a nucleotide sequence having at least about 60%, at least about 65%, at least about 70%, at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 95%, at least about 98%, at least about 99%, or 100% nucleotide sequence identity over a stretch of at least 8 contiguous, at least 9 contiguous, at least 10 contiguous, at least 11 contiguous, at least 12 contiguous, at least 13 contiguous, at least 14 contiguous or at least 15 contiguous nucleotides to any one of the tracrRNA nucleotide sequences set forth in Supplementary Table S5, or complements thereof. The second of the two separate RNA polynucleotide molecules (the targeter-RNA) comprises a nucleotide sequence having at least about 60%, at least about 65%, at least about 70%, at least about 75%, at least about 80%, at least about 85%,

at least about 90%, at least about 95%, at least about 98%, at least about 99%, or 100% nucleotide sequence identity over a stretch of at least 8 contiguous, at least 9 contiguous, at least 10 contiguous, at least 11 contiguous, at least 12 contiguous, at least 13 contiguous, at least 14 contiguous or at least 15 contiguous nucleotides to the cognate CRISPR repeat nucleotide sequence set forth in Supplementary Table S5, or complements thereof. In some embodiments, a suitable guide RNA is a single-molecule RNA polynucleotide and comprises a first nucleotide sequence having at least about 60%, at least about 65%, at least about 70%, at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 95%, at least about 98%, at least about 99%, or 100% nucleotide sequence identity over a stretch of at least 8 contiguous, at least 9 contiguous, at least 10 contiguous, at least 11 contiguous, at least 12 contiguous, at least 13 contiguous, at least 14 contiguous or at least 15 contiguous nucleotides to any one of the tracrRNA nucleotide sequences set forth in Supplementary Table S5 and a second nucleotide sequence having at least about 60%, at least about 65%, at least about 70%, at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 95%, at least about 98%, at least about 99%, or 100% nucleotide sequence identity over a stretch of at least 8 contiguous, at least 9 contiguous, at least 10 contiguous, at least 11 contiguous, at least 12 contiguous, at least 13 contiguous, at least 14 contiguous or at least 15contiguous nucleotides to the cognate CRISPR repeat nucleotide sequence set forth in Supplementary Table S5, or complements thereof.

[00194] In some embodiments, the single-molecule guide RNAs comprise a DNA-targeting segment and a protein-binding segment complementary thereto, wherein the protein-binding segment comprises a tracrRNA set out in Supplementary Table S5 or wherein the protein-binding segment comprises a tracrRNA at least 80% identical over at least 20 nucleotides to a tracrRNA set out in Supplementary Table S5, or at least about 60%, at least about 65%, at least about 70%, at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 95%, at least about 98%, at least about 99%, or 100% nucleotide sequence identity over a stretch of at least 8 contiguous, at least 9 contiguous, at least 10 contiguous, at least 11 contiguous, at least 12 contiguous, at least 13 contiguous, at least 14 contiguous or at least 15 contiguous nucleotides of any one of the tracrRNA nucleotide sequences set forth in Supplementary Table S5. For example, the protein-binding segment may comprise a tracrRNA at least 70% identical over at least 10 contiguous nucleotides, at least 80% identical over at least 10 contiguous nucleotides, at least 70% identical over at least 11 contiguous nucleotides, at least 80% identical over at least 11 contiguous nucleotides, at least 70% identical over at least 12 contiguous nucleotides, or at least 80% identical over at least 12 contiguous nucleotides.

[00195] In some embodiments, the single-molecule guide RNAs comprise a DNA-targeting segment and a protein-binding segment, wherein the protein-binding segment comprises a tracrRNA set out in Supplementary Table S5 or wherein the protein-binding segment comprises a tracrRNA at least 80% identical over at least 20 nucleotides to a tracrRNA set out in Supplementary Table S5. In some embodiments, the protein-binding segment comprises a CRISPR repeat set out in Supplementary Table S5 that is the CRISPR repeat cognate to the tracrRNA of the protein-binding segment. In some

embodiments, the DNA-targeting segment comprises RNA complementary to a protospacer-like sequence in a target DNA 5' to a PAM sequence. In some embodiments, the tracrRNA and CRISPR repeat are respectively the C. jejuni tracrRNA and its cognate CRISPR repeat set out in Supplementary Table S5 and the PAM sequence is NNNNACA. In some embodiments, the tracrRNA and CRISPR repeat are respectively at least 80% identical to the C. jejuni tracrRNA and its cognate CRISPR repeat set out in Supplementary Table S5 and the PAM sequence is NNNNACA. In some embodiments, the single-molecule guide RNA comprises a sequence that hybridizes to a protospacer-like sequence set out in one of SEQ ID NOs: 801-2701.

[00196] In some embodiments, the double-molecule guide RNAs comprise a targeter-RNA and an activator-RNA complementary thereto, wherein the activator-RNA comprises a tracrRNA set out in Supplementary Table S5 or wherein the activator-RNA comprises a tracrRNA at least 80% identical over at least 20 nucleotides to a tracrRNA set out in Supplementary Table S5. In some embodiments, the double-molecule guide RNA comprises a modified backbone, a non-natural internucleoside linkage, a nucleic acid mimetic, a modified sugar moiety, a base modification, a modification or sequence that provides for modified or regulated stability, a modification or sequence that provides for subcellular tracking, a modification or sequence that provides for tracking, or a modification or sequence that provides for a binding site for a protein or protein complex. In some embodiments, the targeter-RNA comprises a CRISPR repeat set out in Supplementary Table S5. In some embodiments, the targeter-RNA comprises a CRISPR repeat set out in Supplementary Table S5 that is the cognate CRISPR repeat of the tracrRNA of the activator-RNA. In some embodiments, the targeter-RNA further comprises RNA complementary to a protospacer-like sequence in a target DNA 5' to a PAM sequence. In some embodiments, the tracrRNA and CRISPR repeat are respectively the C. jejuni tracrRNA and its cognate CRISPR repeat set out in Supplementary Table S5 and the PAM sequence is NNNNACA. In some embodiments, the tracrRNA and CRISPR repeat are at least 80% identical to respectively the C. jejuni tracrRNA and its cognate CRISPR repeat set out in Supplementary Table S5 and the PAM sequence is NNNNACA. In some embodiments, the double-molecule guide RNA comprises a sequence that hybridizes to a protospacer-like sequence set out in one of SEQ ID NOs: 801-2701.

[00197] Nucleic acids encoding a guide RNA and/or a site-directed modifying polypeptide

[00198] The present disclosure provides a nucleic acid comprising a nucleotide sequence encoding a guide RNA and/or a site-directed modifying polypeptide. In some embodiments, a guide RNA-encoding nucleic acid is an expression vector, e.g., a recombinant expression vector.

[00199] In some embodiments, a method involves contacting a target DNA or introducing into a cell (or a population of cells) one or more nucleic acids comprising nucleotide sequences encoding a guide RNA and/or a site-directed modifying polypeptide. In some embodiments a cell comprising a target DNA is *in vitro*. In some embodiments a cell comprising a target DNA is *in vivo*. Suitable nucleic acids comprising nucleotide sequences encoding a guide RNA and/or a site-directed modifying polypeptide include

expression vectors, where an expression vector comprising a nucleotide sequence encoding a guide RNA and/or a site-directed modifying polypeptide is a "recombinant expression vector."

[00200] In some embodiments, the recombinant expression vector is a viral construct, e.g., a recombinant adeno-associated virus construct (see, e.g., U.S. Patent No. 7,078,387), a recombinant adenoviral construct, a recombinant lentiviral construct, a recombinant retroviral construct, etc.

[00201] Suitable expression vectors include, but are not limited to, viral vectors (e.g. viral vectors based on vaccinia virus; poliovirus; adenovirus (see, e.g., Li et al., Invest Opthalmol Vis Sci 35:2543 2549, 1994; Borras et al., Gene Ther 6:515 524, 1999; Li and Davidson, PNAS 92:7700 7704, 1995; Sakamoto et al., H Gene Ther 5:1088 1097, 1999; WO 94/12649, WO 93/03769; WO 93/19191; WO 94/28938; WO 95/11984 and WO 95/00655); adeno-associated virus (see, e.g., Ali et al., Hum Gene Ther 9:81 86, 1998, Flannery et al., PNAS 94:6916 6921, 1997; Bennett et al., Invest Opthalmol Vis Sci 38:2857 2863, 1997; Jomary et al., Gene Ther 4:683 690, 1997, Rolling et al., Hum Gene Ther 10:641 648, 1999; Ali et al., Hum Mol Genet 5:591 594, 1996; Srivastava in WO 93/09239, Samulski et al., J. Vir. (1989) 63:3822-3828; Mendelson et al., Virol. (1988) 166:154-165; and Flotte et al., PNAS (1993) 90:10613-10617); SV40; herpes simplex virus; human immunodeficiency virus (see, e.g., Miyoshi et al., PNAS 94:10319 23, 1997; Takahashi et al., J Virol 73:7812 7816, 1999); a retroviral vector (e.g., Murine Leukemia Virus, spleen necrosis virus, and vectors derived from retroviruses such as Rous Sarcoma Virus, Harvey Sarcoma Virus, avian leukosis virus, a lentivirus, human immunodeficiency virus, myeloproliferative sarcoma virus, and mammary tumor virus); and the like.

[00202] Numerous suitable expression vectors are known to those of skill in the art, and many are commercially available. The following vectors are provided by way of example; for eukaryotic host cells: pXT1, pSG5 (Stratagene), pSVK3, pBPV, pMSG, and pSVLSV40 (Pharmacia). However, any other vector may be used so long as it is compatible with the host cell. Depending on the host/vector system utilized, any of a number of suitable transcription and translation control elements, including constitutive and inducible promoters, transcription enhancer elements, transcription terminators, etc. may be used in the expression vector (see e.g., Bitter et al. (1987) *Methods in Enzymology,* 153:516-544).

[00203] In some embodiments, a nucleotide sequence encoding a guide RNA and/or a site-directed modifying polypeptide is operably linked to a control element, e.g., a transcriptional control element, such as a promoter. The transcriptional control element may be functional in either a eukaryotic cell, e.g., a mammalian cell; or a prokaryotic cell (e.g., bacterial or archaeal cell). In some embodiments, a nucleotide sequence encoding a guide RNA and/or a site-directed modifying polypeptide is operably linked to multiple control elements that allow expression of the nucleotide sequence encoding a guide RNA and/or a site-directed modifying polypeptide in both prokaryotic and eukaryotic cells.

[00204] Non-limiting examples of suitable eukaryotic promoters (promoters functional in a eukaryotic cell) include those from cytomegalovirus (CMV) immediate early, herpes simplex virus (HSV) thymidine kinase, early and late SV40, long terminal repeats (LTRs) from retrovirus, and mouse metallothionein-I.

Selection of the appropriate vector and promoter is well within the level of ordinary skill in the art. The expression vector may also contain a ribosome binding site for translation initiation and a transcription terminator. The expression vector may also include appropriate sequences for amplifying expression. The expression vector may also include nucleotide sequences encoding protein tags (e.g., 6xHis tag, hemagglutinin tag, green fluorescent protein, etc.) that are fused to the site-directed modifying polypeptide, thus resulting in a chimeric polypeptide.

[00205] In some embodiments, a nucleotide sequence encoding a guide RNA and/or a site-directed modifying polypeptide is operably linked to an inducible promoter. In some embodiments, a nucleotide sequence encoding a guide RNA and/or a site-directed modifying polypeptide is operably linked to a constitutive promoter.

[00206] Methods of introducing a nucleic acid into a host cell are known in the art, and any known method can be used to introduce a nucleic acid (e.g., an expression construct) into a cell. Suitable methods include e.g., viral or bacteriophage infection, transfection, conjugation, protoplast fusion, lipofection, electroporation, calcium phosphate precipitation, polyethyleneimine (PEI)-mediated transfection, DEAE-dextran mediated transfection, liposome-mediated transfection, particle gun technology, calcium phosphate precipitation, direct micro injection, nanoparticle-mediated nucleic acid delivery (see, e.g., Panyam et., al Adv Drug Deliv Rev. 2012 Sep 13. pii: S0169-409X(12)00283-9. doi: 10.1016/j.addr.2012.09.023), and the like.

[00207] Chimeric Polypeptides

[00208] The present disclosure provides a chimeric site-directed modifying polypeptide. A chimeric site-directed modifying polypeptide interacts with (e.g., binds to) a guide RNA (described above). The guide RNA guides the chimeric site-directed modifying polypeptide to a target sequence within target DNA (e.g. a chromosomal sequence or an extrachromosomal sequence, e.g. an episomal sequence, a minicircle sequence, a mitochondrial sequence, a chloroplast sequence, etc.). A chimeric site-directed modifying polypeptide modifies target DNA (e.g., cleavage or methylation of target DNA) and/or a polypeptide associated with target DNA (e.g., methylation or acetylation of a histone tail).

[00209] A chimeric site-directed modifying polypeptide modifies target DNA (e.g., cleavage or methylation of target DNA) and/or a polypeptide associated with target DNA (e.g., methylation or acetylation of a histone tail). A chimeric site-directed modifying polypeptide is also referred to herein as a "chimeric site-directed polypeptide" or a "chimeric RNA binding site-directed modifying polypeptide."

[00210] A chimeric site-directed modifying polypeptide comprises two portions, an RNA-binding portion and an activity portion. A chimeric site-directed modifying polypeptide comprises amino acid sequences that are derived from at least two different polypeptides. A chimeric site-directed modifying polypeptide can comprise modified and/or naturally-occurring polypeptide sequences (e.g., a first amino

acid sequence from a modified or unmodified Cas9 protein; and a second amino acid sequence other than the Cas9 protein).

[00211] RNA-binding portion

[00212] In some cases, the RNA-binding portion of a chimeric site-directed modifying polypeptide is a naturally-occurring polypeptide. In other cases, the RNA-binding portion of a chimeric site-directed modifying polypeptide is not a naturally-occurring molecule (modified, e.g., mutation, deletion, insertion). Naturally-occurring RNA-binding portions of interest are derived from site-directed modifying polypeptides known in the art. For example, SEQ ID NOs: 1-800 provide a non-limiting set of naturally occurring Cas9 endonucleases that can be used as site-directed modifying polypeptides. In some cases, the RNA-binding portion of a chimeric site-directed modifying polypeptide comprises an amino acid sequence having at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 95%, at least about 98%, at least about 99%, or 100%, amino acid sequence identity to the RNA-binding portion of a polypeptide set forth in SEQ ID NOs: 1-800.

[00213] In some cases, the site-directed modifying polypeptide comprises an amino acid sequence having at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 95%, at least about 99%, or 100%, amino acid sequence identity to amino acids 7-166 and/or 731-1003 of SEQ ID NO: 8, or to the corresponding portions in any of the amino acid sequences set forth as SEQ ID NOs: 1-800.

[00214] Activity portion

[00215] In addition to the RNA-binding portion, the chimeric site-directed modifying polypeptide comprises an "activity portion." In some embodiments, the activity portion of a chimeric site-directed modifying polypeptide comprises the naturally-occurring activity portion of a site-directed modifying polypeptide (e.g., Cas9 endonuclease). In other embodiments, the activity portion of a subject chimeric site-directed modifying polypeptide comprises a modified amino acid sequence (e.g., substitution, deletion, insertion) of a naturally-occurring activity portion of a site-directed modifying polypeptide. Naturally-occurring activity portions of interest are derived from site-directed modifying polypeptides known in the art. For example, SEQ ID NOs: 1-800 are a non-limiting set of naturally occurring Cas9 endonucleases that can be used as site-directed modifying polypeptides. The activity portion of a chimeric site-directed modifying polypeptide is variable and may comprise any heterologous polypeptide sequence that may be useful in the methods disclosed herein. In some embodiments, the activity portion of a site-directed modifying polypeptide comprises a portion of a Cas9 ortholog (including, but not limited to, the Cas9 orthologs set out in one of SEQ ID NOs: 1-800) that is at least 90% identical to amino acids 7-166 of SEQ ID NO: 8 and/or at least 90% identical to amino acids 731-1003 of SEQ ID NO: 8. In some embodiments, a chimeric site-directed modifying polypeptide comprises: (i) an RNA-binding portion that interacts with a guide RNA, wherein the guide RNA comprises a nucleotide sequence that is

complementary to a sequence in a target DNA; and (ii) an activity portion that exhibits site-directed enzymatic activity (e.g., activity for DNA methylation, activity for DNA cleavage, activity for histone acetylation, activity for histone methylation, etc.), wherein the site of enzymatic activity is determined by the guide RNA.

[00216] In other embodiments, a chimeric site-directed modifying polypeptide comprises: (i) an RNA-binding portion that interacts with a guide RNA, wherein the guide RNA comprises a nucleotide sequence that is complementary to a sequence in a target DNA; and (ii) an activity portion that modulates transcription within the target DNA (e.g., to increase or decrease transcription), wherein the site of modulated transcription within the target DNA is determined by the guide RNA.

[00217] In some cases, the activity portion of a chimeric site-directed modifying polypeptide has enzymatic activity that modifies target DNA (e.g., nuclease activity, methyltransferase activity, demethylase activity, DNA repair activity, DNA damage activity, deamination activity, dismutase activity, alkylation activity, depurination activity, oxidation activity, pyrimidine dimer forming activity, integrase activity, transposase activity, recombinase activity, polymerase activity, ligase activity, helicase activity, photolyase activity or glycosylase activity).

[00218] In other cases, the activity portion of a chimeric site-directed modifying polypeptide has enzymatic activity (e.g., methyltransferase activity, demethylase activity, acetyltransferase activity, deacetylase activity, kinase activity, phosphatase activity, ubiquitin ligase activity, deubiquitinating activity, adenylation activity, deadenylation activity, SUMOylating activity, deSUMOylating activity, ribosylation activity, deribosylation activity, myristoylation activity or demyristoylation activity) that modifies a polypeptide associated with target DNA (e.g., a histone).

[00219] In some cases, the activity portion of a chimeric site-directed modifying polypeptide exhibits enzymatic activity (described above). In other cases, the activity portion of a chimeric site-directed modifying polypeptide modulates transcription of the target DNA (described above). The activity portion of a chimeric site-directed modifying polypeptide is variable and may comprise any heterologous polypeptide sequence that may be useful in the methods disclosed herein.

[00220] Exemplary chimeric site-directed modifying polypeptides

[00221] In some embodiments, the activity portion of the chimeric site-directed modifying polypeptide comprises a modified form of the Cas9 protein, including modified forms of any of the Cas9 orthologs described herein, such as SEQ ID NOs: 1-800). In some instances, the modified form of the Cas9 protein comprises an amino acid change (e.g., deletion, insertion, or substitution) that reduces the naturally-occurring nuclease activity of the Cas9 protein. For example, in some instances, the modified form of the Cas9 protein has less than 50%, less than 40%, less than 30%, less than 20%, less than 10%, less than

5%, or less than 1% of the nuclease activity of the corresponding wild-type Cas9 polypeptide. In some cases, the modified form of the Cas9 polypeptide has no substantial nuclease activity.

[00222] In some embodiments, the modified form of the Cas9 polypeptide is a D10A (aspartate to alanine at amino acid position 10 of SEQ ID NO:8) mutation (or the corresponding mutation of any of the proteins presented in SEQ ID NOs: 1-800) that can cleave the complementary strand of the target DNA but has reduced ability to cleave the non-complementary strand of the target DNA. In some embodiments, the modified form of the SEQ ID NO: 8 Cas9 polypeptide is a H840A (histidine to alanine at amino acid position 840) mutation (or the corresponding mutation of any of the proteins set forth as SEQ ID NOs: 1-800) that can cleave the non-complementary strand of the target DNA but has reduced ability to cleave the complementary strand of the target DNA. In some embodiments, the modified form of the SEQ ID NO: 8 Cas9 polypeptide harbors both the D10A and the H840A mutations (or the corresponding mutations of any of the proteins set forth as SEQ ID NOs: 1-800) such that the polypeptide has a reduced ability to cleave both the complementary and the non-complementary strands of the target DNA. Other residues can be mutated to achieve the above effects (i.e. inactivate one or the other nuclease portions). As non-limiting examples, S. pyogenes Cas9 residues D10, G12, G17, E762, H840, N863, H982, H983, A984, D986, and/or A987 of SEQ ID NO: 8 (or the corresponding mutations of any of the proteins set forth as SEQ ID NOs: 1-800) can be altered (i.e., substituted). Also, mutations other than alanine substitutions are contemplated.

[00223] In some some embodiments, a modified Cas9 endonuclease comprises one or more mutations corresponding to S. pyogenes Cas9 mutation E762A, HH983AA or D986A in SEQ ID NO: 8. In some embodiments, the modified Cas 9 endonuclease further comprises one or more mutations corresponding to S. pyogenes Cas9 mutation D10A, H840A, G12A, G17A, N854A, N863A, N982A or A984A in SEQ ID NO: 8. For example, the modified Cas9 endonuclease may comprise a variant at least about 75% identical to any of SEQ ID NOs: 1-800 that comprises one or more mutations corresponding to a mutation E762A, HH983AA or D986A in SEQ ID NO: 8; and/or one or more mutations corresponding to a mutation D10A, H840A, G12A, G17A, N854A, N863A, N982A or A984A in SEQ ID NO: 8. In some embodiments, such a variant comprises a region at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 95%, at least about 99% or 100% amino acid sequence identity to the regions corresponding to amino acids 7-166 and/or 731-1003 of SEQ ID NO: 8.

[00224] **Table 1.** Table 1 lists four motifs that are present in Cas9 sequences from various species. The amino acids listed here are from the Cas9 from *S. pyogenes* (SEQ ID NO:8).

	Motif	**Amino acids (residue #s)**	Highly conserved
	RuvC-like I	IGLDIGTNSVGWAVI (7-21)	D10, G12, G17
	RuvC-like II	IVIEMARE (759-766)	E762

	HNH-motif	DVDHIVPQSFLKDDSIDNKVLTRSDKN (837-863)	H840, N854, N863
	RuvC-like II	HHAHDAYL (982-989)	H982, H983, A984, D986, A987

[00225] In some cases, the chimeric site-directed modifying polypeptide comprises an amino acid sequence having at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 95%, at least about 99% or 100% amino acid sequence identity to amino acids 7-166 and/or 731-1003 of SEQ ID NO: 8, or to the corresponding portions in any of the amino acid sequences set forth as SEQ ID NOs: 1-800. In some cases, the chimeric site-directed modifying polypeptide comprises 4 motifs (as listed in Table 1), each with amino acid sequences having at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 95%, at least about 99% or 100% amino acid sequence identity to each of the 4 motifs listed in Table 1, or to the corresponding portions in any of the amino acid sequences set forth as SEQ ID NOs: 1-800. In some cases, the chimeric site-directed modifying polypeptide comprises amino acid sequences having at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 95%, at least about 99% or 100% amino acid sequence identity to amino acids 7-166 and/or 731-1003 of SEQ ID NO: 8, or to the corresponding portions in any of the amino acid sequences set forth as SEQ ID NOs: 1-800.

[00226] In some embodiments, the activity portion of the site-directed modifying polypeptide comprises a heterologous polypeptide that has DNA-modifying activity and/or transcription factor activity and/or DNA-associated polypeptide-modifying activity. In some cases, a heterologous polypeptide replaces a portion of the Cas9 polypeptide that provides nuclease activity. In other embodiments, a site-directed modifying polypeptide comprises both a portion of the Cas9 polypeptide that normally provides nuclease activity (and that portion can be fully active or can instead be modified to have less than 100% of the corresponding wild-type activity) and a heterologous polypeptide. In other words, in some cases, a chimeric site-directed modifying polypeptide is a fusion polypeptide comprising both the portion of the Cas9 polypeptide that normally provides nuclease activity and the heterologous polypeptide. In other cases, a chimeric site-directed modifying polypeptide is a fusion polypeptide comprising a modified variant of the activity portion of the Cas9 polypeptide (e.g., amino acid change, deletion, insertion) and a heterologous polypeptide. In yet other cases, a chimeric site-directed modifying polypeptide is a fusion polypeptide comprising a heterologous polypeptide and the RNA-binding portion of a naturally-occurring or a modified site-directed modifying polypeptide.

[00227] For example, in a chimeric Cas9 protein, a naturally-occurring (or modified, e.g., mutation, deletion, insertion) bacterial Cas9 polypeptide may be fused to a heterologous polypeptide sequence (i.e. a polypeptide sequence from a protein other than Cas9 or a polypeptide sequence from another organism).

The heterologous polypeptide sequence may exhibit an activity (e.g., enzymatic activity) that will also be exhibited by the chimeric Cas9 protein (e.g., methyltransferase activity, acetyltransferase activity, kinase activity, ubiquitinating activity, etc.). A heterologous nucleic acid sequence may be linked to another nucleic acid sequence (e.g., by genetic engineering) to generate a chimeric nucleotide sequence encoding a chimeric polypeptide. In some embodiments, a chimeric Cas9 polypeptide is generated by fusing a Cas9 polypeptide (e.g., wild type Cas9 or a Cas9 variant, e.g., a Cas9 with reduced or inactivated nuclease activity) with a heterologous sequence that provides for subcellular localization (e.g., a nuclear localization signal (NLS) for targeting to the nucleus; a mitochondrial localization signal for targeting to the mitochondria; a chloroplast localization signal for targeting to a chloroplast; an ER retention signal; and the like). In some embodiments, the heterologous sequence can provide a tag for ease of tracking or purification (e.g., a fluorescent protein, e.g., green fluorescent protein (GFP), YFP, RFP, CFP, mCherry, tdTomato, and the like; a HIS tag, e.g., a 6XHis tag; a hemagglutinin (HA) tag; a FLAG tag; a Myc tag; and the like). In some embodiments, the heterologous sequence can provide for increased or decreased stability. In some embodiments, the heterologous sequence can provide a binding domain (e.g., to provide the ability of a chimeric Cas9 polypeptide to bind to another protein of interest, e.g., a DNA or histone modifying protein, a transcription factor or transcription repressor, a recruiting protein, etc.).

[00228] Nucleic acid encoding a chimeric site-directed modifying polypeptide

[00229] The present disclosure provides a nucleic acid comprising a nucleotide sequence encoding a chimeric site-directed modifying polypeptide. In some embodiments, the nucleic acid comprising a nucleotide sequence encoding a chimeric site-directed modifying polypeptide is an expression vector, e.g., a recombinant expression vector.

[00230] In some embodiments, a method involves contacting a target DNA or introducing into a cell (or a population of cells) one or more nucleic acids comprising a chimeric site-directed modifying polypeptide. Suitable nucleic acids comprising nucleotide sequences encoding a chimeric site-directed modifying polypeptide include expression vectors, where an expression vector comprising a nucleotide sequence encoding a chimeric site-directed modifying polypeptide is a "recombinant expression vector."

[00231] In some embodiments, the recombinant expression vector is a viral construct, e.g., a recombinant adeno-associated virus construct (see, e.g., U.S. Patent No. 7,078,387), a recombinant adenoviral construct, a recombinant lentiviral construct, etc.

[00232] Suitable expression vectors include, but are not limited to, viral vectors (e.g. viral vectors based on vaccinia virus; poliovirus; adenovirus (see, e.g., Li et al., Invest Opthalmol Vis Sci 35:2543 2549, 1994; Borras et al., Gene Ther 6:515 524, 1999; Li and Davidson, PNAS 92:7700 7704, 1995; Sakamoto et al., H Gene Ther 5:1088 1097, 1999; WO 94/12649, WO 93/03769; WO 93/19191; WO 94/28938; WO 95/11984 and WO 95/00655); adeno-associated virus (see, e.g., Ali et al., Hum Gene Ther 9:81 86, 1998, Flannery et al., PNAS 94:6916 6921, 1997; Bennett et al., Invest Opthalmol Vis Sci

38:2857 2863, 1997; Jomary et al., Gene Ther 4:683 690, 1997, Rolling et al., Hum Gene Ther 10:641 648, 1999; Ali et al., Hum Mol Genet 5:591 594, 1996; Srivastava in WO 93/09239, Samulski et al., J. Vir. (1989) 63:3822-3828; Mendelson et al., Virol. (1988) 166:154-165; and Flotte et al., PNAS (1993) 90:10613-10617); SV40; herpes simplex virus; human immunodeficiency virus (see, e.g., Miyoshi et a1., PNAS 94:10319 23, 1997; Takahashi et al., J Virol 73:7812 7816, 1999); a retroviral vector (e.g., Murine Leukemia Virus, spleen necrosis virus, and vectors derived from retroviruses such as Rous Sarcoma Virus, Harvey Sarcoma Virus, avian leukosis virus, a lentivirus, human immunodeficiency virus, myeloproliferative sarcoma virus, and mammary tumor virus); and the like.

[00233] Numerous suitable expression vectors are known to those of skill in the art, and many are commercially available. The following vectors are provided by way of example; for eukaryotic host cells: pXT1, pSG5 (Stratagene), pSVK3, pBPV, pMSG, and pSVLSV40 (Pharmacia). However, any other vector may be used so long as it is compatible with the host cell.

[00234] Depending on the host/vector system utilized, any of a number of suitable transcription and translation control elements, including constitutive and inducible promoters, transcription enhancer elements, transcription terminators, etc. may be used in the expression vector (see e.g., Bitter et al. (1987) *Methods in Enzymology,* 153:516-544).

[00235] In some embodiments, a nucleotide sequence encoding a chimeric site-directed modifying polypeptide is operably linked to a control element, e.g., a transcriptional control element, such as a promoter. The transcriptional control element may be functional in either a eukaryotic cell, e.g., a mammalian cell; or a prokaryotic cell (e.g., bacterial or archaeal cell). In some embodiments, a nucleotide sequence encoding a chimeric site-directed modifying polypeptide is operably linked to multiple control elements that allow expression of the nucleotide sequence encoding a chimeric site-directed modifying polypeptide in both prokaryotic and eukaryotic cells.

[00236] Non-limiting examples of suitable eukaryotic promoters (promoters functional in a eukaryotic cell) include those from cytomegalovirus (CMV) immediate early, herpes simplex virus (HSV) thymidine kinase, early and late SV40, long terminal repeats (LTRs) from retrovirus, and mouse metallothionein-I. Selection of the appropriate vector and promoter is well within the level of ordinary skill in the art. The expression vector may also contain a ribosome binding site for translation initiation and a transcription terminator. The expression vector may also include appropriate sequences for amplifying expression. The expression vector may also include nucleotide sequences encoding protein tags (e.g., 6xHis tag, hemagglutinin (HA) tag, a fluorescent protein (e.g., a green fluorescent protein; a yellow fluorescent protein, etc.), etc.) that are fused to the chimeric site-directed modifying polypeptide.

[00237] In some embodiments, a nucleotide sequence encoding a chimeric site-directed modifying polypeptide is operably linked to an inducible promoter (e.g., heat shock promoter, Tetracycline-regulated promoter, Steroid-regulated promoter, Metal-regulated promoter, estrogen receptor-regulated promoter,

etc.). In some embodiments, a nucleotide sequence encoding a chimeric site-directed modifying polypeptide is operably linked to a spatially restricted and/or temporally restricted promoter (e.g., a tissue specific promoter, a cell type specific promoter, etc.). In some embodiments, a nucleotide sequence encoding a chimeric site-directed modifying polypeptide is operably linked to a constitutive promoter.

[00238] Methods of introducing a nucleic acid into a host cell are known in the art, and any known method can be used to introduce a nucleic acid (e.g., an expression construct) into a stem cell or progenitor cell. Suitable methods include, include e.g., viral or bacteriophage infection, transfection, conjugation, protoplast fusion, lipofection, electroporation, calcium phosphate precipitation, polyethyleneimine (PEI)-mediated transfection, DEAE-dextran mediated transfection, liposome-mediated transfection, particle gun technology, calcium phosphate precipitation, direct micro injection, nanoparticle-mediated nucleic acid delivery (see, e.g., Panyam et., al Adv Drug Deliv Rev. 2012 Sep 13. pii: 50169-409X(12)00283-9. doi: 10.1016/j.addr.2012.09.023), and the like.

[00239] Methods

[00240] The present disclosure provides methods for modifying a target DNA and/or a target DNA-associated polypeptide. Generally, a method involves contacting a target DNA with a complex (a "targeting complex"), which complex comprises a guide RNA and a site-directed modifying polypeptide.

[00241] As discussed above, a guide RNA and a site-directed modifying polypeptide form a complex. The guide RNA provides target specificity to the complex by comprising a nucleotide sequence that is complementary to a sequence of a target DNA. The site-directed modifying polypeptide of the complex provides the site-specific activity. In some embodiments, a complex modifies a target DNA, leading to, for example, DNA cleavage, DNA methylation, DNA damage, DNA repair, etc. In other embodiments, a complex modifies a target polypeptide associated with target DNA (e.g., a histone, a DNA-binding protein, etc.), leading to, for example, histone methylation, histone acetylation, histone ubiquitination, and the like. The target DNA may be, for example, naked DNA *in vitro,* chromosomal DNA in cells *in vitro,* chromosomal DNA in cells *in vivo,* etc.

[00242] In some cases, the site-directed modifying polypeptide exhibits nuclease activity that cleaves target DNA at a target DNA sequence defined by the region of complementarity between the guide RNA and the target DNA. In some cases, when the site-directed modifying polypeptide is a Cas9 or Cas9 related polypeptide, site-specific cleavage of the target DNA occurs at locations determined by both (i) base-pairing complementarity between the guide RNA and the target DNA; and (ii) a short motif [referred to as the protospacer adjacent motif (PAM)] in the target DNA. In some embodiments (e.g., when Cas9 from S. pyogenes is used), the PAM sequence of the non-complementary strand is 5'-XGG-3', where X is any DNA nucleotide and X is immediately 3' of the target sequence of the non-complementary strand of the target DNA. As such, the PAM sequence of the complementary strand is 5'-CCY-3', where Y is any DNA nucleotide and Y is immediately 5' of the target sequence of the complementary strand of the target

DNA (where the PAM of the non-complementary strand is 5'-GGG-3' and the PAM of the complementary strand is 5'-CCC-3'). In some such embodiments, X and Y can be complementary and the X-Y base pair can be any basepair (e.g., X=C and Y=G; X=G and Y=C; X=A and Y=T, X=T and Y=A).

[00243] In some cases, different Cas9 proteins (i.e., Cas9 proteins from various species) may be advantageous to use in the various provided methods in order to capitalize on various enzymatic characteristics of the different Cas9 proteins (e.g., for different PAM sequence preferences; for increased or decreased enzymatic activity; for an increased or decreased level of cellular toxicity; to change the balance between NHEJ, homology-directed repair, single strand breaks, double strand breaks, etc.).

[00244] Cas9 proteins from various species (see SEQ ID NOs: 1-800) may require different PAM sequences in the target DNA. Thus, for a particular Cas9 protein of choice, the PAM sequence requirement may be different than the 5'-XGG-3' sequence described above. The present disclosure, for example, provides a C. jejuni PAM sequence NNNNACA; P. multocida PAM sequences GNNNCNNA or NNNNC; an F. novicida PAM sequence NG; an S. thermophilus** PAM sequence NNAAAAW; an L. innocua PAM sequence NGG; and an S. dysgalactiae PAM sequence NGG.

[00245] Exemplary methods provided that take advantage of characteristics of Cas9 orthologs include the following.

[00246] A method for manipulating DNA in a cell, comprising contacting the DNA with a Cas9 ortholog-guideRNA complex, wherein the complex comprises: (a) a cognate guide RNA for a first Cas9 endonuclease from a cluster in Supplementary Table S2 and (b) a second Cas9 endonuclease from the cluster that is exchangeable with preserved high cleavage efficiency with the first endonuclease and shares at least 80% identity with the first endonuclease over 80% of their length. In some embodiments, the guide is a single-molecule guide RNA. In some embodiments, the guide RNA is a double-molecule guide RNA.In some embodiments, the first Cas9 endonuclease is from S. pyogenes and the second Cas9 endonuclease is from S. mutans. In some embodiments, the first Cas9 endonuclease is from S. theromophilus* and the second Cas9 endonuclease is from S. mutans. In some embodiments, the first Cas9 endonuclease is from N. meningitidis and the second Cas9 endonuclease is from P. multocida.

[00247] A method for manipulating DNA in a cell, comprising contacting the DNA with a Cas9 ortholog-guideRNA complex, wherein the complex comprises: (a) a cognate guide RNA of a first Cas9 endonuclease from a cluster in Supplementary Table S6 and (b) an Cas9 endonuclease from a cluster in Supplementary Table S6 that is exchangeable with lowered cleavage efficiency with the first endonuclease and shares at least 50% amino acid sequence identity with the first endonuclease over 70% of their length. In some embodiments, the guide is a single-molecule guide RNA. In some embodiments, the guide RNA is a double-molecule guide RNA.In some embodiments, the first Cas9 endonuclease is from C. Jejuni and the second Cas9 endonuclease is from P. multocida. In some

embodiments, the first Cas9 endonuclease is from N. meningitidis and the second Cas9 endonuclease is from P. multocida.

[00248] A method for manipulating DNA in a cell, comprising contacting the DNA with two or more Cas9-guideRNA complexes, wherein each Cas9-guideRNA complex comprises: (a) a Cas9 endonuclease from a different cluster in Supplementary Table S6 exhibiting less than 50% amino acid sequence identity with the other endonucleases of the method over 70% of their length, and (b) a guide RNA specifically complexed with each Cas9 endonuclease. In some embodiments, the guide is a single-molecule guide RNA. In some embodiments, the guide RNA is a double-molecule guide RNA.In some embodiments, the Cas9 endonucleases are from F. novicida and S. pyogenes. In some embodiments, the Cas9 endonucleases are from N. meningitidis and S. mutans. In some embodiments, the S. thermophilus* and S. thermophilus** Cas9 endonucleases.

[00249] Many Cas9 orthologs from a wide variety of species have been identified herein. All identified Cas9 orthologs have the same domain architecture with a central HNH endonuclease domain and a split RuvC/RNaseH domain. Cas9 proteins share four key motifs with a conserved architecture. Motifs 1, 2, and 4 are RuvC like motifs while motif 3 is an HNH-motif. In some cases, a suitable site-directed modifying polypeptide comprises an amino acid sequence having four motifs, each of motifs 1-4 having at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 95%, at least about 99% or 100% amino acid sequence identity to the motifs 1-4 of the Cas9 amino acid sequence depicted in Table 1), or to the corresponding portions in any of the amino acid sequences set forth in SEQ ID NOs: 1-800. In some cases, a suitable site-directed modifying polypeptide comprises an amino acid sequence having at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 95%, at least about 99% or 100% amino acid sequence identity to amino acids 7-166 and/or 731-1003 of SEQ ID NO: 8, or to the corresponding portions in any of the amino acid sequences set forth as SEQ ID NOs: 1-800.

[00250] The nuclease activity cleaves target DNA to produce double strand breaks. These breaks are then repaired by the cell in one of two ways: non-homologous end joining, and homology-directed repair. In non-homologous end joining (NHEJ), the double-strand breaks are repaired by direct ligation of the break ends to one another. As such, no new nucleic acid material is inserted into the site, although some nucleic acid material may be lost, resulting in a deletion. In homology-directed repair, a donor polynucleotide with homology to the cleaved target DNA sequence is used as a template for repair of the cleaved target DNA sequence, resulting in the transfer of genetic information from the donor polynucleotide to the target DNA. As such, new nucleic acid material may be inserted/copied into the site. In some cases, a target DNA is contacted with a donor polynucleotide. In some cases, a donor polynucleotide is introduced into a cell. The modifications of the target DNA due to NHEJ and/or homology-directed repair lead to, for example, gene correction, gene replacement, gene tagging, transgene insertion, nucleotide deletion, gene disruption, gene mutation, sequence replacement, etc.

Accordingly, cleavage of DNA by a site-directed modifying polypeptide may be used to delete nucleic acid material from a target DNA sequence (e.g., to disrupt a gene that makes cells susceptible to infection (e.g. the CCRS or CXCR4 gene, which makes T cells susceptible to HIV infection), to remove disease-causing trinucleotide repeat sequences in neurons, to create gene knockouts and mutations as disease models in research, etc.) by cleaving the target DNA sequence and allowing the cell to repair the sequence in the absence of an exogenously provided donor polynucleotide. Thus, the methods can be used to knock out a gene (resulting in complete lack of transcription or altered transcription) or to knock in genetic material into a locus of choice in the target DNA.

[00251] Alternatively, if a guide RNA and a site-directed modifying polypeptide are coadministered to cells with a donor polynucleotide sequence that includes at least a segment with homology to the target DNA sequence, the subject methods may be used to add, i.e. insert or replace, nucleic acid material to a target DNA sequence (e.g. to "knock in" a nucleic acid that encodes for a protein, an siRNA, an miRNA, etc.), to add a tag (e.g., 6xHis, a fluorescent protein (e.g., a green fluorescent protein; a yellow fluorescent protein, etc.), hemagglutinin (HA), FLAG, etc.), to add a regulatory sequence to a gene (e.g. promoter, polyadenylation signal, internal ribosome entry sequence (IRES), 2A peptide, start codon, stop codon, splice signal, localization signal, etc.), to modify a nucleic acid sequence (e.g., introduce a mutation), and the like. As such, a complex comprising a guide RNA and a site-directed modifying polypeptide is useful in any *in vitro* or *in vivo* application in which it is desirable to modify DNA in a site-specific, i.e. "targeted", way, for example gene knock-out, gene knock-in, gene editing, gene tagging, sequence replacement, etc., as used in, for example, gene therapy, e.g. to treat a disease or as an antiviral, antipathogenic, or anticancer therapeutic, the production of genetically modified organisms in agriculture, the large scale production of proteins by cells for therapeutic, diagnostic, or research purposes, the induction of iPS cells, biological research, the targeting of genes of pathogens for deletion or replacement, etc.

[00252] In some embodiments, the site-directed modifying polypeptide comprises a modified form of the Cas9 protein. In some instances, the modified form of the Cas9 protein comprises an amino acid change (e.g., deletion, insertion, or substitution) that reduces the naturally-occurring nuclease activity of the Cas9 protein. For example, in some instances, the modified form of the Cas9 protein has less than 50%, less than 40%, less than 30%, less than 20%, less than 10%, less than 5%, or less than 1% of the nuclease activity of the corresponding wild-type Cas9 polypeptide. In some cases, the modified form of the Cas9 polypeptide has no substantial nuclease activity. When a site-directed modifying polypeptide is a modified form of the Cas9 polypeptide that has no substantial nuclease activity, it can be referred to as "dCas9."

[00253] In some embodiments, the modified form of the Cas9 polypeptide is a D10A (aspartate to alanine at amino acid position 10 of SEQ ID NO:8) mutation (or the corresponding mutation of any of the proteins set forth as SEQ ID NOs: 1-800) that can cleave the complementary strand of the target DNA but has reduced ability to cleave the non-complementary strand of the target DNA (thus resulting in a single

strand break (SSB) instead of a DSB). In some embodiments, the modified form of the Cas9 polypeptide is a H840A (histidine to alanine at amino acid position 840 of SEQ ID NO:8) mutation (or the corresponding mutation of any of the proteins set forth as SEQ ID NOs: 1-800) that can cleave the non-complementary strand of the target DNA but has reduced ability to cleave the complementary strand of the target DNA (thus resulting in a single strand break (SSB) instead of a DSB). The use of the D10A or H840A variant of SEQ ID NO: 8 Cas9 (or the corresponding mutations in any of the proteins set forth as SEQ ID NOs: 1-800) can alter the expected biological outcome because the non-homologous end joining (NHEJ) is much more likely to occur when DSBs are present as opposed to SSBs. Thus, in some cases where one wishes to reduce the likelihood of DSB (and therefore reduce the likelihood of NHEJ), a D10A or H840A variant of Cas9 can be used. Other residues can be mutated to achieve the same effect (i.e. inactivate one or the other nuclease portions). As non-limiting examples, SEQ ID NO: 8 S. pyogenes Cas9 residues D10, G12, G17, E762, H840, N863, H982, H983, A984, D986, and/or A987 (or the corresponding mutations of any of the proteins set forth as SEQ ID NOs: 1-800) can be altered (i.e., substituted). Also, mutations other than alanine substitutions are contemplated. In some embodiments when a site-directed polypeptide (e.g., site-directred modifying polypeptide) has reduced catalytic activity (e.g., when a SEQ ID NO: 8 Cas9 protein has a D10, G12, G17, E762, H840, N863, H982, H983, A984, D986, and/or a A987 mutation, e.g., D 10A, G12A, G17A, E762A, H840A, N863A, H982A, H983A, A984A, and/or D986A), the polypeptide can still bind to target DNA in a site-specific manner (because it is still guided to a target DNA sequence by a guide RNA) as long as it retains the ability to interact with the guide RNA.

[00254] In some embodiments, the modified form of the SEQ ID NO: 8 Cas9 polypeptide harbors both the D10A and the H840A mutations (or the corresponding mutations of any of the proteins set forth as SEQ ID NOs: 1-800) such that the polypeptide has a reduced ability to cleave both the complementary and the non-complementary strands of the target DNA (i.e., the variant can have no substantial nuclease activity). Other residues can be mutated to achieve the same effect (i.e. inactivate one or the other nuclease portions). As non-limiting examples, SEQ ID NO: 8 residues D10, G12, G17, E762, H840, N863, H982, H983, A984, D986, and/or A987 (or the corresponding mutations of any of the proteins set forth as SEQ ID NOs: 1-800) can be altered (i.e., substituted). Also, mutations other than alanine substitutions are contemplated.

[00255] In some embodiments, the site-directed modifying polypeptide comprises a heterologous sequence (e.g., a fusion). In some embodiments, a heterologous sequence can provide for subcellular localization of the site-directed modifying polypeptide (e.g., a nuclear localization signal (NLS) for targeting to the nucleus; a mitochondrial localization signal for targeting to the mitochondria; a chloroplast localization signal for targeting to a chloroplast; a ER retention signal; and the like). In some embodiments, a heterologous sequence can provide a tag for ease of tracking or purification (e.g., a fluorescent protein, e.g., green fluorescent protein (GFP), YFP, RFP, CFP, mCherry, tdTomato, and the

like; a his tag, e.g., a 6XHis tag; a hemagglutinin (HA) tag; a FLAG tag; a Myc tag; and the like). In some embodiments, the heterologous sequence can provide for increased or decreased stability.

[00256] In some embodiments, a site-directed modifying polypeptide can be codon-optimized. This type of optimization is known in the art and entails the mutation of foreign-derived DNA to mimic the codon preferences of the intended host organism or cell while encoding the same protein. Thus, the codons are changed, but the encoded protein remains unchanged. For example, if the intended target cell was a human cell, a human codon-optimized Cas9 (or variant, e.g., enzymatically inactive variant) would be a suitable site-directed modifying polypeptide. Any suitable site-directed modifying polypeptide (e.g., any Cas9 such as any of the sequences set forth in SEQ ID NOs: 1-800) can be codon optimized. As another non-limiting example, if the intended host cell were a mouse cell, than a mouse codon-optimized Cas9 (or variant, e.g., enzymatically inactive variant) would be a suitable site-directed modifying polypeptide. While codon optimization is not required, it is acceptable and may be preferable in certain cases.

[00257] Polyadenylation signals can also be chosen to optimize expression in the intended host.

[00258] In some embodiments, a guide RNA and a site-directed modifying polypeptide are used as an inducible system for shutting off gene expression in bacterial cells. In some cases, nucleic acids encoding an appropriate guide RNA and/or an appropriate site-directed polypeptide are incorporated into the chromosome of a target cell and are under control of an inducible promoter. When the guide RNA and/or the site-directed polypeptide are induced, the target DNA is cleaved (or otherwise modified) at the location of interest (e.g., a target gene on a separate plasmid), when both the guide RNA and the site-directed modifying polypeptide are present and form a complex. As such, in some cases, bacterial expression strains are engineered to include nucleic acid sequences encoding an appropriate site-directed modifying polypeptide in the bacterial genome and/or an appropriate guide RNA on a plasmid (e.g., under control of an inducible promoter), allowing experiments in which the expression of any targeted gene (expressed from a separate plasmid introduced into the strain) could be controlled by inducing expression of the guide RNA and the site-directed polypeptide.

[00259] In some cases, the site-directed modifying polypeptide has enzymatic activity that modifies target DNA in ways other than introducing double strand breaks. Enzymatic activity of interest that may be used to modify target DNA (e.g., by fusing a heterologous polypeptide with enzymatic activity to a site-directed modifying polypeptide, thereby generating a chimeric site-directed modifying polypeptide) includes, but is not limited methyltransferase activity, demethylase activity, DNA repair activity, DNA damage activity, deamination activity, dismutase activity, alkylation activity, depurination activity, oxidation activity, pyrimidine dimer forming activity, integrase activity, transposase activity, recombinase activity, polymerase activity, ligase activity, helicase activity, photolyase activity or glycosylase activity). Methylation and demethylation is recognized in the art as an important mode of epigenetic gene

regulation while DNA damage and repair activity is essential for cell survival and for proper genome maintenance in response to environmental stresses.

[00260] As such, the methods herein find use in the epigenetic modification of target DNA and may be employed to control epigenetic modification of target DNA at any location in a target DNA by genetically engineering the desired complementary nucleic acid sequence into the DNA-targeting segment of a guide RNA. The methods herein also find use in the intentional and controlled damage of DNA at any desired location within the target DNA. The methods herein also find use in the sequence-specific and controlled repair of DNA at any desired location within the target DNA. Methods to target DNA-modifying enzymatic activities to specific locations in target DNA find use in both research and clinical applications.

[00261] In some cases, the site-directed modifying polypeptide has activity that modulates the transcription of target DNA (e.g., in the case of a chimeric site-directed modifying polypeptide, etc.). In some cases, a chimeric site-directed modifying polypeptides comprising a heterologous polypeptide that exhibits the ability to increase or decrease transcription (e.g., transcriptional activator or transcription repressor polypeptides) is used to increase or decrease the transcription of target DNA at a specific location in a target DNA, which is guided by the DNA-targeting segment of the guide RNA. Examples of source polypeptides for providing a chimeric site-directed modifying polypeptide with transcription modulatory activity include, but are not limited to light-inducible transcription regulators, small molecule/drug-responsive transcription regulators, transcription factors, transcription repressors, etc. In some cases, the method is used to control the expression of a targeted coding-RNA (protein-encoding gene) and/or a targeted non-coding RNA (e.g., tRNA, rRNA, snoRNA, siRNA, miRNA, long ncRNA, etc.). In some cases, the site-directed modifying polypeptide has enzymatic activity that modifies a polypeptide associated with DNA (e.g. histone). In some embodiments, the enzymatic activity is methyltransferase activity, demethylase activity, acetyltransferase activity, deacetylase activity, kinase activity, phosphatase activity, ubiquitin ligase activity (i.e., ubiquitination activity), deubiquitinating activity, adenylation activity, deadenylation activity, SUMOylating activity, deSUMOylating activity, ribosylation activity, deribosylation activity, myristoylation activity, demyristoylation activity glycosylation activity (e.g., from GlcNAc transferase) or deglycosylation activity. The enzymatic activities listed herein catalyze covalent modifications to proteins. Such modifications are known in the art to alter the stability or activity of the target protein (e.g., phosphorylation due to kinase activity can stimulate or silence protein activity depending on the target protein). Of particular interest as protein targets are histones. Histone proteins are known in the art to bind DNA and form complexes known as nucleosomes. Histones can be modified (e.g., by methylation, acetylation, ubuitination, phosphorylation) to elicit structural changes in the surrounding DNA, thus controlling the accessibility of potentially large portions of DNA to interacting factors such as transcription factors, polymerases and the like. A single histone can be modified in many different ways and in many different combinations (e.g., trimethylation of lysine 27 of histone 3, H3K27, is associated with DNA regions of repressed transcription while trimethylation of lysine 4 of histone 3, H3K4, is associated with DNA regions of active transcription). Thus, a site-directed modifying polypeptide with

histone-modifying activity finds use in the site specific control of DNA structure and can be used to alter the histone modification pattern in a selected region of target DNA. Such methods find use in both research and clinical applications.

[00262] In some embodiments, multiple guide RNAs are used simultaneously to simultaneously modify different locations on the same target DNA or on different target DNAs. In some embodiments, two or more guide RNAs target the same gene or transcript or locus. In some embodiments, two or more guide RNAs target different unrelated loci. In some embodiments, two or more guide RNAs target different, but related loci.

[00263] In some cases, the site-directed modifying polypeptide is provided directly as a protein. As one non-limiting example, fungi (e.g., yeast) can be transformed with exogenous protein and/or nucleic acid using spheroplast transformation (see Kawai et al., Bioeng Bugs. 2010 Nov-Dec;1(6):395-403 : "Transformation of Saccharomyces cerevisiae and other fungi: methods and possible underlying mechanism"; and Tanka et al., Nature. 2004 Mar 18;428(6980):323-8: "Conformational variations in an infectious protein determine prion strain differences"; both of which are herein incorporated by reference in their entirety). Thus, a site-directed modifying polypeptide (e.g., Cas9) can be incorporated into a spheroplast (with or without nucleic acid encoding a guide RNA and with or without a donor polynucleotide) and the spheroplast can be used to introduce the content into a yeast cell. A site-directed modifying polypeptide can be introduced into a cell (provided to the cell) by any convenient method; such methods are known to those of ordinary skill in the art. As another non-limiting example, a site-directed modifying polypeptide can be injected directly into a cell (e.g., with or without nucleic acid encoding a guide RNA and with or without a donor polynucleotide), e.g., a cell of a zebrafish embryo, the pronucleus of a fertilized mouse oocyte, etc.

[00264] Target cells of interest

[00265] In some of the above applications, the methods may be employed to induce DNA cleavage, DNA modification, and/or transcriptional modulation in mitotic or post-mitotic cells *in vivo* and/or *ex vivo* and/or *in vitro* (e.g., to produce genetically modified cells that can be reintroduced into an individual). Because the guide RNA provide specificity by hybridizing to target DNA, a mitotic and/or post-mitotic cell of interest in the disclosed methods may include a cell from any organism (e.g. a bacterial cell, an archaeal cell, a cell of a single-cell eukaryotic organism, a plant cell, an algal cell, e.g., *Botryococcus braunii, Chlamydomonas reinhardtii, Nannochloropsis gaditana, Chlorella pyrenoidosa, Sargassum patens C. Agardh,* and the like, a fungal cell (e.g., a yeast cell), an animal cell, a cell from an invertebrate animal (e.g. fruit fly, cnidarian, echinoderm, nematode, etc.), a cell from a vertebrate animal (e.g., fish, amphibian, reptile, bird, mammal), a cell from a mammal, a cell from a rodent, a cell from a primate, a cell from a human, etc.).

[00266] Any type of cell may be of interest (e.g. a stem cell, e.g. an embryonic stem (ES) cell, an induced pluripotent stem (iPS) cell, a germ cell; a somatic cell, e.g. a fibroblast, a hematopoietic cell, a neuron, a muscle cell, a bone cell, a hepatocyte, a pancreatic cell; an in vitro or in vivo embryonic cell of an embryo at any stage, e.g., a 1-cell, 2-cell, 4-cell, 8-cell, etc. stage zebrafish embryo; etc.). Cells may be from established cell lines or they may be primary cells, where "primary cells", "primary cell lines", and "primary cultures" are used interchangeably herein to refer to cells and cells cultures that have been derived from a and allowed to grow in vitro for a limited number of passages, i.e. splittings, of the culture. For example, primary cultures are cultures that may have been passaged 0 times, 1 time, 2 times, 4 times, 5 times, 10 times, or 15 times, but not enough times go through the crisis stage. Typically, the primary cell lines of the present invention are maintained for fewer than 10 passages in vitro. Target cells are in many embodiments unicellular organisms, or are grown in culture.

[00267] If the cells are primary cells, they may be harvest from an individual by any convenient method. For example, leukocytes may be conveniently harvested by apheresis, leukocytapheresis, density gradient separation, etc., while cells from tissues such as skin, muscle, bone marrow, spleen, liver, pancreas, lung, intestine, stomach, etc. are most conveniently harvested by biopsy. An appropriate solution may be used for dispersion or suspension of the harvested cells. Such solution will generally be a balanced salt solution, e.g. normal saline, phosphate-buffered saline (PBS), Hank's balanced salt solution, etc., conveniently supplemented with fetal calf serum or other naturally occurring factors, in conjunction with an acceptable buffer at low concentration, generally from 5-25 mM. Convenient buffers include HEPES, phosphate buffers, lactate buffers, etc. The cells may be used immediately, or they may be stored, frozen, for long periods of time, being thawed and capable of being reused. In such cases, the cells will usually be frozen in 10% DMSO, 50% serum, 40% buffered medium, or some other such solution as is commonly used in the art to preserve cells at such freezing temperatures, and thawed in a manner as commonly known in the art for thawing frozen cultured cells.

[00268] Nucleic acids encoding a guide RNA and/or a site-directed modifying polypeptide

[00269] In some embodiments, a method involves contacting a target DNA or introducing into a cell (or a population of cells) one or more nucleic acids comprising nucleotide sequences encoding a guide RNA and/or a site-directed modifying polypeptide and/or a donor polynucleotide. Suitable nucleic acids comprising nucleotide sequences encoding a guide RNA and/or a site-directed modifying polypeptide include expression vectors, where an expression vector comprising a nucleotide sequence encoding a guide RNA and/or a site-directed modifying polypeptide is a "recombinant expression vector."

[00270] In some embodiments, the recombinant expression vector is a viral construct, e.g., a recombinant adeno-associated virus construct (see, e.g., U.S. Patent No. 7,078,387), a recombinant adenoviral construct, a recombinant lentiviral construct, etc.

[00271] Suitable expression vectors include, but are not limited to, viral vectors (e.g. viral vectors based on vaccinia virus; poliovirus; adenovirus (see, e.g., Li et al., Invest Opthalmol Vis Sci 35:2543 2549, 1994; Borras et al., Gene Ther 6:515 524, 1999; Li and Davidson, PNAS 92:7700 7704, 1995; Sakamoto et al., H Gene Ther 5:1088 1097, 1999; WO 94/12649, WO 93/03769; WO 93/19191; WO 94/28938; WO 95/11984 and WO 95/00655); adeno-associated virus (see, e.g., Ali et al., Hum Gene Ther 9:81 86, 1998, Flannery et al., PNAS 94:6916 6921, 1997; Bennett et al., Invest Opthalmol Vis Sci 38:2857 2863, 1997; Jomary et al., Gene Ther 4:683 690, 1997, Rolling et al., Hum Gene Ther 10:641 648, 1999; Ali et al., Hum Mol Genet 5:591 594, 1996; Srivastava in WO 93/09239, Samulski et al., J. Vir. (1989) 63:3822-3828; Mendelson et al., Virol. (1988) 166:154-165; and Flotte et al., PNAS (1993) 90:10613-10617); SV40; herpes simplex virus; human immunodeficiency virus (see, e.g., Miyoshi et al., PNAS 94:10319 23, 1997; Takahashi et al., J Virol 73:7812 7816, 1999); a retroviral vector (e.g., Murine Leukemia Virus, spleen necrosis virus, and vectors derived from retroviruses such as Rous Sarcoma Virus, Harvey Sarcoma Virus, avian leukosis virus, a lentivirus, human immunodeficiency virus, myeloproliferative sarcoma virus, and mammary tumor virus); and the like.

[00272] Numerous suitable expression vectors are known to those of skill in the art, and many are commercially available. The following vectors are provided by way of example; for eukaryotic host cells: pXT1, pSG5 (Stratagene), pSVK3, pBPV, pMSG, and pSVLSV40 (Pharmacia). However, any other vector may be used so long as it is compatible with the host cell.

[00273] In some embodiments, a nucleotide sequence encoding a guide RNA and/or a site-directed modifying polypeptide is operably linked to a control element, e.g., a transcriptional control element, such as a promoter. The transcriptional control element may be functional in either a eukaryotic cell, e.g., a mammalian cell, or a prokaryotic cell (e.g., bacterial or archaeal cell). In some embodiments, a nucleotide sequence encoding a guide RNA and/or a site-directed modifying polypeptide is operably linked to multiple control elements that allow expression of the nucleotide sequence encoding a guide RNA and/or a site-directed modifying polypeptide in both prokaryotic and eukaryotic cells.

[00274] Depending on the host/vector system utilized, any of a number of suitable transcription and translation control elements, including constitutive and inducible promoters, transcription enhancer elements, transcription terminators, etc. may be used in the expression vector (e.g., U6 promoter, H1 promoter, etc.; see above) (see e.g., Bitter et al. (1987) *Methods in Enzymology,* 153:516-544).

[00275] In some embodiments, a guide RNA and/or a site-directed modifying polypeptide can be provided as RNA. In such cases, the guide RNA and/or the RNA encoding the site-directed modifying polypeptide can be produced by direct chemical synthesis or may be transcribed *in vitro* from a DNA encoding the guide RNA. Methods of synthesizing RNA from a DNA template are well known in the art. In some cases, the guide RNA and/or the RNA encoding the site-directed modifying polypeptide will be synthesized *in vitro* using an RNA polymerase enzyme (e.g., T7 polymerase, T3 polymerase, SP6

polymerase, etc.). Once synthesized, the RNA may directly contact a target DNA or may be introduced into a cell by any of the well-known techniques for introducing nucleic acids into cells (e.g., microinjection, electroporation, transfection, etc).

[00276] Nucleotides encoding a guide RNA (introduced either as DNA or RNA) and/or a site-directed modifying polypeptide (introduced as DNA or RNA) and/or a donor polynucleotide may be provided to the cells using well-developed transfection techniques; see, e.g. Angel and Yanik (2010) PLoS ONE 5(7): e 11756, and the commercially available TransMessenger® reagents from Qiagen, StemfectTM RNA Transfection Kit from Stemgent, and TransIT®-mRNA Transfection Kit from Mims Bio LLC. See also Beumer et al. (2008) Efficient gene targeting in Drosophila by direct embryo injection with zinc-finger nucleases. PNAS 105(50):19821-19826. Alternatively, nucleic acids encoding a guide RNA and/or a site-directed modifying polypeptide and/or a chimeric site-directed modifying polypeptide and/or a donor polynucleotide may be provided on DNA vectors. Many vectors, e.g. plasmids, cosmids, minicircles, phage, viruses, etc., useful for transferring nucleic acids into target cells are available. The vectors comprising the nucleic acid(s) may be maintained episomally, e.g. as plasmids, minicircle DNAs, viruses such cytomegalovirus, adenovirus, etc., or they may be integrated into the target cell genome, through homologous recombination or random integration, e.g. retrovirus-derived vectors such as MMLV, HIV-1, ALV, etc.

[00277] Vectors may be provided directly to the cells. In other words, the cells are contacted with vectors comprising the nucleic acid encoding guide RNA and/or a site-directed modifying polypeptide and/or a chimeric site-directed modifying polypeptide and/or a donor polynucleotide such that the vectors are taken up by the cells. Methods for contacting cells with nucleic acid vectors that are plasmids, including electroporation, calcium chloride transfection, microinjection, and lipofection are well known in the art. For viral vector delivery, the cells are contacted with viral particles comprising the nucleic acid encoding a guide RNA and/or a site-directed modifying polypeptide and/or a chimeric site-directed modifying polypeptide and/or a donor polynucleotide. Retroviruses, for example, lentiviruses, are particularly suitable to the method of the invention. Commonly used retroviral vectors are "defective", i.e. unable to produce viral proteins required for productive infection. Rather, replication of the vector requires growth in a packaging cell line. To generate viral particles comprising nucleic acids of interest, the retroviral nucleic acids comprising the nucleic acid are packaged into viral capsids by a packaging cell line. Different packaging cell lines provide a different envelope protein (ecotropic, amphotropic or xenotropic) to be incorporated into the capsid, this envelope protein determining the specificity of the viral particle for the cells (ecotropic for murine and rat; amphotropic for most mammalian cell types including human, dog and mouse; and xenotropic for most mammalian cell types except murine cells). The appropriate packaging cell line may be used to ensure that the cells are targeted by the packaged viral particles. Methods of introducing the retroviral vectors comprising the nucleic acid encoding the reprogramming factors into packaging cell lines and of collecting the viral particles that are generated by

the packaging lines are well known in the art. Nucleic acids can also introduced by direct micro-injection (e.g., injection of RNA into a zebrafish embryo).

[00278] Vectors used for providing the nucleic acids encoding guide RNA and/or a site-directed modifying polypeptide and/or a chimeric site-directed modifying polypeptide and/or a donor polynucleotide to the cells will typically comprise suitable promoters for driving the expression, that is, transcriptional activation, of the nucleic acid of interest. In other words, the nucleic acid of interest will be operably linked to a promoter. This may include ubiquitously acting promoters, for example, the CMV-13-actin promoter, or inducible promoters, such as promoters that are active in particular cell populations or that respond to the presence of drugs such as tetracycline. By transcriptional activation, it is intended that transcription will be increased above basal levels in the target cell by at least about 10 fold, by at least about 100 fold, more usually by at least about 1000 fold. In addition, vectors used for providing a guide RNA and/or a site-directed modifying polypeptide and/or a chimeric site-directed modifying polypeptide and/or a donor polynucleotide to the cells may include nucleic acid sequences that encode for selectable markers in the target cells, so as to identify cells that have taken up the guide RNA and/or a site-directed modifying polypeptide and/or a chimeric site-directed modifying polypeptide and/or a donor polynucleotide.

[00279] A guide RNA and/or a site-directed modifying polypeptide and/or a chimeric site-directed modifying polypeptide may instead be used to contact DNA or introduced into cells as RNA. Methods of introducing RNA into cells are known in the art and may include, for example, direct injection, transfection, or any other method used for the introduction of DNA. A site-directed modifying polypeptide may instead be provided to cells as a polypeptide. Such a polypeptide may optionally be fused to a polypeptide domain that increases solubility of the product. The domain may be linked to the polypeptide through a defined protease cleavage site, e.g. a TEV sequence, which is cleaved by TEV protease. The linker may also include one or more flexible sequences, e.g. from 1 to 10 glycine residues. In some embodiments, the cleavage of the fusion protein is performed in a buffer that maintains solubility of the product, e.g. in the presence of from 0.5 to 2 M urea, in the presence of polypeptides and/or polynucleotides that increase solubility, and the like. Domains of interest include endosomolytic domains, e.g. influenza HA domain; and other polypeptides that aid in production, e.g. IF2 domain, GST domain, GRPE domain, and the like. The polypeptide may be formulated for improved stability. For example, the peptides may be PEGylated, where the polyethyleneoxy group provides for enhanced lifetime in the blood stream.

[00280] Additionally or alternatively, the site-directed modifying polypeptide may be fused to a polypeptide permeant domain to promote uptake by the cell. A number of permeant domains are known in the art and may be used in the non-integrating polypeptides of the present invention, including peptides, peptidomimetics, and non-peptide carriers. For example, a permeant peptide may be derived from the third alpha helix of Drosophila melanogaster transcription factor Antennapaedia, referred to as penetratin, which comprises the amino acid sequence RQIKIWFQNRRMKWKK. As another example, the

permeant peptide comprises the HIV-1 tat basic region amino acid sequence, which may include, for example, amino acids 49-57 of naturally-occurring tat protein. Other permeant domains include poly-arginine motifs, for example, the region of amino acids 34-56 of HIV-1 rev protein, nona-arginine, octa-arginine, and the like. (See, for example, Futaki et al. (2003) Curr Protein Pept Sci. 2003 Apr; 4(2): 87-9 and 446; and Wender et al. (2000) Proc. Natl. Acad. Sci. U.S.A 2000 Nov. 21; 97(24):13003-8; published U.S. Patent applications 20030220334; 20030083256; 20030032593; and 20030022831, herein specifically incorporated by reference for the teachings of translocation peptides and peptoids). The nona-arginine (R9) sequence is one of the more efficient PTDs that have been characterized (Wender et al. 2000; Uemura et al. 2002). The site at which the fusion is made may be selected in order to optimize the biological activity, secretion or binding characteristics of the polypeptide. The optimal site will be determined by routine experimentation.

[00281] A site-directed modifying polypeptide may be produced *in* vitro or by eukaryotic cells or by prokaryotic cells, and it may be further processed by unfolding, e.g. heat denaturation, DTT reduction, etc. and may be further refolded, using methods known in the art.

[00282] Modifications of interest that do not alter primary sequence include chemical derivatization of polypeptides, e.g., acylation, acetylation, carboxylation, amidation, etc. Also included are modifications of glycosylation, e.g. those made by modifying the glycosylation patterns of a polypeptide during its synthesis and processing or in further processing steps; e.g. by exposing the polypeptide to enzymes which affect glycosylation, such as mammalian glycosylating or deglycosylating enzymes. Also embraced are sequences that have phosphorylated amino acid residues, e.g. phosphotyrosine, phosphoserine, or phosphothreonine.

[00283] Also included in the invention are guide RNAs and site-directed modifying polypeptides that have been modified using ordinary molecular biological techniques and synthetic chemistry so as to improve their resistance to proteolytic degradation, to change the target sequence specificity, to optimize solubility properties, to alter protein activity (e.g., transcription modulatory activity, enzymatic activity, etc) or to render them more suitable as a therapeutic agent. Analogs of such polypeptides include those containing residues other than naturally occurring L-amino acids, e.g. D-amino acids or non-naturally occurring synthetic amino acids. D-amino acids may be substituted for some or all of the amino acid residues. The site-directed modifying polypeptides may be prepared by in vitro synthesis, using conventional methods as known in the art. Various commercial synthetic apparatuses are available, for example, automated synthesizers by Applied Biosystems, Inc., Beckman, etc. By using synthesizers, naturally occurring amino acids may be substituted with unnatural amino acids. The particular sequence and the manner of preparation will be determined by convenience, economics, purity required, and the like.

[00284] If desired, various groups may be introduced into the peptide during synthesis or during expression, which allow for linking to other molecules or to a surface. Thus cysteines can be used to make thioethers, histidines for linking to a metal ion complex, carboxyl groups for forming amides or esters, amino groups for forming amides, and the like.

[00285] The site-directed modifying polypeptides may also be isolated and purified in accordance with conventional methods of recombinant synthesis. A lysate may be prepared of the expression host and the lysate purified using HPLC, exclusion chromatography, gel electrophoresis, affinity chromatography, or other purification technique. For the most part, the compositions which are used will comprise at least 20% by weight of the desired product, more usually at least about 75% by weight, preferably at least about 95% by weight, and for therapeutic purposes, usually at least about 99.5% by weight, in relation to contaminants related to the method of preparation of the product and its purification. Usually, the percentages will be based upon total protein. To induce DNA cleavage and recombination, or any desired modification to a target DNA, or any desired modification to a polypeptide associated with target DNA, the guide RNA and/or the site-directed modifying polypeptide and/or the donor polynucleotide, whether they be introduced as nucleic acids or polypeptides, are provided to the cells for about 30 minutes to about 24 hours, e.g., 1 hour, 1.5 hours, 2 hours, 2.5 hours, 3 hours, 3.5 hours 4 hours, 5 hours, 6 hours, 7 hours, 8 hours, 12 hours, 16 hours, 18 hours, 20 hours, or any other period from about 30 minutes to about 24 hours, which may be repeated with a frequency of about every day to about every 4 days, e.g., every 1.5 days, every 2 days, every 3 days, or any other frequency from about every day to about every four days. The agent(s) may be provided to the cells one or more times, e.g. one time, twice, three times, or more than three times, and the cells allowed to incubate with the agent(s) for some amount of time following each contacting event e.g. 16-24 hours, after which time the media is replaced with fresh media and the cells are cultured further. In cases in which two or more different targeting complexes are provided to the cell (e.g., two different guide RNAs that are complementary to different sequences within the same or different target DNA), the complexes may be provided simultaneously (e.g. as two polypeptides and/or nucleic acids), or delivered simultaneously. Alternatively, they may be provided consecutively, e.g. the targeting complex being provided first, followed by the second targeting complex, etc. or vice versa.

[00286] Typically, an effective amount of the guide RNA and/or site-directed modifying polypeptide and/or donor polynucleotide is provided to the target DNA or cells to induce target modification. An effective amount of the guide RNA and/or site-directed modifying polypeptide and/or donor polynucleotide is the amount to induce a 2-fold increase or more in the amount of target modification observed between two homologous sequences relative to a negative control, e.g. a cell contacted with an empty vector or irrelevant polypeptide. That is to say, an effective amount or dose of the guide RNA and/or site-directed modifying polypeptide and/or donor polynucleotide will induce a 2-fold increase, a 3-fold increase, a 4-fold increase or more in the amount of target modification observed at a target DNA region, in some instances a 5-fold increase, a 6-fold increase or more, sometimes a 7-fold or 8-fold increase or more in the amount of

recombination observed, e.g. an increase of 10-fold, 50-fold, or 100-fold or more, in some instances, an increase of 200-fold, 500-fold, 700-fold, or 1000-fold or more, e.g. a 5000-fold, or 10,000-fold increase in the amount of recombination observed. The amount of target modification may be measured by any convenient method. For example, a silent reporter construct comprising complementary sequence to the targeting segment (targeting sequence) of the guide RNA flanked by repeat sequences that, when recombined, will reconstitute a nucleic acid encoding an active reporter may be cotransfected into the cells, and the amount of reporter protein assessed after contact with the guide RNA and/or site-directed modifying polypeptide and/or donor polynucleotide, e.g. 2 hours, 4 hours, 8 hours, 12 hours, 24 hours, 36 hours, 48 hours, 72 hours or more after contact with the guide RNA and/or site-directed modifying polypeptide and/or donor polynucleotide. As another, more sensitivity assay, for example, the extent of recombination at a genomic DNA region of interest comprising target DNA sequences may be assessed by PCR or Southern hybridization of the region after contact with a guide RNA and/or site-directed modifying polypeptide and/or donor polynucleotide, e.g. 2 hours, 4 hours, 8 hours, 12 hours, 24 hours, 36 hours, 48 hours, 72 hours or more after contact with the guide RNA and/or site-directed modifying polypeptide and/or donor polynucleotide.

[00287] Contacting the cells with a guide RNA and/or site-directed modifying polypeptide and/or donor polynucleotide may occur in any culture media and under any culture conditions that promote the survival of the cells. For example, cells may be suspended in any appropriate nutrient medium that is convenient, such as Iscove's modified DMEM or RPMI 1640, supplemented with fetal calf serum or heat inactivated goat serum (about 5-10%), L-glutamine, a thiol, particularly 2-mercaptoethanol, and antibiotics, e.g. penicillin and streptomycin. The culture may contain growth factors to which the cells are responsive. Growth factors, as defined herein, are molecules capable of promoting survival, growth and/or differentiation of cells, either in culture or in the intact tissue, through specific effects on a transmembrane receptor. Growth factors include polypeptides and non-polypeptide factors. Conditions that promote the survival of cells are typically permissive of nonhomologous end joining and homology-directed repair. In applications in which it is desirable to insert a polynucleotide sequence into a target DNA sequence, a polynucleotide comprising a donor sequence to be inserted is also provided to the cell. By a "donor sequence" or "donor polynucleotide" it is meant a nucleic acid sequence to be inserted at the cleavage site induced by a site-directed modifying polypeptide. The donor polynucleotide will contain sufficient homology to a genomic sequence at the cleavage site, e.g. 70%, 80%, 85%, 90%, 95%, or 100% homology with the nucleotide sequences flanking the cleavage site, e.g. within about 50 bases or less of the cleavage site, e.g. within about 30 bases, within about 15 bases, within about 10 bases, within about 5 bases, or immediately flanking the cleavage site, to support homology-directed repair between it and the genomic sequence to which it bears homology. Approximately 25, 50, 100, or 200 nucleotides, or more than 200 nucleotides, of sequence homology between a donor and a genomic sequence (or any integral value between 10 and 200 nucleotides, or more) will support homology-directed repair. Donor sequences can be

of any length, e.g. 10 nucleotides or more, 50 nucleotides or more, 100 nucleotides or more, 250 nucleotides or more, 500 nucleotides or more, 1000 nucleotides or more, 5000 nucleotides or more, etc.

[00288] The donor sequence is typically not identical to the genomic sequence that it replaces. Rather, the donor sequence may contain at least one or more single base changes, insertions, deletions, inversions or rearrangements with respect to the genomic sequence, so long as sufficient homology is present to support homology-directed repair. In some embodiments, the donor sequence comprises a non-homologous sequence flanked by two regions of homology, such that homology-directed repair between the target DNA region and the two flanking sequences results in insertion of the non-homologous sequence at the target region. Donor sequences may also comprise a vector backbone containing sequences that are not homologous to the DNA region of interest and that are not intended for insertion into the DNA region of interest. Generally, the homologous region(s) of a donor sequence will have at least 50% sequence identity to a genomic sequence with which recombination is desired. In certain embodiments, 60%, 70%, 80%, 90%, 95%, 98%, 99%, or 99.9% sequence identity is present. Any value between 1% and 100% sequence identity can be present, depending upon the length of the donor polynucleotide. The donor sequence may comprise certain sequence differences as compared to the genomic sequence, e.g. restriction sites, nucleotide polymorphisms, selectable markers (e.g., drug resistance genes, fluorescent proteins, enzymes etc.), etc., which may be used to assess for successful insertion of the donor sequence at the cleavage site or in some cases may be used for other purposes (e.g., to signify expression at the targeted genomic locus). In some cases, if located in a coding region, such nucleotide sequence differences will not change the amino acid sequence, or will make silent amino acid changes (i.e., changes which do not affect the structure or function of the protein). Alternatively, these sequences differences may include flanking recombination sequences such as FLPs, loxP sequences, or the like, that can be activated at a later time for removal of the marker sequence.

[00289] The donor sequence may be provided to the cell as single-stranded DNA, single-stranded RNA, double-stranded DNA, or double-stranded RNA. It may be introduced into a cell in linear or circular form. If introduced in linear form, the ends of the donor sequence may be protected (e.g., from exonucleolytic degradation) by methods known to those of skill in the art. For example, one or more dideoxynucleotide residues are added to the 3' terminus of a linear molecule and/or self-complementary oligonucleotides are ligated to one or both ends. See, for example, Chang et al. (1987) Proc. Natl. Acad Sci USA 84:4959-4963; Nehls et al. (1996) Science 272:886-889. Additional methods for protecting exogenous polynucleotides from degradation include, but are not limited to, addition of terminal amino group(s) and the use of modified internucleotide linkages such as, for example, phosphorothioates, phosphoramidates, and 0-methyl ribose or deoxyribose residues. As an alternative to protecting the termini of a linear donor sequence, additional lengths of sequence may be included outside of the regions of homology that can be degraded without impacting recombination. A donor sequence can be introduced into a cell as part of a vector molecule having additional sequences such as, for example, replication

origins, promoters and genes encoding antibiotic resistance. Moreover, donor sequences can be introduced as naked nucleic acid, as nucleic acid complexed with an agent such as a liposome or poloxamer, or can be delivered by viruses (e.g., adenovirus, AAV), as described above for nucleic acids encoding a guide RNA and/or site-directed modifying polypeptide and/or donor polynucleotide.

[00290] Following the methods described above, a DNA region of interest may be cleaved and modified, i.e. "genetically modified", ex vivo. In some embodiments, as when a selectable marker has been inserted into the DNA region of interest, the population of cells may be enriched for those comprising the genetic modification by separating the genetically modified cells from the remaining population. Prior to enriching, the "genetically modified" cells may make up only about 1% or more (e.g., 2% or more, 3% or more, 4% or more, 5% or more, 6% or more, 7% or more, 8% or more, 9% or more, 10% or more, 15% or more, or 20% or more) of the cellular population. Separation of "genetically modified" cells may be achieved by any convenient separation technique appropriate for the selectable marker used. For example, if a fluorescent marker has been inserted, cells may be separated by fluorescence activated cell sorting, whereas if a cell surface marker has been inserted, cells may be separated from the heterogeneous population by affinity separation techniques, e.g. magnetic separation, affinity chromatography, "panning" with an affinity reagent attached to a solid matrix, or other convenient technique. Techniques providing accurate separation include fluorescence activated cell sorters, which can have varying degrees of sophistication, such as multiple color channels, low angle and obtuse light scattering detecting channels, impedance channels, etc. The cells may be selected against dead cells by employing dyes associated with dead cells (e.g. propidium iodide). Any technique may be employed which is not unduly detrimental to the viability of the genetically modified cells. Cell compositions that are highly enriched for cells comprising modified DNA are achieved in this manner. By "highly enriched", it is meant that the genetically modified cells will be 70% or more, 75% or more, 80% or more, 85% or more, 90% or more of the cell composition, for example, about 95% or more, or 98% or more of the cell composition. In other words, the composition may be a substantially pure composition of genetically modified cells.

[00291] Genetically modified cells produced by the methods described herein may be used immediately. Alternatively, the cells may be frozen at liquid nitrogen temperatures and stored for long periods of time, being thawed and capable of being reused. In such cases, the cells will usually be frozen in 10% dimethylsulfoxide (DMSO), 50% serum, 40% buffered medium, or some other such solution as is commonly used in the art to preserve cells at such freezing temperatures, and thawed in a manner as commonly known in the art for thawing frozen cultured cells.

[00292] The genetically modified cells may be cultured in vitro under various culture conditions. The cells may be expanded in culture, i.e. grown under conditions that promote their proliferation. Culture medium may be liquid or semi-solid, e.g. containing agar, methylcellulose, etc. The cell population may be suspended in an appropriate nutrient medium, such as Iscove's modified DMEM or RPMI 1640,

normally supplemented with fetal calf serum (about 5-10%), L-glutamine, a thiol, particularly 2-mercaptoethanol, and antibiotics, e.g. penicillin and streptomycin. The culture may contain growth factors to which the regulatory T cells are responsive. Growth factors, as defined herein, are molecules capable of promoting survival, growth and/or differentiation of cells, either in culture or in the intact tissue, through specific effects on a transmembrane receptor. Growth factors include polypeptides and non-polypeptide factors.

[00293] Cells that have been genetically modified in this way may be transplanted to a subject for purposes such as gene therapy, e.g. to treat a disease or as an antiviral, antipathogenic, or anticancer therapeutic, for the production of genetically modified organisms in agriculture, or for biological research. The subject may be a neonate, a juvenile, or an adult. Of particular interest are mammalian subjects. Mammalian species that may be treated with the present methods include canines and felines; equines; bovines; ovines; etc. and primates, particularly humans. Animal models, particularly small mammals (e.g. mouse, rat, guinea pig, hamster, lagomorpha (e.g., rabbit), etc.) may be used for experimental investigations.

[00294] Cells may be provided to the subject alone or with a suitable substrate or matrix, e.g. to support their growth and/or organization in the tissue to which they are being transplanted. Usually, at least $1x10^3$ cells will be administered, for example $5x10^3$ cells, $1x10^4$ cells, $5x10^4$ cells, $1x10^5$ cells, 1×10^6 cells or more. The cells may be introduced to the subject via any of the following routes: parenteral, subcutaneous, intravenous, intracranial, intraspinal, intraocular, or into spinal fluid. The cells may be introduced by injection, catheter, or the like. Examples of methods for local delivery, that is, delivery to the site of injury, include, e.g. through an Ommaya reservoir, e.g. for intrathecal delivery (see e.g. US Patent Nos. 5,222,982 and 5385582, incorporated herein by reference); by bolus injection, e.g. by a syringe, e.g. into a joint; by continuous infusion, e.g. by cannulation, e.g. with convection (see e.g. US Application No. 20070254842, incorporated herein by reference); or by implanting a device upon which the cells have been reversably affixed (see e.g. US Application Nos. 20080081064 and 20090196903, incorporated herein by reference). Cells may also be introduced into an embryo (e.g., a blastocyst) for the purpose of generating a transgenic animal (e.g., a transgenic mouse).

[00295] The number of administrations of treatment to a subject may vary. Introducing the genetically modified cells into the subject may be a one-time event; but in certain situations, such treatment may elicit improvement for a limited period of time and require an on-going series of repeated treatments. In other situations, multiple administrations of the genetically modified cells may be required before an effect is observed. The exact protocols depend upon the disease or condition, the stage of the disease and parameters of the individual subject being treated.

[00296] In other aspects of the disclosure, the guide RNA and/or site-directed modifying polypeptide and/or donor polynucleotide are employed to modify cellular DNA in vivo, again for purposes such as

gene therapy, e.g. to treat a disease or as an antiviral, antipathogenic, or anticancer therapeutic, for the production of genetically modified organisms in agriculture, or for biological research. In these in vivo embodiments, a guide RNA and/or site-directed modifying polypeptide and/or donor polynucleotide are administered directly to the individual. A guide RNA and/or site-directed modifying polypeptide and/or donor polynucleotide may be administered by any of a number of well-known methods in the art for the administration of peptides, small molecules and nucleic acids to a subject. A guide RNA and/or site-directed modifying polypeptide and/or donor polynucleotide can be incorporated into a variety of formulations. More particularly, a guide RNA and/or site-directed modifying polypeptide and/or donor polynucleotide of the present invention can be formulated into pharmaceutical compositions by combination with appropriate pharmaceutically acceptable carriers or diluents.

[00297] Pharmaceutical preparations are compositions that include one or more a guide RNA and/or site-directed modifying polypeptide and/or donor polynucleotide present in a pharmaceutically acceptable vehicle. "Pharmaceutically acceptable vehicles" may be vehicles approved by a regulatory agency of the Federal or a state government or listed in the U.S. Pharmacopeia or other generally recognized pharmacopeia for use in mammals, such as humans. The term "vehicle" refers to a diluent, adjuvant, excipient, or carrier with which a compound of the invention is formulated for administration to a mammal. Such pharmaceutical vehicles can be lipids, e.g. liposomes, e.g. liposome dendrimers; liquids, such as water and oils, including those of petroleum, animal, vegetable or synthetic origin, such as peanut oil, soybean oil, mineral oil, sesame oil and the like, saline; gum acacia, gelatin, starch paste, talc, keratin, colloidal silica, urea, and the like. In addition, auxiliary, stabilizing, thickening, lubricating and coloring agents may be used. Pharmaceutical compositions may be formulated into preparations in solid, semi-solid, liquid or gaseous forms, such as tablets, capsules, powders, granules, ointments, solutions, suppositories, injections, inhalants, gels, microspheres, and aerosols. As such, administration of the a guide RNA and/or site-directed modifying polypeptide and/or donor polynucleotide can be achieved in various ways, including oral, buccal, rectal, parenteral, intraperitoneal, intradermal, transdermal, intratracheal, intraocular, etc., administration. The active agent may be systemic after administration or may be localized by the use of regional administration, intramural administration, or use of an implant that acts to retain the active dose at the site of implantation. The active agent may be formulated for immediate activity or it may be formulated for sustained release.

[00298] For some conditions, particularly central nervous system conditions, it may be necessary to formulate agents to cross the blood-brain barrier (BBB). One strategy for drug delivery through the BBB entails disruption of the BBB, either by osmotic means such as mannitol or leukotrienes, or biochemically by the use of vasoactive substances such as bradykinin. The potential for using BBB opening to target specific agents to brain tumors is also an option. A BBB disrupting agent can be co-administered with the therapeutic compositions of the invention when the compositions are administered by intravascular injection. Other strategies to go through the BBB may entail the use of endogenous transport systems,

including Caveolin-1 mediated transcytosis, carrier-mediated transporters such as glucose and amino acid carriers, receptor-mediated transcytosis for insulin or transferrin, and active efflux transporters such as p-glycoprotein. Active transport moieties may also be conjugated to the therapeutic compounds for use in the invention to facilitate transport across the endothelial wall of the blood vessel. Alternatively, drug delivery of therapeutics agents behind the BBB may be by local delivery, for example by intrathecal delivery, e.g. through an Ommaya reservoir (see e.g. US Patent Nos. 5,222,982 and 5385582, incorporated herein by reference); by bolus injection, e.g. by a syringe, e.g. intravitreally or intracranially; by continuous infusion, e.g. by cannulation, e.g. with convection (see e.g. US Application No. 20070254842, incorporated here by reference); or by implanting a device upon which the agent has been reversably affixed (see e.g. US Application Nos. 20080081064 and 20090196903, incorporated herein by reference).

[00299] Typically, an effective amount of a guide RNA and/or site-directed modifying polypeptide and/or donor polynucleotide are provided. As discussed above with regard to ex vivo methods, an effective amount or effective dose of a guide RNA and/or site-directed modifying polypeptide and/or donor polynucleotide in vivo is the amount to induce a 2 fold increase or more in the amount of recombination observed between two homologous sequences relative to a negative control, e.g. a cell contacted with an empty vector or irrelevant polypeptide. The amount of recombination may be measured by any convenient method, e.g. as described above and known in the art. The calculation of the effective amount or effective dose of a guide RNA and/or site-directed modifying polypeptide and/or donor polynucleotide to be administered is within the skill of one of ordinary skill in the art, and will be routine to those persons skilled in the art. The final amount to be administered will be dependent upon the route of administration and upon the nature of the disorder or condition that is to be treated.

[00300] The effective amount given to a particular patient will depend on a variety of factors, several of which will differ from patient to patient. A competent clinician will be able to determine an effective amount of a therapeutic agent to administer to a patient to halt or reverse the progression the disease condition as required. Utilizing LD50 animal data, and other information available for the agent, a clinician can determine the maximum safe dose for an individual, depending on the route of administration. For instance, an intravenously administered dose may be more than an intrathecally administered dose, given the greater body of fluid into which the therapeutic composition is being administered. Similarly, compositions which are rapidly cleared from the body may be administered at higher doses, or in repeated doses, in order to maintain a therapeutic concentration. Utilizing ordinary skill, the competent clinician will be able to optimize the dosage of a particular therapeutic in the course of routine clinical trials.

[00301] For inclusion in a medicament, a guide RNA and/or site-directed modifying polypeptide and/or donor polynucleotide may be obtained from a suitable commercial source. As a general proposition, the total pharmaceutically effective amount of the a guide RNA and/or site-directed modifying polypeptide and/or donor polynucleotide administered parenterally per dose will be in a range that can be measured by a dose response curve.

[00302] Therapies based on a guide RNA and/or site-directed modifying polypeptide and/or donor polynucleotides, i.e. preparations of a guide RNA and/or site-directed modifying polypeptide and/or donor polynucleotide to be used for therapeutic administration, must be sterile. Sterility is readily accomplished by filtration through sterile filtration membranes (e.g., 0.2 μm membranes). Therapeutic compositions generally are placed into a container having a sterile access port, for example, an intravenous solution bag or vial having a stopper pierceable by a hypodermic injection needle. The therapies based on a guide RNA and/or site-directed modifying polypeptide and/or donor polynucleotide may be stored in unit or multi-dose containers, for example, sealed ampules or vials, as an aqueous solution or as a lyophilized formulation for reconstitution. As an example of a lyophilized formulation, 10-ml vials are filled with 5 ml of sterile-filtered 1% (w/v) aqueous solution of compound, and the resulting mixture is lyophilized. The infusion solution is prepared by reconstituting the lyophilized compound using bacteriostatic Water-for-Injection.

[00303] Pharmaceutical compositions can include, depending on the formulation desired, pharmaceutically-acceptable, non-toxic carriers of diluents, which are defined as vehicles commonly used to formulate pharmaceutical compositions for animal or human administration. The diluent is selected so as not to affect the biological activity of the combination. Examples of such diluents are distilled water, buffered water, physiological saline, PBS, Ringer's solution, dextrose solution, and Hank's solution. In addition, the pharmaceutical composition or formulation can include other carriers, adjuvants, or non-toxic, nontherapeutic, nonimmunogenic stabilizers, excipients and the like. The compositions can also include additional substances to approximate physiological conditions, such as pH adjusting and buffering agents, toxicity adjusting agents, wetting agents and detergents.

[00304] The composition can also include any of a variety of stabilizing agents, such as an antioxidant for example. When the pharmaceutical composition includes a polypeptide, the polypeptide can be complexed with various well-known compounds that enhance the in vivo stability of the polypeptide, or otherwise enhance its pharmacological properties (e.g., increase the half-life of the polypeptide, reduce its toxicity, enhance solubility or uptake). Examples of such modifications or complexing agents include sulfate, gluconate, citrate and phosphate. The nucleic acids or polypeptides of a composition can also be complexed with molecules that enhance their in vivo attributes. Such molecules include, for example, carbohydrates, polyamines, amino acids, other peptides, ions (e.g., sodium, potassium, calcium, magnesium, manganese), and lipids.

[00305] Further guidance regarding formulations that are suitable for various types of administration can be found in Remington's Pharmaceutical Sciences, Mace Publishing Company, Philadelphia, Pa., 17th ed. (1985). For a brief review of methods for drug delivery, see, Langer, Science 249:1527-1533 (1990).

[00306] The pharmaceutical compositions can be administered for prophylactic and/or therapeutic treatments. Toxicity and therapeutic efficacy of the active ingredient can be determined according to

standard pharmaceutical procedures in cell cultures and/or experimental animals, including, for example, determining the LD50 (the dose lethal to 50% of the population) and the ED50 (the dose therapeutically effective in 50% of the population). The dose ratio between toxic and therapeutic effects is the therapeutic index and it can be expressed as the ratio LD50/ED50. Therapies that exhibit large therapeutic indices are preferred.

[00307] The data obtained from cell culture and/or animal studies can be used in formulating a range of dosages for humans. The dosage of the active ingredient typically lines within a range of circulating concentrations that include the ED50 with low toxicity. The dosage can vary within this range depending upon the dosage form employed and the route of administration utilized. The components used to formulate the pharmaceutical compositions are preferably of high purity and are substantially free of potentially harmful contaminants (e.g., at least National Food (NF) grade, generally at least analytical grade, and more typically at least pharmaceutical grade). Moreover, compositions intended for in vivo use are usually sterile. To the extent that a given compound must be synthesized prior to use, the resulting product is typically substantially free of any potentially toxic agents, particularly any endotoxins, which may be present during the synthesis or purification process. Compositions for parental administration are also sterile, substantially isotonic and made under GMP conditions.

[00308] The effective amount of a therapeutic composition to be given to a particular patient will depend on a variety of factors, several of which will differ from patient to patient. A competent clinician will be able to determine an effective amount of a therapeutic agent to administer to a patient to halt or reverse the progression the disease condition as required. Utilizing LD50 animal data, and other information available for the agent, a clinician can determine the maximum safe dose for an individual, depending on the route of administration. For instance, an intravenously administered dose may be more than an intrathecally administered dose, given the greater body of fluid into which the therapeutic composition is being administered. Similarly, compositions which are rapidly cleared from the body may be administered at higher doses, or in repeated doses, in order to maintain a therapeutic concentration. Utilizing ordinary skill, the competent clinician will be able to optimize the dosage of a particular therapeutic in the course of routine clinical trials.

[00309] Genetically Modified Host Cells

[00310] The present disclosure provides genetically modified host cells, including isolated genetically modified host cells, where a genetically modified host cell comprises (has been genetically modified with: 1) an exogenous guide RNA; 2) an exogenous nucleic acid comprising a nucleotide sequence encoding a guide RNA; 3) an exogenous site-directed modifying polypeptide (e.g., a naturally occurring Cas9; a modified, i.e., mutated or variant, Cas9; a chimeric Cas9; etc.); 4) an exogenous nucleic acid comprising a nucleotide sequence encoding a site-directed modifying polypeptide; or 5) any combination of the above. A genetically modified cell is generated by genetically modifying a host cell with, for example: 1) an exogenous guide RNA;

2) an exogenous nucleic acid comprising a nucleotide sequence encoding a guide RNA; 3) an exogenous site-directed modifying polypeptide; 4) an exogenous nucleic acid comprising a nucleotide sequence encoding a site-directed modifying polypeptide; or 5) any combination of the above.).

[00311] All cells suitable to be a target cell are also suitable to be a genetically modified host cell. For example, a genetically modified host cells of interest can be a cell from any organism (e.g. a bacterial cell, an archaeal cell, a cell of a single-cell eukaryotic organism, a plant cell, an algal cell, e.g., *Botryococcus braunii, Chlamydomonas reinhardtii, Nannochloropsis gaditana, Chlorella pyrenoidosa, Sargassum patens C. Agardh,* and the like, a fungal cell (e.g., a yeast cell), an animal cell, a cell from an invertebrate animal (e.g. fruit fly, cnidarian, echinoderm, nematode, etc.), a cell from a vertebrate animal (e.g., fish, amphibian, reptile, bird, mammal), a cell from a mammal (e.g., a pig, a cow, a goat, a sheep, a rodent, a rat, a mouse, a non-human primate, a human, etc.), etc.

[00312] In some embodiments, a genetically modified host cell has been genetically modified with an exogenous nucleic acid comprising a nucleotide sequence encoding a site-directed modifying polypeptide (e.g., a naturally occurring Cas9; a modified, i.e., mutated or variant, Cas9; a chimeric Cas9; etc.). The DNA of a genetically modified host cell can be targeted for modification by introducing into the cell a guide RNA (or a DNA encoding a guide RNA, which determines the genomic location/sequence to be modified) and optionally a donor nucleic acid. In some embodiments, the nucleotide sequence encoding a site-directed modifying polypeptide is operably linked to an inducible promoter (e.g., heat shock promoter, Tetracycline-regulated promoter, Steroid-regulated promoter, Metal-regulated promoter, estrogen receptor-regulated promoter, etc.). In some embodiments, the nucleotide sequence encoding a site-directed modifying polypeptide is operably linked to a spatially restricted and/or temporally restricted promoter (e.g., a tissue specific promoter, a cell type specific promoter, etc.). In some embodiments, the nucleotide sequence encoding a site-directed modifying polypeptide is operably linked to a constitutive promoter.

[00313] In some embodiments, a genetically modified host cell is *in vitro.* In some embodiments, a genetically modified host cell is *in vivo.* In some embodiments, a genetically modified host cell is a prokaryotic cell or is derived from a prokaryotic cell. In some embodiments, a genetically modified host cell is a bacterial cell or is derived from a bacterial cell. In some embodiments, a genetically modified host cell is an archaeal cell or is derived from an archaeal cell. In some embodiments, a genetically modified host cell is a eukaryotic cell or is derived from a eukaryotic cell. In some embodiments, a genetically modified host cell is a plant cell or is derived from a plant cell. In some embodiments, a genetically modified host cell is an animal cell or is derived from an animal cell. In some embodiments, a genetically modified host cell is an invertebrate cell or is derived from an invertebrate cell. In some embodiments, a genetically modified host cell is a vertebrate cell or is derived from a vertebrate cell. In some embodiments, a genetically modified host cell is a mammalian cell or is derived from a mammalian cell. In some embodiments, a

genetically modified host cell is a rodent cell or is derived from a rodent cell. In some embodiments, a genetically modified host cell is a human cell or is derived from a human cell.

[00314] The present disclosure further provides progeny of a genetically modified cell, where the progeny can comprise the same exogenous nucleic acid or polypeptide as the genetically modified cell from which it was derived. The present disclosure further provides a composition comprising a genetically modified host cell.

[00315] Genetically modified stem cells and genetically modified progenitor cells

[00316] In some embodiments, a genetically modified host cell is a genetically modified stem cell or progenitor cell. Suitable host cells include, e.g., stem cells (adult stem cells, embryonic stem cells, iPS cells, etc.) and progenitor cells (e.g., cardiac progenitor cells, neural progenitor cells, etc.). Suitable host cells include mammalian stem cells and progenitor cells, including, e.g., rodent stem cells, rodent progenitor cells, human stem cells, human progenitor cells, etc. Suitable host cells include *in vitro* host cells, e.g., isolated host cells.

[00317] In some embodiments, a genetically modified host cell comprises an exogenous guide RNA nucleic acid. In some embodiments, a genetically modified host cell comprises an exogenous nucleic acid comprising a nucleotide sequence encoding a guide RNA. In some embodiments, a genetically modified host cell comprises an exogenous site-directed modifying polypeptide (e.g., a naturally occurring Cas9; a modified, i.e., mutated or variant, Cas9; a chimeric Cas9; etc.). In some embodiments, a genetically modified host cell comprises an exogenous nucleic acid comprising a nucleotide sequence encoding a site-directed modifying polypeptide. In some embodiments, a genetically modified host cell comprises exogenous nucleic acid comprising a nucleotide sequence encoding 1) a guide RNA and 2) a site-directed modifying polypeptide.

[00318] In some cases, the site-directed modifying polypeptide comprises an amino acid sequence having at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 95%, at least about 99%, or 100%, amino acid sequence identity to amino acids 7-166 and/or 731-1003 of SEQ ID NO: 8, or to the corresponding portions in any of the amino acid sequences set forth as SEQ ID NOs: 1-800.

[00319] Compositions

[00320] The present disclosure provides a composition comprising a guide RNA and/or a site-directed modifying polypeptide. In some cases, the site-directed modifying polypeptide is a chimeric polypeptide. A composition is useful for carrying out a method of the present disclosure, e.g., a method for site-specific modification of a target DNA; a method for site-specific modification of a polypeptide associated with a target DNA; etc.

[00321] Compositions comprising a guide RNA

[00322] The present disclosure provides a composition comprising a guide RNA. The composition can comprise, in addition to the guide RNA, one or more of: a salt, e.g., NaCl, $MgCl_2$, KCl, $MgSO_4$, etc.; a buffering agent, e.g., a Tris buffer, N-(2-Hydroxyethyl)piperazine-N'-(2-ethanesulfonic acid) (HEPES), 2-(N-Morpholino)ethanesulfonic acid (MES), MES sodium salt, 3-(N-Morpholino)propanesulfonic acid (MOPS), N-tris[Hydroxymethyl]methy1-3-aminopropanesulfonic acid (TAPS), etc.; a solubilizing agent; a detergent, e.g., a non-ionic detergent such as Tween-20, etc.; a nuclease inhibitor; and the like. For example, in some cases, a composition comprises a guide RNA and a buffer for stabilizing nucleic acids.

[00323] In some embodiments, a guide RNA present in a composition is pure, e.g., at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 95%, at least about 98%, at least about 99%, or more than 99% pure, where "% purity" means that guide RNA is the recited percent free from other macromolecules, or contaminants that may be present during the production of the guide RNA.

[00324] Compositions comprising a chimeric polypeptide

[00325] The present disclosure provides a composition a chimeric polypeptide. The composition can comprise, in addition to the guide RNA, one or more of: a salt, e.g., NaCl, $MgCl_2$, KCl, $MgSO_4$, etc.; a buffering agent, e.g., a Tris buffer, HEPES, MES, MES sodium salt, MOPS, TAPS, etc.; a solubilizing agent; a detergent, e.g., a non-ionic detergent such as Tween-20, etc.; a protease inhibitor; a reducing agent (e.g., dithiothreitol); and the like.

[00326] In some embodiments, a chimeric polypeptide present in a composition is pure, e.g., at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 95%, at least about 98%, at least about 99%, or more than 99% pure, where "% purity" means that the site-directed modifying polypeptide is the recited percent free from other proteins, other macromolecules, or contaminants that may be present during the production of the chimeric polypeptide.

[00327] Compositions comprising a guide RNA and a site-directed modifying polypeptide

[00328] The present disclosure provides a composition comprising: (i) a guide RNA or a DNA polynucleotide encoding the same; and ii) a site-directed modifying polypeptide, or a polynucleotide encoding the same. In some cases, the site-directed modifying polypeptide is a chimeric site-directed modifying polypeptide. In other cases, the site-directed modifying polypeptide is a naturally-occurring site-directed modifying polypeptide. In some instances, the site-directed modifying polypeptide exhibits enzymatic activity that modifies a target DNA. In other cases, the site-directed modifying polypeptide exhibits enzymatic activity that modifies a polypeptide that is associated with a target DNA. In still other cases, the site-directed modifying polypeptide modulates transcription of the target DNA.

[00329] The present disclosure provides a composition comprising: (i) a guide RNA, as described above, or a DNA polynucleotide encoding the same, the guide RNA comprising: (a) a first segment comprising a nucleotide sequence that is complementary to a sequence in a target DNA; and (b) a second segment that interacts with a site-directed modifying polypeptide; and (ii) the site-directed modifying polypeptide, or a polynucleotide encoding the same, the site-directed modifying polypeptide comprising: (a) an RNA-binding portion that interacts with the guide RNA; and (b) an activity portion that exhibits site-directed enzymatic activity, wherein the site of enzymatic activity is determined by the guide RNA.

[00330] In some instances, a composition comprises: a composition comprising: (i) a guide RNA, the guide RNA comprising: (a) a first segment comprising a nucleotide sequence that is complementary to a sequence in a target DNA; and (b) a second segment that interacts with a site-directed modifying polypeptide; and (ii) the site-directed modifying polypeptide, the site-directed modifying polypeptide comprising: (a) an RNA-binding portion that interacts with the guide RNA; and (b) an activity portion that exhibits site-directed enzymatic activity, wherein the site of enzymatic activity is determined by the guide RNA.

[00331] In other embodiments, a composition comprises: (i) a polynucleotide encoding a guide RNA, the guide RNA comprising: (a) a first segment comprising a nucleotide sequence that is complementary to a sequence in a target DNA; and (b) a second segment that interacts with a site-directed modifying polypeptide; and (ii) a polynucleotide encoding the site-directed modifying polypeptide, the site-directed modifying polypeptide comprising: (a) an RNA-binding portion that interacts with the guide RNA; and (b) an activity portion that exhibits site-directed enzymatic activity, wherein the site of enzymatic activity is determined by the guide RNA.

[00332] In some embodiments, a composition includes both RNA molecules of a double-molecule guide RNA. As such, in some embodiments, a composition includes an activator-RNA that comprises a duplex-forming segment that is complementary to the duplex-forming segment of a targeter-. The duplex-forming segments of the activator-RNA and the targeter-RNA hybridize to form the dsRNA duplex of the protein-binding segment of the guide RNA. The targeter-RNA further provides the DNA-targeting segment (single stranded) of the guide RNA and therefore targets the guide RNA to a specific sequence within the target DNA. As one non-limiting example, the duplex-forming segment of the activator-RNA comprises a nucleotide sequence that has at least about 70%, at least about 80%, at least about 90%, at least about 95%, at least about 98%, or 100% identity with a tracrRNA sequence set out in Supplementary Table S5. As another non-limiting example, the duplex-forming segment of the targeter-RNA comprises a nucleotide sequence that has at least about 70%, at least about 80%, at least about 90%, at least about 95%, at least about 98%, or 100% identity with a CRISPR repeat sequence set out in Supplementary Table S5.

[00333] The present disclosure provides a composition comprising: (i) a guide RNA, or a DNA polynucleotide encoding the same, the guide RNA comprising: (a) a first segment comprising a nucleotide sequence that is complementary to a sequence in a target DNA; and (b) a second segment that interacts with a site-directed modifying polypeptide; and (ii) the site-directed modifying polypeptide, or a polynucleotide encoding the same, the site-directed modifying polypeptide comprising: (a) an RNA-binding portion that interacts with the guide RNA; and (b) an activity portion that modulates transcription within the target DNA, wherein the site of modulated transcription within the target DNA is determined by the guide RNA.

[00334] For example, in some cases, a composition comprises: (i) a guide RNA, the guide RNA comprising: (a) a first segment comprising a nucleotide sequence that is complementary to a sequence in a target DNA; and (b) a second segment that interacts with a site-directed modifying polypeptide; and (ii) the site-directed modifying polypeptide, the site-directed modifying polypeptide comprising: (a) an RNA-binding portion that interacts with the guide RNA; and (b) an activity portion that modulates transcription within the target DNA, wherein the site of modulated transcription within the target DNA is determined by the guide RNA.

[00335] As another example, in some cases, a composition comprises: (i) a DNA polynucleotide encoding a guide RNA, the guide RNA comprising: (a) a first segment comprising a nucleotide sequence that is complementary to a sequence in a target DNA; and (b) a second segment that interacts with a site-directed modifying polypeptide; and (ii) a polynucleotide encoding the site-directed modifying polypeptide, the site-directed modifying polypeptide comprising: (a) an RNA-binding portion that interacts with the guide RNA; and (b) an activity portion that modulates transcription within the target DNA, wherein the site of modulated transcription within the target DNA is determined by the guide RNA. A composition can comprise, in addition to i) a guide RNA, or a DNA polynucleotide encoding the same; and ii) a site-directed modifying polypeptide, or a polynucleotide encoding the same, one or more of: a salt, e.g., NaCl, $MgCl_2$, KCl, $MgSO_4$, etc.; a buffering agent, e.g., a Tris buffer, HEPES, MES, MES sodium salt, MOPS, TAPS, etc.; a solubilizing agent; a detergent, e.g., a non-ionic detergent such as Tween-20, etc.; a protease inhibitor; a reducing agent (e.g., dithiothreitol); and the like.

[00336] In some cases, the components of the composition are individually pure, e.g., each of the components is at least about 75%, at least about 80%, at least about 90%, at least about 95%, at least about 98%, at least about 99%, or at least 99%, pure. In some cases, the individual components of a composition are pure before being added to the composition.

[00337] For example, in some embodiments, a site-directed modifying polypeptide present in a composition is pure, e.g., at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 95%, at least about 98%, at least about 99%, or more than 99% pure, where "% purity" means that the site-directed modifying polypeptide is the recited percent free from other proteins (e.g.,

proteins other than the site-directed modifying polypeptide), other macromolecules, or contaminants that may be present during the production of the site-directed modifying polypeptide.

[00338] Kits

[00339] The present disclosure provides kits for carrying out a method. A kit can include one or more of: a site-directed modifying polypeptide; a nucleic acid comprising a nucleotide encoding a site-directed modifying polypeptide; a guide RNA; a nucleic acid comprising a nucleotide sequence encoding a guide RNA; an activator-RNA; a nucleic acid comprising a nucleotide sequence encoding an activator-RNA; a targeter-RNA; and a nucleic acid comprising a nucleotide sequence encoding a targeter-RNA. A site-directed modifying polypeptide; a nucleic acid comprising a nucleotide encoding a site-directed modifying polypeptide; a guide RNA; a nucleic acid comprising a nucleotide sequence encoding a guide RNA; an activator-RNA; a nucleic acid comprising a nucleotide sequence encoding an activator-RNA; a targeter-RNA; and a nucleic acid comprising a nucleotide sequence encoding a targeter-RNA, are described in detail above. A kit may comprise a complex that comprises two or more of: a site-directed modifying polypeptide; a nucleic acid comprising a nucleotide encoding a site-directed modifying polypeptide; a guide RNA; a nucleic acid comprising a nucleotide sequence encoding a guide RNA; an activator-RNA; a nucleic acid comprising a nucleotide sequence encoding an activator-RNA; a targeter-RNA; and a nucleic acid comprising a nucleotide sequence encoding a targeter-RNA. In some embodiments, a kit comprises a site-directed modifying polypeptide, or a polynucleotide encoding the same. In some embodiments, the site-directed modifying polypeptide comprises: (a) an RNA-binding portion that interacts with the guide RNA; and (b) an activity portion that modulates transcription within the target DNA, wherein the site of modulated transcription within the target DNA is determined by the guide RNA. In some cases, the activity portion of the site-directed modifying polypeptide exhibits reduced or inactivated nuclease activity. In some cases, the site-directed modifying polypeptide is a chimeric site-directed modifying polypeptide.

[00340] In some embodiments, a kit comprises: a site-directed modifying polypeptide, or a polynucleotide encoding the same, and a reagent for reconstituting and/or diluting the site-directed modifying polypeptide. In other embodiments, a kit comprises a nucleic acid (e.g., DNA, RNA) comprising a nucleotide encoding a site-directed modifying polypeptide. In some embodiments, a kit comprises: a nucleic acid (e.g., DNA, RNA) comprising a nucleotide encoding a site-directed modifying polypeptide; and a reagent for reconstituting and/or diluting the site-directed modifying polypeptide.

[00341] A kit comprising a site-directed modifying polypeptide, or a polynucleotide encoding the same, can further include one or more additional reagents, where such additional reagents can be selected from: a buffer for introducing the site-directed modifying polypeptide into a cell; a wash buffer; a control reagent; a control expression vector or RNA polynucleotide; a reagent for *in vitro* production of the site-directed modifying polypeptide from DNA, and the like. In some cases, the site-directed modifying polypeptide included in a kit is a chimeric site-directed modifying polypeptide, as described above.

[00342] In some embodiments, a kit comprises a guide RNA, or a DNA polynucleotide encoding the same, the guide RNA comprising: (a) a first segment comprising a nucleotide sequence that is complementary to a sequence in a target DNA; and (b) a second segment that interacts with a site-directed modifying polypeptide. In some embodiments, the guide RNA further comprises a third segment (as described above). In some embodiments, a kit comprises: (i) a guide RNA, or a DNA polynucleotide encoding the same, the guide RNA comprising: (a) a first segment comprising a nucleotide sequence that is complementary to a sequence in a target DNA; and (b) a second segment that interacts with a site-directed modifying polypeptide; and (ii) a site-directed modifying polypeptide, or a polynucleotide encoding the same, the site-directed modifying polypeptide comprising: (a) an RNA-binding portion that interacts with the guide RNA; and (b) an activity portion that exhibits site-directed enzymatic activity, wherein the site of enzymatic activity is determined by the guide RNA. In some embodiments, the activity portion of the site-directed modifying polypeptide does not exhibit enzymatic activity (comprises an inactivated nuclease, e.g., via mutation). In some cases, the kit comprises a guide RNA and a site-directed modifying polypeptide. In other cases, the kit comprises: (i) a nucleic acid comprising a nucleotide sequence encoding a guide RNA; and (ii) a nucleic acid comprising a nucleotide sequence encoding site-directed modifying polypeptide. As another example, a kit can include: (i) a guide RNA, or a DNA polynucleotide encoding the same, comprising: (a) a first segment comprising a nucleotide sequence that is complementary to a sequence in a target DNA; and (b) a second segment that interacts with a site-directed modifying polypeptide; and (ii) the site-directed modifying polypeptide, or a polynucleotide encoding the same, comprising: (a) an RNA-binding portion that interacts with the guide RNA; and (b) an activity portion that that modulates transcription within the target DNA, wherein the site of modulated transcription within the target DNA is determined by the guide RNA In some cases, the kit comprises: (i) a guide RNA; and a site-directed modifying polypeptide. In other cases, the kit comprises: (i) a nucleic acid comprising a nucleotide sequence encoding a guide RNA; and (ii) a nucleic acid comprising a nucleotide sequence encoding site-directed modifying polypeptide. The present disclosure provides a kit comprising: (1) a recombinant expression vector comprising (i) a nucleotide sequence encoding a guide RNA, wherein the guide RNA comprises: (a) a first segment comprising a nucleotide sequence that is complementary to a sequence in a target DNA; and (b) a second segment that interacts with a site-directed modifying polypeptide; and (ii) a nucleotide sequence encoding the site-directed modifying polypeptide, wherein the site-directed modifying polypeptide comprises: (a) an RNA-binding portion that interacts with the guide RNA; and (b) an activity portion that exhibits site-directed enzymatic activity, wherein the site of enzymatic activity is determined by the guide RNA.; and (2) a reagent for reconstitution and/or dilution of the expression vector.

[00343] The present disclosure provides a kit comprising: (1) a recombinant expression vector comprising: (i) a nucleotide sequence encoding a guide RNA, wherein the guide RNA comprises: (a) a first segment comprising a nucleotide sequence that is complementary to a sequence in a target DNA; and (b) a second segment that interacts with a site-directed modifying polypeptide; and (ii) a nucleotide sequence

encoding the site-directed modifying polypeptide, wherein the site-directed modifying polypeptide comprises: (a) an RNA-binding portion that interacts with the guide RNA; and (b) an activity portion that modulates transcription within the target DNA, wherein the site of modulated transcription within the target DNA is determined by the guide RNA; and (2) a reagent for reconstitution and/or dilution of the recombinant expression vector.

[00344] The present disclosure provides a kit comprising: (1) a recombinant expression vector comprising a nucleic acid comprising a nucleotide sequence that encodes a DNA targeting RNA comprising: (i) a first segment comprising a nucleotide sequence that is complementary to a sequence in a target DNA; and (ii) a second segment that interacts with a site-directed modifying polypeptide; and (2) a reagent for reconstitution and/or dilution of the recombinant expression vector. In some embodiments of this kit, the kit comprises: a recombinant expression vector comprising a nucleotide sequence that encodes a site-directed modifying polypeptide, wherein the site-directed modifying polypeptide comprises: (a) an RNA-binding portion that interacts with the guide RNA; and (b) an activity portion that exhibits site-directed enzymatic activity, wherein the site of enzymatic activity is determined by the guide RNA. In other embodiments of this kit, the kit comprises: a recombinant expression vector comprising a nucleotide sequence that encodes a site-directed modifying polypeptide, wherein the site-directed modifying polypeptide comprises: (a) an RNA-binding portion that interacts with the guide RNA; and (b) an activity portion that modulates transcription within the target DNA, wherein the site of modulated transcription within the target DNA is determined by the guide RNA.

[00345] In some embodiments of any of the above kits, the kit comprises an activator-RNA or a targeter-RNA. In some embodiments of any of the above kits, the kit comprises a single-molecule guide RNA. In some embodiments of any of the above kits, the kit comprises two or more double-molecule or single-molecule guide RNAs. In some embodiments of any of the above kits, a guide RNA (e.g., including two or more guide RNAs) can be provided as an array (e.g., an array of RNA molecules, an array of DNA molecules encoding the guide RNA(s), etc.). Such kits can be useful, for example, for use in conjunction with the above described genetically modified host cells that comprise a site-directed modifying polypeptide. In some embodiments of any of the above kits, the kit further comprises a donor polynucleotide to effect the desired genetic modification. Components of a kit can be in separate containers; or can be combined in a single container.

[00346] In some cases, a kit further comprises one or more variant Cas9 site-directed polypeptides that exhibits reduced endodeoxyribonuclease activity relative to wild-type Cas9.

[00347] In some cases, a kit further comprises one or more nucleic acids comprising a nucleotide sequence encoding a variant Cas9 site-directed polypeptide that exhibits reduced endodeoxyribonuclease activity relative to wild-type Cas9.

[00348] Any of the above-described kits can further include one or more additional reagents, where such additional reagents can be selected from: a dilution buffer; a reconstitution solution; a wash buffer; a control reagent; a control expression vector or RNA polynucleotide; a reagent for *in vitro* production of the site-directed modifying polypeptide from DNA, and the like.

[00349] In addition to above-mentioned components, a kit can further include instructions for using the components of the kit to practice the methods. The instructions for practicing the methods are generally recorded on a suitable recording medium. For example, the instructions may be printed on a substrate, such as paper or plastic, etc. As such, the instructions may be present in the kits as a package insert, in the labeling of the container of the kit or components thereof (i.e., associated with the packaging or subpackaging) etc. In other embodiments, the instructions are present as an electronic storage data file present on a suitable computer readable storage medium, e.g. CD-ROM, diskette, flash drive, etc. In yet other embodiments, the actual instructions are not present in the kit, but means for obtaining the instructions from a remote source, e.g. via the internet, are provided. An example of this embodiment is a kit that includes a web address where the instructions can be viewed and/or from which the instructions can be downloaded. As with the instructions, this means for obtaining the instructions is recorded on a suitable substrate.

[00350] Non-Human Genetically Modified Organisms

[00351] In some embodiments, a genetically modified host cell has been genetically modified with an exogenous nucleic acid comprising a nucleotide sequence encoding a site-directed modifying polypeptide (e.g., a naturally occurring Cas9; a modified, i.e., mutated or variant, Cas9; a chimeric Cas9; etc.). If such a cell is a eukaryotic single-cell organism, then the modified cell can be considered a genetically modified organism. In some embodiments, the non-human genetically modified organism is a Cas9 transgenic multicellular organism.

[00352] In some embodiments, a genetically modified non-human host cell (e.g., a cell that has been genetically modified with an exogenous nucleic acid comprising a nucleotide sequence encoding a site-directed modifying polypeptide, e.g., a naturally occurring Cas9; a modified, i.e., mutated or variant, Cas9; a chimeric Cas9; etc.) can generate a genetically modified nonhuman organism (e.g., a mouse, a fish, a frog, a fly, a worm, etc.). For example, if the genetically modified host cell is a pluripotent stem cell (i.e., PSC) or a germ cell (e.g., sperm, oocyte, etc.), an entire genetically modified organism can be derived from the genetically modified host cell. In some embodiments, the genetically modified host cell is a pluripotent stem cell (e.g., ESC, iPSC, pluripotent plant stem cell, etc.) or a germ cell (e.g., sperm cell, oocyte, etc.), either in vivo or in vitro, that can give rise to a genetically modified organism. In some embodiments the genetically modified host cell is a vertebrate PSC (e.g., ESC, iPSC, etc.) and is used to generate a genetically modified organism (e.g. by injecting a PSC into a blastocyst to produce a chimeric/mosaic animal, which could then be mated to generate non-chimeric/non-mosaic genetically modified organisms; grafting in the case of

plants; etc.). Any convenient method/protocol for producing a genetically modified organism, including the methods described herein, is suitable for producing a genetically modified host cell comprising an exogenous nucleic acid comprising a nucleotide sequence encoding a site-directed modifying polypeptide (e.g., a naturally occurring Cas9; a modified, i.e., mutated or variant, Cas9; a chimeric Cas9; etc.). Methods of producing genetically modified organisms are known in the art. For example, see Cho et al., Curr Protoc Cell Biol. 2009 Mar;Chapter 19:Unit 19.11: Generation of transgenic mice; Gama et al., Brain Struct Funct. 2010 Mar;214(2-3):91-109. Epub 2009 Nov 25: Animal transgenesis: an overview; Husaini et al., GM Crops. 2011 Jun-Dec;2(3):150-62. Epub 2011 Jun 1: Approaches for gene targeting and targeted gene expression in plants.

[00353] In some embodiments, a genetically modified organism comprises a target cell for methods of the invention, and thus can be considered a source for target cells. For example, if a genetically modified cell comprising an exogenous nucleic acid comprising a nucleotide sequence encoding a site-directed modifying polypeptide (e.g., a naturally occurring Cas9; a modified, i.e., mutated or variant, Cas9; a chimeric Cas9; etc.) is used to generate a genetically modified organism, then the cells of the genetically modified organism comprise the exogenous nucleic acid comprising a nucleotide sequence encoding a site-directed modifying polypeptide (e.g., a naturally occurring Cas9; a modified, i.e., mutated or variant, Cas9; a chimeric Cas9; etc.). In some such embodiments, the DNA of a cell or cells of the genetically modified organism can be targeted for modification by introducing into the cell or cells a guide RNA (or a DNA encoding a guide RNA) and optionally a donor nucleic acid. For example, the introduction of a guide RNA (or a DNA encoding a guide RNA) into a subset of cells (e.g., brain cells, intestinal cells, kidney cells, lung cells, blood cells, etc.) of the genetically modified organism can target the DNA of such cells for modification, the genomic location of which will depend on the DNA-targeting sequence of the introduced guide RNA.

[00354] In some embodiments, a genetically modified organism is a source of target cells for methods of the invention. For example, a genetically modified organism comprising cells that are genetically modified with an exogenous nucleic acid comprising a nucleotide sequence encoding a site-directed modifying polypeptide (e.g., a naturally occurring Cas9; a modified, i.e., mutated or variant, Cas9; a chimeric Cas9; etc.) can provide a source of genetically modified cells, for example PSCs (e.g., ESCs, iPSCs, sperm, oocytes, etc.), neurons, progenitor cells, cardiomyocytes, etc.

[00355] In some embodiments, a genetically modified cell is a PSC comprising an exogenous nucleic acid comprising a nucleotide sequence encoding a site-directed modifying polypeptide (e.g., a naturally occurring Cas9; a modified, i.e., mutated or variant, Cas9; a chimeric Cas9; etc.). As such, the PSC can be a target cell such that the DNA of the PSC can be targeted for modification by introducing into the PSC a guide RNA (or a DNA encoding a guide RNA) and optionally a donor nucleic acid, and the genomic location of the modification will depend on the DNA-targeting sequence of the introduced guide RNA. Thus, in some embodiments, the methods described herein can be used to modify the DNA (e.g., delete and/or

replace any desired genomic location) of PSCs derived from a genetically modified organism. Such modified PSCs can then be used to generate organisms having both (i) an exogenous nucleic acid comprising a nucleotide sequence encoding a site-directed modifying polypeptide (e.g., a naturally occurring Cas9; a modified, i.e., mutated or variant, Cas9; a chimeric Cas9; etc.) and (ii) a DNA modification that was introduced into the PSC.

[00356] An exogenous nucleic acid comprising a nucleotide sequence encoding a site-directed modifying polypeptide (e.g., a naturally occurring Cas9; a modified, i.e., mutated or variant, Cas9; a chimeric Cas9; etc.) can be under the control of (i.e., operably linked to) an unknown promoter (e.g., when the nucleic acid randomly integrates into a host cell genome) or can be under the control of (i.e., operably linked to) a known promoter. Suitable known promoters can be any known promoter and include constitutively active promoters (e.g., CMV promoter), inducible promoters (e.g., heat shock promoter, Tetracycline-regulated promoter, Steroid-regulated promoter, Metal-regulated promoter, estrogen receptor-regulated promoter, etc.), spatially restricted and/or temporally restricted promoters (e.g., a tissue specific promoter, a cell type specific promoter, etc.), etc.

[00357] A genetically modified organism (e.g. an organism whose cells comprise a nucleotide sequence encoding a site-directed modifying polypeptide, e.g., a naturally occurring Cas9; a modified, i.e., mutated or variant, Cas9; a chimeric Cas9; etc.) can be any organism including for example, a plant; algae; an invertebrate (e.g., a cnidarian, an echinoderm, a worm, a fly, etc.); a vertebrate (e.g., a fish (e.g., zebrafish, puffer fish, gold fish, etc.), an amphibian (e.g., salamander, frog, etc.), a reptile, a bird, a mammal, etc.); an ungulate (e.g., a goat, a pig, a sheep, a cow, etc.); a rodent (e.g., a mouse, a rat, a hamster, a guinea pig); a lagomorpha (e.g., a rabbit); etc.

[00358] In some cases, the site-directed modifying polypeptide comprises an amino acid sequence having at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 95%, at least about 99%, or 100%, amino acid sequence identity to amino acids 7-166 and/or 731-1003 of SEQ ID NO: 8, or to the corresponding portions in any of the amino acid sequences set forth as SEQ ID NOs: 1-800.

[00359] Transgenic non-human animals

[00360] As described above, in some embodiments, a nucleic acid (e.g., a nucleotide sequence encoding a site-directed modifying polypeptide, e.g., a naturally occurring Cas9; a modified, i.e., mutated or variant, Cas9; a chimeric Cas9; etc.) or a recombinant expression vector is used as a transgene to generate a transgenic animal that produces a site-directed modifying polypeptide. Thus, the present disclosure further provides a transgenic non-human animal, which animal comprises a transgene comprising a nucleic acid comprising a nucleotide sequence encoding a site-directed modifying polypeptide, e.g., a naturally occurring Cas9; a modified, i.e., mutated or variant, Cas9; a chimeric Cas9; etc., as described above. In some embodiments, the genome of the transgenic non-human animal

comprises a nucleotide sequence encoding a site-directed modifying polypeptide. In some embodiments, the transgenic non-human animal is homozygous for the genetic modification. In some embodiments, the transgenic non-human animal is heterozygous for the genetic modification. In some embodiments, the transgenic non-human animal is a vertebrate, for example, a fish (e.g., zebra fish, gold fish, puffer fish, cave fish, etc.), an amphibian (frog, salamander, etc.), a bird (e.g., chicken, turkey, etc.), a reptile (e.g., snake, lizard, etc.), a mammal (e.g., an ungulate, e.g., a pig, a cow, a goat, a sheep, etc.; a lagomorph (e.g., a rabbit); a rodent (e.g., a rat, a mouse); a nonhuman primate; etc.), etc.

[00361] An exogenous nucleic acid comprising a nucleotide sequence encoding a site-directed modifying polypeptide (e.g., a naturally occurring Cas9; a modified, i.e., mutated or variant, Cas9; a chimeric Cas9; etc.) can be under the control of (i.e., operably linked to) an unknown promoter (e.g., when the nucleic acid randomly integrates into a host cell genome) or can be under the control of (i.e., operably linked to) a known promoter. Suitable known promoters can be any known promoter and include constitutively active promoters (e.g., CMV promoter), inducible promoters (e.g., heat shock promoter, Tetracycline-regulated promoter, Steroid-regulated promoter, Metal-regulated promoter, estrogen receptor-regulated promoter, etc.), spatially restricted and/or temporally restricted promoters (e.g., a tissue specific promoter, a cell type specific promoter, etc.), etc.

[00362] In some cases, the site-directed modifying polypeptide comprises an amino acid sequence having at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 95%, at least about 99%, or 100%, amino acid sequence identity to amino acids 7-166 and/or 731-1003 of SEQ ID NO: 8, or to the corresponding portions in any of the amino acid sequences set forth as SEQ ID NOs: 1-800.

[00363] Transgenic plants

[00364] As described above, in some embodiments, a nucleic acid (e.g., a nucleotide sequence encoding a site-directed modifying polypeptide, e.g., a naturally occurring Cas9; a modified, i.e., mutated or variant, Cas9; a chimeric Cas9; etc.) or a recombinant expression vector is used as a transgene to generate a transgenic plant that produces a site-directed modifying polypeptide. Thus, the present disclosure further provides a transgenic plant, which plant comprises a transgene comprising a nucleic acid comprising a nucleotide sequence encoding site-directed modifying polypeptide, e.g., a naturally occurring Cas9; a modified, i.e., mutated or variant, Cas9; a chimeric Cas9; etc., as described above. In some embodiments, the genome of the transgenic plant comprises a nucleic acid. In some embodiments, the transgenic plant is homozygous for the genetic modification. In some embodiments, the transgenic plant is heterozygous for the genetic modification.

[00365] Methods of introducing exogenous nucleic acids into plant cells are well known in the art. Such plant cells are considered "transformed," as defined above. Suitable methods include viral infection (such as double stranded DNA viruses), transfection, conjugation, protoplast fusion, electroporation,

particle gun technology, calcium phosphate precipitation, direct microinjection, silicon carbide whiskers technology, *Agrobacterium-mediated* transformation and the like. The choice of method is generally dependent on the type of cell being transformed and the circumstances under which the transformation is taking place (i.e. *in vitro, ex vivo,* or *in vivo).* Transformation methods based upon the soil bacterium *Agrobacterium tumefaciens* are particularly useful for introducing an exogenous nucleic acid molecule into a vascular plant. The wild type form of *Agrobacterium* contains a Ti (tumor-inducing) plasmid that directs production of tumorigenic crown gall growth on host plants. Transfer of the tumor-inducing T-DNA region of the Ti plasmid to a plant genome requires the Ti plasmid-encoded virulence genes as well as T-DNA borders, which are a set of direct DNA repeats that delineate the region to be transferred. An *Agrobacterium-based* vector is a modified form of a Ti plasmid, in which the tumor inducing functions are replaced by the nucleic acid sequence of interest to be introduced into the plant host.

[00366] *Agrobacterium-mediated* transformation generally employs cointegrate vectors or binary vector systems, in which the components of the Ti plasmid are divided between a helper vector, which resides permanently in the *Agrobacterium* host and carries the virulence genes, and a shuttle vector, which contains the gene of interest bounded by T-DNA sequences. A variety of binary vectors are well known in the art and are commercially available, for example, from Clontech (Palo Alto, Calif.). Methods of coculturing *Agrobacterium* with cultured plant cells or wounded tissue such as leaf tissue, root explants, hypocotyledons, stem pieces or tubers, for example, also are well known in the art. See., e.g., Glick and Thompson, (eds.), *Methods in Plant Molecular Biology and Biotechnology,* Boca Raton, Fla.: CRC Press (1993).

[00367] Microprojectile-mediated transformation also can be used to produce a transgenic plant. This method, first described by Klein et al. *(Nature* 327:70-73 (1987)), relies on microprojectiles such as gold or tungsten that are coated with the desired nucleic acid molecule by precipitation with calcium chloride, spermidine or polyethylene glycol. The microprojectile particles are accelerated at high speed into an angiosperm tissue using a device such as the BIOLISTIC PD-1000 (Biorad; Hercules Calif.).

[00368] A nucleic acid may be introduced into a plant in a manner such that the nucleic acid is able to enter a plant cell(s), e.g., via an *in vivo* or *ex vivo* protocol. By *"in vivo,"* it is meant in the nucleic acid is administered to a living body of a plant *e.g.* infiltration. By *"ex vivo"* it is meant that cells or explants are modified outside of the plant, and then such cells or organs are regenerated to a plant. A number of vectors suitable for stable transformation of plant cells or for the establishment of transgenic plants have been described, including those described in Weissbach and Weissbach, (1989) Methods for Plant Molecular Biology Academic Press, and Gelvin et al., (1990) Plant Molecular Biology Manual, Kluwer Academic Publishers. Specific examples include those derived from a Ti plasmid of *Agrobacterium tumefaciens,* as well as those disclosed by Herrera-Estrella et al. (1983) Nature 303: 209, Bevan (1984) Nucl Acid Res. 12: 8711-8721, Klee (1985) Bio/Technolo 3: 637-642. Alternatively, non-Ti vectors can be used to transfer the DNA into plants and cells by using free DNA delivery techniques. By using these

methods transgenic plants such as wheat, rice (Christou (1991) Bio/Technology 9:957-9 and 4462) and corn (Gordon-Kamm (1990) Plant Cell 2: 603-618) can be produced. An immature embryo can also be a good target tissue for monocots for direct DNA delivery techniques by using the particle gun (Weeks et al. (1993) Plant Physiol 102: 1077-1084; Vasil (1993) Bio/Technolo 10: 667-674; Wan and Lemeaux (1994) Plant Physiol 104: 37-48 and for Agrobacterium-mediated DNA transfer (Ishida et al. (1996) Nature Biotech 14: 745-750). Exemplary methods for introduction of DNA into chloroplasts are biolistic bombardment, polyethylene glycol transformation of protoplasts, and microinjection (Daniell et al Nat. Biotechnol 16:345-348, 1998; Staub et al Nat. Biotechnol 18: 333-338, 2000; O'Neill et al Plant J. 3:729-738, 1993; Knoblauch et al Nat. Biotechnol 17: 906-909; U.S. Pat. Nos. 5,451,513, 5,545,817, 5,545,818, and 5,576,198; in Intl. Application No. WO 95/16783; and in Boynton et al., Methods in Enzymology 217: 510-536 (1993), Svab et al., Proc. Natl. Acad. Sci. USA 90: 913-917 (1993), and McBride et al., Proc. Nati. Acad. Sci. USA 91: 7301-7305 (1994)). Any vector suitable for the methods of biolistic bombardment, polyethylene glycol transformation of protoplasts and microinjection will be suitable as a targeting vector for chloroplast transformation. Any double stranded DNA vector may be used as a transformation vector, especially when the method of introduction does not utilize *Agrobacterium*.

[00369] Plants which can be genetically modified include grains, forage crops, fruits, vegetables, oil seed crops, palms, forestry, and vines. Specific examples of plants which can be modified follow: maize, banana, peanut, field peas, sunflower, tomato, canola, tobacco, wheat, barley, oats, potato, soybeans, cotton, carnations, sorghum, lupin and rice.

[00370] Also provided by the disclosure are transformed plant cells, tissues, plants and products that contain the transformed plant cells. A feature of the transformed cells, and tissues and products that include the same is the presence of a nucleic acid integrated into the genome, and production by plant cells of a site-directed modifying polypeptide, e.g., a naturally occurring Cas9; a modified, i.e., mutated or variant, Cas9; a chimeric Cas9; etc. Recombinant plant cells of the present invention are useful as populations of recombinant cells, or as a tissue, seed, whole plant, stem, fruit, leaf, root, flower, stem, tuber, grain, animal feed, a field of plants, and the like.

[00371] A nucleic acid comprising a nucleotide sequence encoding a site-directed modifying polypeptide (e.g., a naturally occurring Cas9; a modified, i.e., mutated or variant, Cas9; a chimeric Cas9; etc.) can be under the control of (i.e., operably linked to) an unknown promoter (e.g., when the nucleic acid randomly integrates into a host cell genome) or can be under the control of (i.e., operably linked to) a known promoter. Suitable known promoters can be any known promoter and include constitutively active promoters, inducible promoters, spatially restricted and/or temporally restricted promoters, etc.

[00372] In some cases, the site-directed modifying polypeptide comprises an amino acid sequence having at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 95%, at least about 99%, or 100%, amino acid sequence identity to amino acids 7-166 and/or 731-1003 of SEQ ID NO: 8, or to the

corresponding portions in any of the amino acid sequences set forth as SEQ ID NOs: 1-800. Also provided by the disclosure is reproductive material of a transgenic plant, where reproductive material includes seeds, progeny plants and clonal material.

[00373] Detailed Description — Part II

[00374] The present disclosure provides methods of modulating transcription of a target nucleic acid in a host cell. The methods generally involve contacting the target nucleic acid with an enzymatically inactive Cas9 polypeptide and a single-guide RNA. The methods are useful in a variety of applications, which are also provided.

[00375] A transcriptional modulation method of the present disclosure overcomes some of the drawbacks of methods involving RNAi. A transcriptional modulation method of the present disclosure finds use in a wide variety of applications, including research applications, drug discovery (e.g., high throughput screening), target validation, industrial applications (e.g., crop engineering; microbial engineering, etc.), diagnostic applications, therapeutic applications, and imaging techniques.

[00376] Methods of Modulating Transcription

[00377] The present disclosure provides a method of selectively modulating transcription of a target DNA in a host cell. The method generally involves: a) introducing into the host cell: i) a guide RNA, or a nucleic acid comprising a nucleotide sequence encoding the guide RNA; and ii) a variant Cas9 site-directed polypeptide ("variant Cas9 polypeptide"), or a nucleic acid comprising a nucleotide sequence encoding the variant Cas9 polypeptide, where the variant Cas9 polypeptide exhibits reduced endodeoxyribonuclease activity.

[00378] The guide RNA (also referred to herein as "guide RNA"; or "gRNA") comprises: i) a first segment comprising a nucleotide sequence that is complementary to a target sequence in a target DNA; ii) a second segment that interacts with a site-directed polypeptide; and iii) a transcriptional terminator. The first segment, comprising a nucleotide sequence that is complementary to a target sequence in a target DNA, is referred to herein as a "targeting segment". The second segment, which interacts with a site-directed polypeptide, is also referred to herein as a "protein-binding sequence" or "dCas9-binding hairpin," or "dCas9 handle." By "segment" it is meant a segment/section/region of a molecule, e.g., a contiguous stretch of nucleotides in an RNA. The definition of "segment," unless otherwise specifically defined in a particular context, is not limited to a specific number of total base pairs, and may include regions of RNA molecules that are of any total length and may or may not include regions with complementarity to other molecules. As described above, guide RNA according to the present disclosure can be a single-molecule guide RNA or a two-moleculte guide RNA. The term "guide RNA" or "gRNA" is inclusive, referring both to two-molecule guide RNAs and to single-molecule guide RNAs (i.e., sgRNAs).

[00379] The variant Cas9 site-directed polypeptide comprises: i) an RNA-binding portion that interacts with the guide RNA; and an activity portion that exhibits reduced endodeoxyribonuclease activity.

[00380] The guide RNA and the variant Cas9 polypeptide form a complex in the host cell; the complex selectively modulates transcription of a target DNA in the host cell.

[00381] In some cases, a transcription modulation method of the present disclosure provides for selective modulation (e.g., reduction or increase) of a target nucleic acid in a host cell. For example, "selective" reduction of transcription of a target nucleic acid reduces transcription of the target nucleic acid by at least about 10%, at least about 20%, at least about 30%, at least about 40%, at least about 50%, at least about 60%, at least about 70%, at least about 80%, at least about 90%, or greater than 90%, compared to the level of transcription of the target nucleic acid in the absence of a guide RNA/variant Cas9 polypeptide complex. Selective reduction of transcription of a target nucleic acid reduces transcription of the target nucleic acid, but does not substantially reduce transcription of a non-target nucleic acid, e.g., transcription of a non-target nucleic acid is reduced, if at all, by less than 10% compared to the level of transcription of the non-target nucleic acid in the absence of the guide RNA/variant Cas9 polypeptide complex.

[00382] Increased transcription

[00383] "Selective" increased transcription of a target DNA can increase transcription of the target DNA by at least about 1.1 fold (e.g., at least about 1.2 fold, at least about 1.3 fold, at least about 1.4 fold, at least about 1.5 fold, at least about 1.6 fold, at least about 1.7 fold, at least about 1.8 fold, at least about 1.9 fold, at least about 2 fold, at least about 2.5 fold, at least about 3 fold, at least about 3.5 fold, at least about 4 fold, at least about 4.5 fold, at least about 5 fold, at least about 6 fold, at least about 7 fold, at least about 8 fold, at least about 9 fold, at least about 10 fold, at least about 12 fold, at least about 15 fold, or at least about 20-fold) compared to the level of transcription of the target DNA in the absence of a guide RNA/variant Cas9 polypeptide complex. Selective increase of transcription of a target DNA increases transcription of the target DNA, but does not substantially increase transcription of a non-target DNA, e.g., transcription of a non-target DNA is increased, if at all, by less than about 5-fold (e.g., less than about 4-fold, less than about 3-fold, less than about 2-fold, less than about 1.8-fold, less than about 1.6-fold, less than about 1.4-fold, less than about 1.2-fold, or less than about 1.1-fold) compared to the level of transcription of the non-targeted DNA in the absence of the guide RNA/variant Cas9 polypeptide complex.

[00384] As a non-limiting example, increased transcription can be achieved by fusing dCas9 to a heterologous sequence. Suitable fusion partners include, but are not limited to, a polypeptide that provides an activity that indirectly increases transcription by acting directly on the target DNA or on a polypeptide (e.g., a histone or other DNA-binding protein) associated with the target DNA. Suitable fusion partners include, but are not limited to, a polypeptide that provides for methyltransferase activity, demethylase

activity, acetyltransferase activity, deacetylase activity, kinase activity, phosphatase activity, ubiquitin ligase activity, deubiquitinating activity, adenylation activity, deadenylation activity, SUMOylating activity, deSUMOylating activity, ribosylation activity, deribosylation activity, myristoylation activity, or demyristoylation activity.

[00385] Additional suitable fusion partners include, but are not limited to, a polypeptide that directly provides for increased transcription of the target nucleic acid (e.g., a transcription activator or a fragment thereof, a protein or fragment thereof that recruits a transcription activator, a small molecule/drug-responsive transcription regulator, etc.).

[00386] A non-limiting example of a method using a dCas9 fusion protein to increase transcription in a prokaryote includes a modification of the bacterial one-hybrid (B1H) or two-hybrid (B2H) system. In the B1H system, a DNA binding domain (BD) is fused to a bacterial transcription activation domain (AD, e.g., the alpha subunit of the *Escherichia coli* RNA polymerase (RNAPa)). Thus, a dCas9 can be fused to a heterologous sequence comprising an AD. When the dCas9 fusion protein arrives at the upstream region of a promoter (targeted there by the guide RNA) the AD (e.g., RNAPa) of the dCas9 fusion protein recruits the RNAP holoenzyme, leading to transcription activation. In the B2H system, the BD is not directly fused to the AD; instead, their interaction is mediated by a protein-protein interaction (e.g., GAL11P - GAL4 interaction). To modify such a system for use in the methods, dCas9 can be fused to a first protein sequence that provides for protein-protein interaction (e.g., the yeast GAL11P and/or GAL4 protein) and RNAa can be fused to a second protein sequence that completes the protein-protein interaction (e.g., GAL4 if GAL11P is fused to dCas9, GAL11P if GAL4 is fused to dCas9, etc.). The binding affinity between GAL11P and GAL4 increases the efficiency of binding and transcription firing rate.

[00387] A non-limiting example of a method using a dCas9 fusion protein to increase transcription in a eukaryotes includes fusion of dCas9 to an activation domain (AD) (e.g., GAL4, herpesvirus activation protein VP16 or VP64, human nuclear factor NF-κB p65 subunit, etc.). To render the system inducible, expression of the dCas9 fusion protein can be controlled by an inducible promoter (e.g., Tet-ON, Tet-OFF, etc.). The guide RNA can be design to target known transcription response elements (e.g., promoters, enhancers, etc.), known upstream activating sequences (UAS), sequences of unknown or known function that are suspected of being able to control expression of the target DNA, etc.

[00388] Additional fusion partners

[00389] Non-limiting examples of fusion partners to accomplish increased or decreased transcription include, but are not limited to, transcription activator and transcription repressor domains (e.g., the Kriippel associated box (KRAB or SKD); the Mad mSIN3 interaction domain (SID); the ERF repressor domain (ERD), etc). In some such cases, the dCas9 fusion protein is targeted by the guide RNA to a specific location (i.e., sequence) in the target DNA and exerts locus-specific regulation such as blocking RNA polymerase binding to a promoter (which selectively inhibits transcription activator function), and/or

modifying the local chromatin status (e.g., when a fusion sequence is used that modifies the target DNA or modifies a polypeptide associated with the target DNA). In some cases, the changes are transient (e.g., transcription repression or activation). In some cases, the changes are inheritable (e.g., when epigenetic modifications are made to the target DNA or to proteins associated with the target DNA, e.g., nucleosomal histones).

[00390] In some embodiments, the heterologous sequence can be fused to the C-terminus of the dCas9 polypeptide. In some embodiments, the heterologous sequence can be fused to the N-terminus of the dCas9 polypeptide. In some embodiments, the heterologous sequence can be fused to an internal portion (i.e., a portion other than the N- or C- terminus) of the dCas9 polypeptide.

[00391] The biological effects of a method using a dCas9 fusion protein can be detected by any convenient method (e.g., gene expression assays; chromatin-based assays, e.g., Chromatin immunoPrecipitation (ChiP), Chromatin in vivo Assay (CiA), etc.; and the like).

[00392] In some cases, a method involves use of two or more different guide RNAs. For example, two different guide RNAs can be used in a single host cell, where the two different guide RNAs target two different target sequences in the same target nucleic acid.

[00393] Thus, for example, a transcriptional modulation method can further comprise introducing into the host cell a second guide RNA, or a nucleic acid comprising a nucleotide sequence encoding the second guide RNA, where the second guide RNA comprises: i) a first segment comprising a nucleotide sequence that is complementary to a second target sequence in the target DNA; ii) a second segment that interacts with the site-directed polypeptide; and iii) a transcriptional terminator. In some cases, use of two different guide RNAs targeting two different targeting sequences in the same target nucleic acid provides for increased modulation (e.g., reduction or increase) in transcription of the target nucleic acid.

[00394] As another example, two different guide RNAs can be used in a single host cell, where the two different guide RNAs target two different target nucleic acids. Thus, for example, a transcriptional modulation method can further comprise introducing into the host cell a second guide RNA, or a nucleic acid comprising a nucleotide sequence encoding the second guide RNA, where the second guide RNA comprises: i) a first segment comprising a nucleotide sequence that is complementary to a target sequence in at least a second target DNA; ii) a second segment that interacts with the site-directed polypeptide; and iii) a transcriptional terminator.

[00395] In some embodiments, a nucleic acid (e.g., a guide RNA, e.g., a single-molecule guide RNA, an activator-RNA, a targeter-RNA, etc.; a donor polynucleotide; a nucleic acid encoding a site-directed modifying polypeptide; etc.) comprises a modification or sequence that provides for an additional desirable feature (e.g., modified or regulated stability; subcellular targeting; tracking, e.g., a fluorescent label; a binding site for a protein or protein complex; etc.). Non-limiting examples include: a 5' cap (e.g., a 7-

methylguanylate cap (m^7G)); a 3' polyadenylated tail (i.e., a 3' poly(A) tail); a riboswitch sequence or an aptamer sequence (e.g., to allow for regulated stability and/or regulated accessibility by proteins and/or protein complexes); a terminator sequence; a sequence that forms a dsRNA duplex (i.e., a hairpin)); a modification or sequence that targets the RNA to a subcellular location (e.g., nucleus, mitochondria, chloroplasts, and the like); a modification or sequence that provides for tracking (e.g., direct conjugation to a fluorescent molecule, conjugation to a moiety that facilitates fluorescent detection, a sequence that allows for fluorescent detection, etc.); a modification or sequence that provides a binding site for proteins (e.g., proteins that act on DNA, including transcriptional activators, transcriptional repressors, DNA methyltransferases, DNA demethylases, histone acetyltransferases, histone deacetylases, and the like); and combinations thereof.

[00396] DNA-targeting segment

[00397] The DNA-targeting segment (or "DNA-targeting sequence") of a guide RNA comprises a nucleotide sequence that is complementary to a specific sequence within a target DNA (the complementary strand of the target DNA).

[00398] In other words, the DNA-targeting segment of a guide RNA interacts with a target DNA in a sequence-specific manner via hybridization (i.e., base pairing). As such, the nucleotide sequence of the DNA-targeting segment may vary and determines the location within the target DNA that the guide RNA and the target DNA will interact. The DNA-targeting segment of a guide RNA can be modified (e.g., by genetic engineering) to hybridize to any desired sequence within a target DNA.

[00399] Stability control sequence (e.g., transcriptional terminator segment)

[00400] A stability control sequence influences the stability of an RNA (e.g., a guide RNA, a targeter-RNA, an activator-RNA, etc.). One example of a suitable stability control sequence is a transcriptional terminator segment (i.e., a transcription termination sequence). A transcriptional terminator segment of a guide RNA can have a total length of from about 10 nucleotides to about 100 nucleotides, e.g., from about 10 nucleotides (nt) to about 20 nt, from about 20 nt to about 30 nt, from about 30 nt to about 40 nt, from about 40 nt to about 50 nt, from about 50 nt to about 60 nt, from about 60 nt to about 70 nt, from about 70 nt to about 80 nt, from about 80 nt to about 90 nt, or from about 90 nt to about 100 nt. For example, the transcriptional terminator segment can have a length of from about 15 nucleotides (nt) to about 80 nt, from about 15 nt to about 50 nt, from about 15 nt to about 40 nt, from about 15 nt to about 30 nt or from about 15 nt to about 25 nt.

[00401] In some cases, the transcription termination sequence is one that is functional in a eukaryotic cell. In some cases, the transcription termination sequence is one that is functional in a prokaryotic cell.

[00402] Nucleotide sequences that can be included in a stability control sequence (e.g., transcriptional termination segment, or in any segment of the guide RNA to provide for increased stability)

include, for example, 5'-UAAUCCCACAGCCGCCAGUUCCGCUGGCGGCAUUUU-5' (a Rho-independent *trp* termination site).

[00403] Additional sequences

[00404] In some embodiments, a guide RNA comprises at least one additional segment at either the 5' or 3' end. For example, a suitable additional segment can comprise a 5' cap (e.g., a 7-methylguanylate cap (m^7G)); a 3' polyadenylated tail (i.e., a 3' poly(A) tail); a riboswitch sequence (e.g., to allow for regulated stability and/or regulated accessibility by proteins and protein complexes); a sequence that forms a dsRNA duplex (i.e., a hairpin)); a sequence that targets the RNA to a subcellular location (e.g., nucleus, mitochondria, chloroplasts, and the like); a modification or sequence that provides for tracking (e.g., direct conjugation to a fluorescent molecule, conjugation to a moiety that facilitates fluorescent detection, a sequence that allows for fluorescent detection, etc.); a modification or sequence that provides a binding site for proteins (e.g., proteins that act on DNA, including transcriptional activators, transcriptional repressors, DNA methyltransferases, DNA demethylases, histone acetyltransferases, histone deacetylases, and the like) a modification or sequence that provides for increased, decreased, and/or controllable stability; and combinations thereof.

[00405] Multiple simultaneous guide RNAs

[00406] In some embodiments, multiple guide RNAs are used simultaneously in the same cell to simultaneously modulate transcription at different locations on the same target DNA or on different target DNAs. In some embodiments, two or more guide RNAs target the same gene or transcript or locus. In some embodiments, two or more guide RNAs target different unrelated loci. In some embodiments, two or more guide RNAs target different, but related loci.

[00407] Because the guide RNAs are small and robust they can be simultaneously present on the same expression vector and can even be under the same transcriptional control if so desired. In some embodiments, two or more (e.g., 3 or more, 4 or more, 5 or more, 10 or more, 15 or more, 20 or more, 25 or more, 30 or more, 35 or more, 40 or more, 45 or more, or 50 or more) guide RNAs are simultaneously expressed in a target cell (from the same or different vectors). The expressed guide RNAs can be differently recognized by Cas9 proteins from different bacteria, such as *S. pyogenes, S. thermophilus, L. innocua, and N. meningitidis.*

[00408] In some cases, multiple guide RNAs can be encoded in an array mimicking naturally occurring CRISPR arrays of targeter RNAs and corresponding tracrRNAs (activator RNAs). The targeting segments are encoded as approximately 30 nucleotide long sequences (can be about 16 to about 100 nt) and are separated by CRISPR repeat sequences. In some cases, the array and tracrRNAs are introduced to a cell by DNAs encoding the RNAs. In some cases, they are introduced to the cell as RNAs.

[00409] To express multiple guide RNAs, an artificial RNA processing system mediated by the Csy4 endoribonuclease can be used. Multiple guide RNAs can be concatenated into a tandem array on a precursor transcript (e.g., expressed from a U6 promoter), and separated by Csy4-specific RNA sequence. Co-expressed Csy4 protein cleaves the precursor transcript into multiple guide RNAs. Advantages for using an RNA processing system include: first, there is no need to use multiple promoters; second, since all guide RNAs are processed from a precursor transcript, their concentrations are normalized for similar dCas9-binding.

[00410] Csy4 is a small endoribonuclease (RNase) protein derived from bacteria *Pseudomonas aeruginosa.* Csy4 specifically recognizes a minimal 17-bp RNA hairpin, and exhibits rapid (<1 min) and highly efficient (>99.9%) RNA cleavage. Unlike most RNases, the cleaved RNA fragment remains stable and functionally active. The Csy4-based RNA cleavage can be repurposed into an artificial RNA processing system. In this system, the 17-bp RNA hairpins are inserted between multiple RNA fragments that are transcribed as a precursor transcript from a single promoter. Co-expression of Csy4 is effective in generating individual RNA fragments.

[00411] Site-directed polypeptide

[00412] As noted above, a guide RNA and a variant Cas9 site-directed polypeptide form a complex. The guide RNA provides target specificity to the complex by comprising a nucleotide sequence that is complementary to a sequence of a target DNA. The variant Cas9 site-directed polypeptide has reduced endodeoxyribonuclease activity. For example, a variant Cas9 site-directed polypeptide suitable for use in a transcription modulation method of the present disclosure exhibits less than about 20%, less than about 15%, less than about 10%, less than about 5%, less than about 1%, or less than about 0.1%, of the endodeoxyribonuclease activity of a wild-type Cas9 polypeptide, e.g., a wild-type Cas9 polypeptide comprising an amino acid sequence set out in SEQ ID NO:8. In some embodiments, the variant Cas9 site-directed polypeptide has substantially no detectable endodeoxyribonuclease activity. In some embodiments when a site-directed polypeptide has reduced catalytic activity (e.g., when a SEQ ID NO: 8 S. pyogenes Cas9 protein has a D10, G12, G17, E762, H840, N863, H982, H983, A984, D986, and/or a A987 mutation, e.g., D10A, G12A, G17A, E762A, H840A, N863A, H982A, H983A, A984A, and/or D986A), the polypeptide can still bind to target DNA in a site-specific manner (because it is still guided to a target DNA sequence by a guide RNA) as long as it retains the ability to interact with the guide RNA.

[00413] In some cases, a suitable variant Cas9 site-directed polypeptide comprises an amino acid sequence having at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 95%, at least about 99% or 100% amino acid sequence identity to amino acids 7-166 and/or 731-1003 of SEQ ID NO: 8, or to the corresponding portions in any one of the amino acid sequences SEQ ID NOs: 1-800.

[00414] In some cases, the variant Cas9 site-directed polypeptide is a nickase that can cleave the complementary strand of the target DNA but has reduced ability to cleave the non-complementary strand of the target DNA. For example, the variant Cas9 site-directed polypeptide can have a mutation (amino acid substitution) that reduces the function of the RuvC domain. As a non-limiting example, in some cases, the variant Cas9 site-directed polypeptide is a D10A (aspartate to alanine) mutation of SEQ ID NO: 8 (or the corresponding mutation of any of the amino acid sequences set forth in SEQ ID NOs: 1-800).

[00415] In some cases, the variant Cas9 site-directed polypeptide in a nickase that can cleave the non-complementary strand of the target DNA but has reduced ability to cleave the complementary strand of the target DNA. For example, the variant Cas9 site-directed polypeptide can have a mutation (amino acid substitution) that reduces the function of the HNH domain (RuvC/HNH/RuvC domain motifs, "domain 2"). As a non-limiting example, in some cases, the variant Cas9 site-directed polypeptide is a H840A (histidine to alanine at amino acid position 840 of SEQ ID NO:8) or the corresponding mutation of any of the amino acid sequences set forth in SEQ ID NOs: 1-800).

[00416] In some cases, the variant Cas9 site-directed polypeptide has a reduced ability to cleave both the complementary and the non-complementary strands of the target DNA. As a non-limiting example, in some cases, the variant Cas9 site-directed polypeptide harbors both D10A and H840A mutations of SEQ ID NO: 8 (or the corresponding mutations of any of the amino acid sequences set forth in SEQ ID NOs: 1-800). Other residues can be mutated to achieve the same effect (i.e. inactivate one or the other nuclease portions). As non-limiting examples, S. pyogenes Cas9 residues D10, G12, G17, E762, H840, N863, H982, H983, A984, D986, and/or A987 of SEQ ID NO: 8 (or the corresponding mutations of any of the proteins set forth as SEQ ID NOs: 1-800) can be altered (i.e., substituted) (see Table 1 for examples of the conservation of Cas9 amino acid residues). Also, mutations other than alanine substitutions are contemplated.

[00417] In some embodiments, a variant Cas9 endonuclease comprises one or more mutations corresponding to a S. pyogenes Cas9 mutation E762A, HH983AA or D986A in SEQ ID NO: 8. In some embodiments, the modified Cas 9 endonuclease further comprises one or more mutations corresponding to a S. pyogenes Cas9 mutation D10A, H840A, G12A, G17A, N854A, N863A, N982A or A984A in SEQ ID NO: 8.

[00418] In some cases, the variant Cas9 site-directed polypeptide is a fusion polypeptide (a "variant Cas9 fusion polypeptide"), i.e., a fusion polypeptide comprising: i) a variant Cas9 site-directed polypeptide; and ii) a covalently linked heterologous polypeptide (also referred to as a "fusion partner").

[00419] The heterologous polypeptide may exhibit an activity (e.g., enzymatic activity) that will also be exhibited by the variant Cas9 fusion polypeptide (e.g., methyltransferase activity, acetyltransferase

activity, kinase activity, ubiquitinating activity, etc.). A heterologous nucleic acid sequence may be linked to another nucleic acid sequence (e.g., by genetic engineering) to generate a chimeric nucleotide sequence encoding a chimeric polypeptide. In some embodiments, a variant Cas9 fusion polypeptide is generated by fusing a variant Cas9 polypeptide with a heterologous sequence that provides for subcellular localization (i.e., the heterologous sequence is a subcellular localization sequence, e.g., a nuclear localization signal (NLS) for targeting to the nucleus; a mitochondrial localization signal for targeting to the mitochondria; a chloroplast localization signal for targeting to a chloroplast; an ER retention signal; and the like). In some embodiments, the heterologous sequence can provide a tag (i.e., the heterologous sequence is a detectable label) for ease of tracking and/or purification (e.g., a fluorescent protein, e.g., green fluorescent protein (GFP), YFP, RFP, CFP, mCherry, tdTomato, and the like; a histidine tag, e.g., a 6XHis tag; a hemagglutinin (HA) tag; a FLAG tag; a Myc tag; and the like). In some embodiments, the heterologous sequence can provide for increased or decreased stability (i.e., the heterologous sequence is a stability control peptide, e.g., a degron, which in some cases is controllable (e.g., a temperature sensitive or drug controllable degron sequence, see below). In some embodiments, the heterologous sequence can provide for increased or decreased transcription from the target DNA (i.e., the heterologous sequence is a transcription modulation sequence, e.g., a transcription factor/activator or a fragment thereof, a protein or fragment thereof that recruits a transcription factor/activator, a transcription repressor or a fragment thereof, a protein or fragment thereof that recruits a transcription repressor, a small molecule/drug-responsive transcription regulator, etc.). In some embodiments, the heterologous sequence can provide a binding domain (i.e., the heterologous sequence is a protein binding sequence, e.g., to provide the ability of a chimeric dCas9 polypeptide to bind to another protein of interest, e.g., a DNA or histone modifying protein, a transcription factor or transcription repressor, a recruiting protein, etc.).

[00420] Suitable fusion partners that provide for increased or decreased stability include, but are not limited to degron sequences. Degrons are readily understood by one of ordinary skill in the art to be amino acid sequences that control the stability of the protein of which they are part. For example, the stability of a protein comprising a degron sequence is controlled at least in part by the degron sequence. In some cases, a suitable degron is constitutive such that the degron exerts its influence on protein stability independent of experimental control (i.e., the degron is not drug inducible, temperature inducible, etc.) In some cases, the degron provides the variant Cas9 polypeptide with controllable stability such that the variant Cas9 polypeptide can be turned "on" (i.e., stable) or "off" (i.e., unstable, degraded) depending on the desired conditions. For example, if the degron is a temperature sensitive degron, the variant Cas9 polypeptide may be functional (i.e., "on", stable) below a threshold temperature (e.g., 42°C, 41°C, 40°C, 39°C, 38°C, 37°C, 36°C, 35°C, 34°C, 33°C, 32°C, 31°C, 30°C, etc.) but non-functional (i.e., "off', degraded) above the threshold temperature. As another example, if the degron is a drug inducible degron, the presence or absence of drug can switch the protein from an "off' (i.e., unstable) state to an "on" (i.e., stable) state or vice versa. An exemplary drug inducible degron is derived from the FKBP12

protein. The stability of the degron is controlled by the presence or absence of a small molecule that binds to the degron.

[00421] Examples of suitable degrons include, but are not limited to those degrons controlled by Shield-1, DHFR, auxins, and/or temperature. Non-limiting examples of suitable degrons are known in the art (e.g., Dohmen et al., Science, 1994. 263(5151): p. 1273-1276: Heat-inducible degron: a method for constructing temperature-sensitive mutants; Schoeber et al., Am J Physiol Renal Physiol. 2009 Jan;296(1):F204-11 : Conditional fast expression and function of multimeric TRPV5 channels using Shield-1 ; Chu et al., Bioorg Med Chem Lett. 2008 Nov 15;18(22):5941-4: Recent progress with FKBP-derived destabilizing domains ; Kanemaki, Pflugers Arch. 2012 Dec 28: Frontiers of protein expression control with conditional degrons; Yang et al., Mol Cell. 2012 Nov 30;48(4):487-8: Titivated for destruction: the methyl degron; Barbour et al., Biosci Rep. 2013 Jan 18;33(1).: Characterization of the bipartite degron that regulates ubiquitin-independent degradation of thymidylate synthase; and Greussing et al., J Vis Exp. 2012 Nov 10;(69): Monitoring of ubiquitin-proteasome activity in living cells using a Degron (dgn)-destabilized green fluorescent protein (GFP)-based reporter protein; all of which are hereby incorporated in their entirety by reference).

[00422] Exemplary degron sequences have been well-characterized and tested in both cells and animals. Thus, fusing Cas9 to a degron sequence produces a "tunable" and "inducible" Cas9 polypeptide. Any of the fusion partners described herein can be used in any desirable combination. As one non-limiting example to illustrate this point, a Cas9 fusion protein can comprise a YFP sequence for detection, a degron sequence for stability, and transcription activator sequence to increase transcription of the target DNA. Furthermore, the number of fusion partners that can be used in a Cas9 fusion protein is unlimited. In some cases, a Cas9 fusion protein comprises one or more (e.g. two or more, three or more, four or more, or five or more) heterologous sequences.

[00423] Suitable fusion partners include, but are not limited to, a polypeptide that provides for methyltransferase activity, demethylase activity, acetyltransferase activity, deacetylase activity, kinase activity, phosphatase activity, ubiquitin ligase activity, deubiquitinating activity, adenylation activity, deadenylation activity, SUMOylating activity, deSUMOylating activity, ribosylation activity, deribosylation activity, myristoylation activity, or demyristoylation activity, any of which can be directed at modifying the DNA directly (e.g., methylation of DNA) or at modifying a DNA-associated polypeptide (e.g., a histone or DNA binding protein). Further suitable fusion partners include, but are not limited to boundary elements (e.g., CTCF), proteins and fragments thereof that provide periphery recruitment (e.g., Lamin A, Lamin B, etc.), and protein docking elements (e.g., FKBP/FRB, Pil 1/Aby 1, etc.).

[00424] In some embodiments, a site-directed modifying polypeptide can be codon-optimized. This type of optimization is known in the art and entails the mutation of foreign-derived DNA to mimic the codon preferences of the intended host organism or cell while encoding the same protein. Thus, the

codons are changed, but the encoded protein remains unchanged. For example, if the intended target cell was a human cell, a human codon-optimized dCas9 (or dCas9 variant) would be a suitable site-directed modifying polypeptide. As another non-limiting example, if the intended host cell were a mouse cell, than a mouse codon-optimized Cas9 (or variant, e.g., enzymatically inactive variant) would be a suitable Cas9 site-directed polypeptide. While codon optimization is not required, it is acceptable and may be preferable in certain cases.

[00425] Polyadenylation signals can also be chosen to optimize expression in the intended host.

[00426] Host cells

[00427] A method of the present disclosure to modulate transcription may be employed to induce transcriptional modulation in mitotic or post-mitotic cells *in vivo* and/or *ex vivo* and/or *in vitro.* Because the guide RNA provides specificity by hybridizing to target DNA, a mitotic and/or post-mitotic cell can be any of a variety of host cell, where suitable host cells include, but are not limited to, a bacterial cell; an archaeal cell; a single-celled eukaryotic organism; a plant cell; an algal cell, e.g., *Botryococcus braunii, Chlamydomonas reinhardtii, Nannochloropsis gaditana, Chlorella pyrenoidosa, Sargassum patens, C. agardh,* and the like; a fungal cell; an animal cell; a cell from an invertebrate animal (e.g., an insect, a cnidarian, an echinoderm, a nematode, etc.); a eukaryotic parasite (e.g., a malarial parasite, e.g., *Plasmodium fakiparum;* a helminth; etc.); a cell from a vertebrate animal (e.g., fish, amphibian, reptile, bird, mammal); a mammalian cell, e.g., a rodent cell, a human cell, a non-human primate cell, etc. Suitable host cells include naturally-occurring cells; genetically modified cells (e.g., cells genetically modified in a laboratory, e.g., by the "hand of man"); and cells manipulated *in vitro* in any way. In some cases, a host cell is isolated.

[00428] Any type of cell may be of interest (e.g. a stem cell, e.g. an embryonic stem (ES) cell, an induced pluripotent stem (iPS) cell, a germ cell; a somatic cell, e.g. a fibroblast, a hematopoietic cell, a neuron, a muscle cell, a bone cell, a hepatocyte, a pancreatic cell; an in vitro or in vivo embryonic cell of an embryo at any stage, e.g., a 1-cell, 2-cell, 4-cell, 8-cell, etc. stage zebrafish embryo; etc.). Cells may be from established cell lines or they may be primary cells, where "primary cells", "primary cell lines", and "primary cultures" are used interchangeably herein to refer to cells and cells cultures that have been derived from a subject and allowed to grow in vitro for a limited number of passages, i.e. splittings, of the culture. For example, primary cultures include cultures that may have been passaged 0 times, 1 time, 2 times, 4 times, 5 times, 10 times, or 15 times, but not enough times go through the crisis stage. Primary cell lines can be are maintained for fewer than 10 passages *in vitro.* Target cells are in many embodiments unicellular organisms, or are grown in culture.

[00429] If the cells are primary cells, such cells may be harvest from an individual by any convenient method. For example, leukocytes may be conveniently harvested by apheresis, leukocytapheresis, density gradient separation, etc., while cells from tissues such as skin, muscle, bone marrow, spleen,

liver, pancreas, lung, intestine, stomach, etc. are most conveniently harvested by biopsy. An appropriate solution may be used for dispersion or suspension of the harvested cells. Such solution will generally be a balanced salt solution, e.g. normal saline, phosphate-buffered saline (PBS), Hank's balanced salt solution, etc., conveniently supplemented with fetal calf serum or other naturally occurring factors, in conjunction with an acceptable buffer at low concentration, e.g., from 5-25 mM. Convenient buffers include HEPES, phosphate buffers, lactate buffers, etc. The cells may be used immediately, or they may be stored, frozen, for long periods of time, being thawed and capable of being reused. In such cases, the cells will usually be frozen in 10% dimethyl sulfoxide (DMSO), 50% serum, 40% buffered medium, or some other such solution as is commonly used in the art to preserve cells at such freezing temperatures, and thawed in a manner as commonly known in the art for thawing frozen cultured cells.

[00430] Introducing nucleic acid into a host cell

[00431] A guide RNA, or a nucleic acid comprising a nucleotide sequence encoding same, can be introduced into a host cell by any of a variety of well-known methods. Similarly, where a method involves introducing into a host cell a nucleic acid comprising a nucleotide sequence encoding a variant Cas9 site-directed polypeptide, such a nucleic acid can be introduced into a host cell by any of a variety of well-known methods.

[00432] Methods of introducing a nucleic acid into a host cell are known in the art, and any known method can be used to introduce a nucleic acid (e.g., an expression construct) into a stem cell or progenitor cell. Suitable methods include, include e.g., viral or bacteriophage infection, transfection, conjugation, protoplast fusion, lipofection, electroporation, calcium phosphate precipitation, polyethyleneimine (PEI)-mediated transfection, DEAE-dextran mediated transfection, liposome-mediated transfection, particle gun technology, calcium phosphate precipitation, direct micro injection, nanoparticle-mediated nucleic acid delivery (see, e.g., Panyam et., al Adv Drug Deliv Rev. 2012 Sep 13. pii: 50169-409X(12)00283-9. doi: 10.1016/j.addr.2012.09.023), and the like.

[00433] Nucleic Acids

[00434] The present disclosure provides an isolated nucleic acid comprising a nucleotide sequence encoding a guide RNA. In some cases, a nucleic acid also comprises a nucleotide sequence encoding a variant Cas9 site-directed polypeptide.

[00435] In some embodiments, a method involves introducing into a host cell (or a population of host cells) one or more nucleic acids comprising nucleotide sequences encoding a guide RNA and/or a variant Cas9 site-directed polypeptide. In some embodiments a cell comprising a target DNA is *in vitro*. In some embodiments a cell comprising a target DNA is *in vivo*. Suitable nucleic acids comprising nucleotide sequences encoding a guide RNA and/or a site-directed polypeptide include expression vectors, where

an expression vector comprising a nucleotide sequence encoding a guide RNA and/or a site-directed polypeptide is a "recombinant expression vector."

[00436] In some embodiments, the recombinant expression vector is a viral construct, e.g., a recombinant adeno-associated virus construct (see, e.g., U.S. Patent No. 7,078,387), a recombinant adenoviral construct, a recombinant lentiviral construct, a recombinant retroviral construct, etc.

Suitable expression vectors include, but are not limited to, viral vectors (e.g. viral vectors based on vaccinia virus; poliovirus; adenovirus (see, e.g., Li et al., Invest Opthalmol Vis Sci 35:2543 2549, 1994; Borras et al., Gene Ther 6:515 524, 1999; Li and Davidson, PNAS 92:7700 7704, 1995; Sakamoto et al., H Gene Ther 5:1088 1097, 1999; WO 94/12649, WO 93/03769; WO 93/19191; WO 94/28938; WO 95/11984 and WO 95/00655); adeno-associated virus (see, e.g., Ali et al., Hum Gene Ther 9:81 86, 1998, Flannery et al., PNAS 94:6916 6921, 1997; Bennett et al., Invest Opthalmol Vis Sci 38:2857 2863, 1997; Jomary et al., Gene Ther 4:683-690, 1997, Rolling et al., Hum Gene Ther 10:641 648, 1999; Ali et al., Hum Mol Genet 5:591 594, 1996; Srivastava in WO 93/09239, Samulski et al., J. Vir. (1989) 63:3822-3828; Mendelson et al., Virol. (1988) 166:154-165; and Flotte et al., PNAS (1993) 90:10613-10617); SV40; herpes simplex virus; human immunodeficiency virus (see, e.g., Miyoshi et al., PNAS 94:10319 23, 1997; Takahashi et al., J Virol 73:7812 7816, 1999); a retroviral vector (e.g., Murine Leukemia Virus, spleen necrosis virus, and vectors derived from retroviruses such as Rous Sarcoma Virus, Harvey Sarcoma Virus, avian leukosis virus, a lentivirus, human immunodeficiency virus, myeloproliferative sarcoma virus, and mammary tumor virus); and the like.

[00437] Numerous suitable expression vectors are known to those of skill in the art, and many are commercially available. The following vectors are provided by way of example; for eukaryotic host cells: pXT1, pSG5 (Stratagene), pSVK3, pBPV, pMSG, and pSVLSV40 (Pharmacia). However, any other vector may be used so long as it is compatible with the host cell.

[00438] Depending on the host/vector system utilized, any of a number of suitable transcription and translation control elements, including constitutive and inducible promoters, transcription enhancer elements, transcription terminators, etc. may be used in the expression vector (see e.g., Bitter et al. (1987) *Methods in Enzymology,* 153:516-544).

[00439] In some embodiments, a nucleotide sequence encoding a guide RNA and/or a variant Cas9 site-directed polypeptide is operably linked to a control element, e.g., a transcriptional control element, such as a promoter. The transcriptional control element may be functional in either a eukaryotic cell, e.g., a mammalian cell; or a prokaryotic cell (e.g., bacterial or archaeal cell). In some embodiments, a nucleotide sequence encoding a guide RNA and/or a variant Cas9 site-directed polypeptide is operably linked to multiple control elements that allow expression of the nucleotide sequence encoding a guide RNA and/or a variant Cas9 site-directed polypeptide in both prokaryotic and eukaryotic cells.

[00440] A promoter can be a constitutively active promoter (i.e., a promoter that is constitutively in an active/"ON" state), it may be an inducible promoter (i.e., a promoter whose state, active/"ON" or inactive/"OFF", is controlled by an external stimulus, e.g., the presence of a particular temperature, compound, or protein.), it may be a spatially restricted promoter (i.e., transcriptional control element, enhancer, etc.)(e.g., tissue specific promoter, cell type specific promoter, etc.), and it may be a temporally restricted promoter (i.e., the promoter is in the "ON" state or "OFF" state during specific stages of embryonic development or during specific stages of a biological process, e.g., hair follicle cycle in mice).

[00441] Suitable promoters can be derived from viruses and can therefore be referred to as viral promoters, or they can be derived from any organism, including prokaryotic or eukaryotic organisms. Suitable promoters can be used to drive expression by any RNA polymerase (e.g., pol I, pol II, pol III). Exemplary promoters include, but are not limited to the SV40 early promoter, mouse mammary tumor virus long terminal repeat (LTR) promoter; adenovirus major late promoter (Ad MLP); a herpes simplex virus (HSV) promoter, a cytomegalovirus (CMV) promoter such as the CMV immediate early promoter region (CMVIE), a rous sarcoma virus (RSV) promoter, a human U6 small nuclear promoter (U6) (Miyagishi et al. , Nature Biotechnology 20, 497 - 500 (2002)), an enhanced U6 promoter (e.g., Xia et al., Nucleic Acids Res. 2003 Sep 1;31(17)), a human H1 promoter (H1), and the like.

[00442] Examples of inducible promoters include, but are not limited toT7 RNA polymerase promoter, T3 RNA polymerase promoter, Isopropyl-beta-D-thiogalactopyranoside (IPTG)-regulated promoter, lactose induced promoter, heat shock promoter, Tetracycline-regulated promoter (e.g., Tet-ON, Tet-OFF, etc.), Steroid-regulated promoter, Metal-regulated promoter, estrogen receptor-regulated promoter, etc. Inducible promoters can therefore be regulated by molecules including, but not limited to, doxycycline; RNA polymerase, e.g., T7 RNA polymerase; an estrogen receptor; an estrogen receptor fusion; etc.

[00443] In some embodiments, the promoter is a spatially restricted promoter (i.e., cell type specific promoter, tissue specific promoter, etc.) such that in a multi-cellular organism, the promoter is active (i.e., "ON") in a subset of specific cells. Spatially restricted promoters may also be referred to as enhancers, transcriptional control elements, control sequences, etc. Any convenient spatially restricted promoter may be used and the choice of suitable promoter (e.g., a brain specific promoter, a promoter that drives expression in a subset of neurons, a promoter that drives expression in the germline, a promoter that drives expression in the lungs, a promoter that drives expression in muscles, a promoter that drives expression in islet cells of the pancreas, etc.) will depend on the organism. For example, various spatially restricted promoters are known for plants, flies, worms, mammals, mice, etc. Thus, a spatially restricted promoter can be used to regulate the expression of a nucleic acid encoding a site-directed polypeptide in a wide variety of different tissues and cell types, depending on the organism. Some spatially restricted promoters are also temporally restricted such that the promoter is in the "ON" state or "OFF' state during specific stages of embryonic development or during specific stages of a biological process (e.g., hair follicle cycle in mice).

[00444] For illustration purposes, examples of spatially restricted promoters include, but are not limited to, neuron-specific promoters, adipocyte-specific promoters, cardiomyocyte-specific promoters, smooth muscle-specific promoters, photoreceptor-specific promoters, etc. Neuron-specific spatially restricted promoters include, but are not limited to, a neuron-specific enolase (NSE) promoter (see, e.g., EMBL HSENO2, X51956); an aromatic amino acid decarboxylase (AADC) promoter; a neurofilament promoter (see, e.g., GenBank HUMNFL, L04147); a synapsin promoter (see, e.g., GenBank HUMSYNIB, M55301); a thy-1 promoter (see, e.g., Chen et al. (1987) *Cell* 51:7-19; and Llewellyn, et al. (2010) *Nat. Med.* 16(10):1161-1166); a serotonin receptor promoter (see, e.g., GenBank S62283); a tyrosine hydroxylase promoter (TH) (see, e.g., Oh et al. (2009) *Gene Ther* 16:437; Sasaoka et al. (1992) *Mol. Brain Res.* 16:274; Boundy et al. (1998) *J. Neurosci.* 18:9989; and Kaneda et al. (1991) *Neuron* 6:583-594); a GnRH promoter (see, e.g., Radovick et al. (1991) *Proc. Natl. Acad. Sci. USA* 88:3402-3406); an L7 promoter (see, e.g., Oberdick et al. (1990) *Science* 248:223-226); a DNMT promoter (see, e.g., Bartge et al. (1988) *Proc. Natl. Acad. Sci. USA* 85:3648-3652); an enkephalin promoter (see, e.g., Comb et al. (1988) *EMBO J.* 17:3793-3805); a myelin basic protein (MBP) promoter; a Ca^{2+}-calmodulin-dependent protein kinase II-alpha (CamKIIa) promoter (see, e.g., Mayford et al. (1996) *Proc. Natl. Acad. Sci. USA* 93:13250; and Casanova et al. (2001) *Genesis* 31:37); a CMV enhancer/platelet-derived growth factor-0 promoter (see, e.g., Liu et al. (2004) *Gene Therapy* 11:52-60); and the like.

[00445] Adipocyte-specific spatially restricted promoters include, but are not limited to aP2 gene promoter/enhancer, e.g., a region from -5.4 kb to +21 bp of a human aP2 gene (see, e.g., Tozzo et al. (1997) *Endocrinol.* 138:1604; Ross et al. (1990) *Proc. Natl. Acad. Sci. USA* 87:9590; and Pavjani et al. (2005) *Nat. Med.* 11:797); a glucose transporter-4 (GLUT4) promoter (see, e.g., Knight et al. (2003) *Proc. Natl. Acad. Sci. USA* 100:14725); a fatty acid translocase (FAT/CD36) promoter (see, e.g., Kuriki et al. (2002) *Biol. Pharm. Bull.* 25:1476; and Sato et al. (2002) *J. Biol. Chem.* 277:15703); a stearoyl-CoA desaturase-1 (SCD1) promoter (Tabor et al. (1999) *J. Biol. Chem.* 274:20603); a leptin promoter (see, e.g., Mason et al. (1998) *Endocrinol.* 139:1013; and Chen et al. (1999) *Biochem. Biophys. Res. Comm.* 262:187); an adiponectin promoter (see, e.g., Kita et al. (2005) *Biochem. Biophys. Res. Comm.* 331:484; and Chakrabarti (2010) *Endocrinol.* 151:2408); an adipsin promoter (see, e.g., Platt et al. (1989) *Proc. Natl. Acad. Sci. USA* 86:7490); a resistin promoter (see, e.g., Seo et al. (2003) *Molec. Endocrinol.* 17:1522); and the like.

[00446] Cardiomyocyte-specific spatially restricted promoters include, but are not limited to control sequences derived from the following genes: myosin light chain-2, a-myosin heavy chain, AE3, cardiac troponin C, cardiac actin, and the like. Franz et al. (1997) Cardiovasc. Res. 35:560-566; Robbins et al. (1995) Ann. N.Y. Acad. Sci. 752:492-505; Linn et al. (1995) Circ. Res. 76:584591; Parmacek et al. (1994) Mol. Cell. Biol. 14:1870-1885; Hunter et al. (1993) Hypertension 22:608-617; and Sartorelli et al. (1992) Proc. Natl. Acad. Sci. USA 89:4047-4051.

[00447] Smooth muscle-specific spatially restricted promoters include, but are not limited to an SM22a promoter (see, e.g., Akyilrek et al. (2000) *Mol. Med.* 6:983; and U.S. Patent No. 7,169,874); a smoothelin promoter (see, e.g., WO 2001/018048); an a-smooth muscle actin promoter; and the like. For example, a 0.4 kb region of the SM22a promoter, within which lie two CArG elements, has been shown to mediate vascular smooth muscle cell-specific expression (see, e.g., Kim, et al. (1997) Mol. Cell. Biol. 17, 2266-2278; Li, et al., (1996) J. Cell Biol. 132, 849-859; and Moessler, et al. (1996) Development 122, 2415-2425).

[00448] Photoreceptor-specific spatially restricted promoters include, but are not limited to, a rhodopsin promoter; a rhodopsin kinase promoter (Young et al. (2003) *Ophthalmol. Vis. Sci.* 44:4076); a beta phosphodiesterase gene promoter (Nicoud et al. (2007) *J. Gene Med.* 9:1015); a retinitis pigmentosa gene promoter (Nicoud et al. (2007) *supra);* an interphotoreceptor retinoid-binding protein (IRBP) gene enhancer (Nicoud et al. (2007) *supra);* an IRBP gene promoter (Yokoyama et al. (1992) *Exp Eye Res.* 55:225); and the like.

[00449] Libraries

[00450] The present disclosure provides a library of guide RNAs. The present disclosure provides a library of nucleic acids comprising nucleotides encoding guide RNAs. A library of nucleic acids comprising nucleotides encoding guide RNAs can comprises a library of recombinant expression vectors comprising nucleotides encoding the guide RNAs.

[00451] A library can comprise from about 10 individual members to about 10^{12} individual members; e.g., a library can comprise from about 10 individual members to about 10^2 individual members, from about 10^2 individual members to about 10^3 individual members, from about 10^3 individual members to about 10^5 individual members, from about 10^5 individual members to about 10^7 individual members, from about 10^7 individual members to about 10^9 individual members, or from about 10^9 individual members to about 10^{12} individual members.

[00452] An "individual member" of a library differs from other members of the library in the nucleotide sequence of the DNA targeting segment of the guide RNA. Thus, e.g., each individual member of a library can comprise the same or substantially the same nucleotide sequence of the protein-binding segment as all other members of the library; and can comprise the same or substantially the same nucleotide sequence of the transcriptional termination segment as all other members of the library; but differs from other members of the library in the nucleotide sequence of the DNA targeting segment of the guide RNA. In this way, the library can comprise members that bind to different target nucleic acids.

[00453] Uses

[00454] A method for modulating transcription according to the present disclosure finds use in a variety of applications, which are also provided. Applications include research applications; diagnostic applications; industrial applications; and treatment applications.

[00455] Research applications include, e.g., determining the effect of reducing or increasing transcription of a target nucleic acid on, e.g., development, metabolism, expression of a downstream gene, and the like.

[00456] High through-put genomic analysis can be carried out using a transcription modulation method, in which only the DNA-targeting segment of the guide RNA needs to be varied, while the protein-binding segment and the transcription termination segment can (in some cases) be held constant. A library (e.g., a library) comprising a plurality of nucleic acids used in the genomic analysis would include: a promoter operably linked to a guide RNA-encoding nucleotide sequence, where each nucleic acid would include a different DNA-targeting segment, a common protein-binding segment, and a common transcription termination segment. A chip could contain over 5×10^4 unique guide RNAs. Applications would include large-scale phenotyping, gene-to-function mapping, and meta-genomic analysis.

[00457] The methods disclosed herein find use in the field of metabolic engineering. Because transcription levels can be efficiently and predictably controlled by designing an appropriate guide RNA, as disclosed herein, the activity of metabolic pathways (e.g., biosynthetic pathways) can be precisely controlled and tuned by controlling the level of specific enzymes (e.g., via increased or decreased transcription) within a metabolic pathway of interest. Metabolic pathways of interest include those used for chemical (fine chemicals, fuel, antibiotics, toxins, agonists, antagonists, etc.) and/or drug production.

[00458] Biosynthetic pathways of interest include but are not limited to (1) the mevalonate pathway (e.g., HMG-CoA reductase pathway) (converts acetyl-CoA to dimethylallyl pyrophosphate (DMAPP) and isopentenyl pyrophosphate (IPP), which are used for the biosynthesis of a wide variety of biomolecules including terpenoids/isoprenoids), (2) the non-mevalonate pathway (i.e., the "2-C-methyl-D-erythritol 4-phosphate/1-deoxy-D-xylulose 5-phosphate pathway" or "MEP/DOXP pathway" or "DXP pathway")(also produces DMAPP and IPP, instead by converting pyruvate and glyceraldehyde 3-phosphate into DMAPP and IPP via an alternative pathway to the mevalonate pathway), (3) the polyketide synthesis pathway (produces a variety of polyketides via a variety of polyketide synthase enzymes. Polyketides include naturally occurring small molecules used for chemotherapy (e. g., tetracyclin, and macrolides) and industrially important polyketides include rapamycin (immunosuppressant), erythromycin (antibiotic), lovastatin (anticholesterol drug), and epothilone B (anticancer drug)), (4) fatty acid synthesis pathways, (5) the DAHP (3-deoxy-D-arabino-heptulosonate 7-phosphate) synthesis pathway, (6) pathways that produce potential biofuels (such as short-chain alcohols and alkane, fatty acid methyl esters and fatty alcohols, isoprenoids, etc.), etc.

[00459] Networks and Cascades

[00460] The methods disclosed herein can be used to design integrated networks (i.e., a cascade or cascades) of control. For example, a guide RNA / variant Cas9 site-directed polypeptide may be used to control (i.e., modulate, e.g., increase, decrease) the expression of another DNA-tageting RNA or another variant Cas9 site-directed polypeptide. For example, a first guide RNA may be designed to target the modulation of transcription of a second chimeric dCas9 polypeptide with a function that is different than the first variant Cas9 site-directed polypeptide (e.g., methyltransferase activity, demethylase activity, acetyltansferase activity, deacetylase activity, etc.). In addition, because different dCas9 proteins (e.g., derived from different species) may require a different Cas9 handle (i.e., protein binding segment), the second chimeric dCas9 polypeptide can be derived from a different species than the first dCas9 polypeptide above. Thus, in some cases, the second chimeric dCas9 polypeptide can be selected such that it may not interact with the first guide RNA. In other cases, the second chimeric dCas9 polypeptide can be selected such that it does interact with the first guide RNA. In some such cases, the activities of the two (or more) dCas9 proteins may compete (e.g., if the polypeptides have opposing activities) or may synergize (e.g., if the polypeptides have similar or synergistic activities). Likewise, as noted above, any of the complexes (i.e., guide RNA / dCas9 polypeptide) in the network can be designed to control other guide RNAs or dCas9 polypeptides. Because a guide RNA and variant Cas9 site-directed polypeptide can be targeted to any desired DNA sequence, the methods described herein can be used to control and regulate the expression of any desired target. The integrated networks (i.e., cascades of interactions) that can be designed range from very simple to very complex, and are without limit.

[00461] In a network wherein two or more components (e.g., guide RNAs, activator-RNAs, targeter-RNAs, or dCas9 polypeptides) are each under regulatory control of another guide RNA/dCas9 polypeptide complex, the level of expression of one component of the network may affect the level of expression (e.g., may increase or decrease the expression) of another component of the network. Through this mechanism, the expression of one component may affect the expression of a different component in the same network, and the network may include a mix of components that increase the expression of other components, as well as components that decrease the expression of other components. As would be readily understood by one of skill in the art, the above examples whereby the level of expression of one component may affect the level of expression of one or more different component(s) are for illustrative purposes, and are not limiting. An additional layer of complexity may be optionally introduced into a network when one or more components are modified (as described above) to be manipulable (i.e., under experimental control, e.g., temperature control; drug control, i.e., drug inducible control; light control; etc.).

[00462] As one non-limiting example, a first guide RNA can bind to the promoter of a second guide RNA, which controls the expression of a target therapeutic/metabolic gene. In such a case, conditional expression of the first guide RNA indirectly activates the therapeutic/metabolic gene. RNA cascades of

this type are useful, for example, for easily converting a repressor into an activator, and can be used to control the logics or dynamics of expression of a target gene.

[00463] A transcription modulation method can also be used for drug discovery and target validation.

Examples

[00464] Various aspects of the invention make use of the following materials and methods and are illustrated by the following non-limiting examples, wherein Example 1 relates to Cas9 orthologs, Example 2 elates to exchangeability of bacterial RNase III enzymes, Example 3 relates to the Cas9 HNH and RuvC domains, Example 4 relates to exchangeability of Cas9 endonucleases in tracrRNA-directed pre-crRNA maturation by RNase III, Example 5 relates to PAMs of Cas9 orthologs and Example 6 relates to exchangeability of guide RNA and Cas9 endonucleases.

Materials and Methods

[00465] Bacterial strains and culture conditions

[00466] Supplementary Table S1 lists bacterial strains used in this study. *S. pyogenes*, *Streptococcus mutans*, *Campylobacter jejuni*, *N. meningitidis*, *Escherichia coli* and *Francisella novicida* were grown as previously described (15,16). BHI (Brain Heart Infusion, Becton Dickinson) agar and BHI broth medium supplemented with 1% glucose and 1% lactose were used to culture *S. thermophilus* at 42°C in a 5% CO_2 environment (16). *Pasteurella multocida* and *Staphylococcus aureus* were grown at 37°C on BHI agar plates and in BHI broth with shaking. Cell growth was monitored by measuring the optical density of cultures at 620 nm (OD_{620}) using a microplate reader (BioTek PowerWave).

[00467] Bacterial transformation

[00468] *E. coli* was transformed with plasmid DNA according to standard protocols (35). Transformation of *S. pyogenes* was performed as previously described (36) with some modifications. *S. pyogenes* pre-cultures were diluted 1:100 in fresh THY medium and grown at 37°C, 5% CO_2 until OD_{620} reached 0.3. Glycine was added to the medium to 10% final concentration and growth was maintained for an additional hour. Cells were spun down at 4°C at 2500 x g and washed three times with electroporation buffer (5 mM KH_2PO_4, 0.4 M D-sorbitol, 10% glycerol, pH 4.5), finally suspended in the same buffer and equalized to the same OD_{620}. For electroporation, 1 µg of plasmid was incubated with the competent cells on ice for 10 min. The conditions were 25 µF, 600 Ω and 1.5 V using 1 mm electroporation cuvettes (Biorad). After a regeneration time of 3 h, bacteria were spread on agar medium supplemented with kanamycin (300 µg/ml). Transformation assays were performed at least three times independently with technical triplicates. The efficiencies were calculated as CFU (colony-forming units) per µg of plasmid

DNA. Positive and negative control transformations were done with backbone plasmid pEC85 and sterile H_2O, respectively.

[00469] DNA manipulations

[00470] DNA manipulations including DNA preparation (QIAprep Spin MiniPrep Kit, Qiagen), PCR (*Phusion*® High-Fidelity DNA *Polymerase,* Finnzyme), DNA digestion (restriction enzymes, Fermentas), DNA ligation (T4 DNA ligase, Fermentas), DNA purification (QIAquick PCR Purification Kit, Qiagen) and agarose gel electrophoresis were performed according to standard techniques or manufacturers' protocols with some modifications (35). Site-directed mutagenesis was done using QuikChange II XL kit (Stratagene) or PCR-based mutagenesis (37). Synthetic oligonucleotides (Sigma-Aldrich & Biomers) and plasmids used and generated in this study are listed in Supplementary Table S1. The integrity of all constructed plasmids was verified by enzymatic digestion and sequencing at LGC Genomics.

[00471] Construction of plasmids for complementation studies in *S. pyogenes*

[00472] The backbone shuttle vector pEC85 was used for complementation study (38,39). The RNase-III encoding genes (*rnc* genes) of *S. pyogenes*, *S. mutans*, *S. thermophilus*, *C. jejuni*, *N. meningitidis*, *P. multocida*, *F. novicida*, *E. coli* and *S. aureus*, and the genes encoding truncated and inactive RNase III variants (truncated and inactive (D51A) *rnc* mutants) of *S. pyogenes* were cloned in pEC483 (pEC85 containing the native promoter of *S. pyogenes rnc*) using NcoI and EcoRI restriction sites (Supplementary Table S1, Supplementary Figure S6). The ortholog and mutant *cas9* genes were cloned in pEC342 (pEC85 containing a sequence encoding tracrRNA-171 nt (16) and the native promoter of the *S. pyogenes cas* operon) using SalI and SmaI restriction sites (Supplementary Table S1). Note that in a previous study, we observed low abundance of tracrRNA in the *cas9* deletion mutant. For this reason, plasmids used in cas9 complementation studies were designed to encode tracrRNA in addition to *cas9* (16). The generated *rnc* and *cas9* recombinant plasmids were introduced in *S. pyogenes* Δ*rnc* and Δ*cas9* deletion strains, respectively (Supplementary Table S1). Plasmid integrity in all complemented strains was checked by plasmid DNA extraction and digestion.

[00473] Construction of plasmids for transformation studies in *S. pyogenes*

[00474] Plasmid pEC85 was used as backbone vector for transformation studies. A DNA fragment containing WT *speM* protospacer sequence was cloned in the PstI site of plasmids containing coding sequences of WT or mutated *cas9* from *S. pyogenes* (Supplementary Table S1).

[00475] Construction of plasmids for protein purification

[00476] The overexpression vector pET16b (Novagen) was modified by inserting three additional restriction sites (SalI, SacI, NotI) into the NdeI restriction site, generating pEC621. The genes coding for the orthologous Cas9 proteins were PCR amplified from genomic DNA of the corresponding strains using

primers containing a SalI and a NotI restriction site (Supplementary Table S1). The *S. pyogenes cas9* mutant genes were PCR amplified from the complementation plasmids mentioned above. All orthologous and mutant *cas9* genes were cloned into the SalI and NotI sites of pEC621.

[00477] Construction of substrate plasmids for *in vitro* cleavage assays

[00478] Plasmid pEC287 that contains the *speM* protospacer sequence was used as a vector to construct all substrate plasmids. The PAM sequence located in 3' just next to the crRNA-targeted sequence of the *speM* protospacer (GGG on this plasmid) was modified by PCR-mediated site-directed mutagenesis (37) using one standard oligonucleotide (OLEC 3140 or OLEC3194) that either introduced or removed a XbaI restriction site for screening purposes, and a second mutagenic oligonucleotide to exchange the protospacer adjacent sequence (Supplementary Table S1).

[00479] RNA preparation

[00480] Total RNA from *S. pyogenes* SF370 WT, deletion mutants and complemented strains was prepared from culture samples collected at the mid-logarithmic phase of growth using TRIzol (Sigma-Aldrich). The total RNA samples were treated with DNase I (Fermentas) according to the manufacturer's instructions. The concentration of RNA in each sample was measured using NanoDrop.

[00481] Northern blot analysis

[00482] Northern blot analysis was carried out essentially as described previously (40-42). Total RNA was separated on 10% polyacrylamide 8 M urea gels and further processed for blotting on nylon membranes (Hybond™ N+, GE healthcare; Trans-Blot® SD semi-dry transfer apparatus, Biorad; 1X TBE, 2 h at 10 V/cm), chemical crosslinking with EDC (1-Ethyl-3-(3-dimethylaminopropyl) carbodiimide hydrochloride) (41) and prehybridization (Rapid-hyb buffer, GE healthcare; 1 h at 42°C). Oligonucleotide probes (40 pmol) were labeled with ^{32}P (20 µCi) using the T4-polynucleotide kinase (10 U, Fermentas) and purified using G-25 columns (GE Healthcare) prior use. Visualization of the radioactive signal was done using a phosphorimager. 5S rRNA served as loading control.

[00483] Protein purification

[00484] *E. coli* Rosetta2(DE3) and *E. coli* NiCo21(DE3) (New England Biolabs) were transformed with overexpression plasmids coding for *S. pyogenes* WT and mutant or orthologous Cas9, respectively. Cells were grown at 37°C to reach an OD_{600} of 0.7 - 0.8, protein expression was induced by adding IPTG to a final concentration of 0.5 mM and cultures were further grown at 13°C overnight. The cells were harvested by centrifugation and the pellet was resuspended in lysis-buffer (20 mM HEPES pH 7.5, 500 mM KCl [1 M for *S. thermophilus** Cas9], 0.1% Triton X-100, 25 mM imidazole) and lysed by sonication. The lysate was cleared by centrifugation (> 20 000 x g) and incubated with Ni-NTA (Qiagen) for 1 h at 4°C. After washing the Ni-NTA with lysis-buffer and wash-buffer (20 mM HEPES pH 7.5, 300 mM KCl,

0.1% Triton X-100, 25 mM imidazole), the recombinant protein was eluted with elution-buffer (20 mM HEPES pH 7.5, 150 mM KCl, 0.1 mM DTT, 250 mM imidazole, 1 mM EDTA) and the fractions were analyzed by SDS-PAGE. In the case of *S. pyogenes* Cas9 WT and mutants, the protein containing eluates were pooled and further purified via HiTrap SP FF (GE Healthcare) cation-exchange chromatography. Briefly, the protein was loaded on the column equilibrated with buffer A (20 mM HEPES pH 7.5, 100 mM KCl) using an FPLC system (Äkta, GE Healthcare). Cas9 was eluted with a gradient of buffer B (20 mM HEPES pH 7.5, 1 M KCl) over 12 ml. 1 ml fractions were collected and analyzed by SDS-PAGE. The protein containing fractions were pooled and dialyzed overnight (20 mM HEPES pH 7.5, 150 mM KCl, 50% glycerol). For Cas9 orthologs, the eluates from Ni-NTA purification were checked for purity by SDS-PAGE. In case of contaminants, a second purification over chitin beads was performed as described in the manual for NiCo21(DE3) cells from New England Biolabs. Briefly, 1 ml chitin beads (New England Biolabs) equilibrated with buffer A was incubated with the Ni^{2+}-IMAC eluates for 1 h at 4°C. Afterwards the beads were added onto a column and the Cas9 containing flowthroughs were collected and again checked for purity by SDS-PAGE (Supplementary Figure S1). The purified proteins were dialyzed overnight. The protein concentration was calculated by measuring the OD_{280} using the extinction coefficient. The detailed characteristics of purified proteins are summarized in Supplementary Figure S1A.

[00485] *In vitro* transcription

[00486] RNA for *in vitro* DNA cleavage assays was generated by *in vitro* transcription using the AmpliScribe™ T7-*Flash*™ Transcription Kit (Epicentre) according to the manufacturer's instructions. PCR products or synthetic oligonucleotides used as templates are listed in Supplementary Table S1. The synthesized tracrRNA and repeat region of crRNA from each bacterial species correspond to the mature forms of RNAs as determined by deep RNA sequencing (15) or bioinformatics predictions. The spacer region of all crRNAs used in this study targets the *speM* protospacer (encoding superantigen; targeted by spacer 2 of *S. pyogenes* SF370 CRISPR array, Spyo1h_002 (16)). Transcribed RNAs were precipitated and further purified from 10% polyacrylamide 8 M urea denaturing gel. The RNA concentration was determined by measuring the OD_{260} and the molarity was calculated. Equimolar amounts of crRNA and tracrRNA were mixed in 5X RNA annealing buffer (1 M NaCl, 100 mM HEPES pH 7.5), heated up to 95°C for 5 min and slowly cooled to room temperature before use.

[00487] *In vitro* DNA cleavage assays

[00488] For the cleavage assays using Cas9 mutant proteins, 25 nM of Cas9 were incubated with equimolar amounts of prehybridized *S. pyogenes* dual-RNA in cleavage buffer (20 mM HEPES pH 7.5, 150 mM KCl, 10 mM $MgCl_2$, 0.5 mM DTT, 0.1 mM EDTA) for 15 min at 37°C. Plasmid DNA (5 nM) containing *speM* (NGG PAM) was added and further incubated for 1 h at 37°C. The reaction was stopped by addition of 5X loading buffer (250 mM EDTA, 30% glycerol, 1.2% SDS, 0.1% (w/v) bromophenol blue)

and analyzed by 1% agarose gel electrophoresis in 1X TAE. Cleavage products were visualized by ethidium bromide staining. All other cleavage assays were carried out using the same conditions with the following modifications: KGB (43) (100 mM potassium glutamate, 25 mM Tris/acetate pH 7.5, 10 mM Mg-acetate, 0.5 mM 2-mercaptoethanol, 10 µg/ml BSA) was used as cleavage buffer and different concentrations of dual-RNA:Cas9 complex were analyzed. The concentration of plasmid DNA was kept constant in all experiments, i.e. 5 nM.

[00489] Search for PAM motifs

[00490] Spacer sequences of the selected bacterial species were extracted from the CRISPRdatabase (http://crispr.u-psud.fr/crispr/) and used to find cognate protospacer candidates using megaBLAST (http://blast.ncbi.nih.gov/Blast). Protospacer candidates were defined as containing a sequence with ≥90% similarity to the crRNA spacer sequence and originating from phage, plasmid or genomic DNA related to the bacterial species of the targeting CRISPR-Cas. For the investigated CRISPR-Cas loci, the orientation of transcription was determined previously by RNA sequencing or Northern blot analysis (15,16). It was also shown before that in type II CRISPR-Cas, the PAM sequence is located in 3' of the protospacer, juxtaposed to the sequence targeted by cognate crRNA on the non-target strand (14,18,23,44). To identify possible PAMs in each bacterial species, 10 nt sequences on the non-target strand directly downstream of each protospacer sequence were aligned. A logo plot (http://weblogo.berkeley.edu/) showing the most abundant nucleotides was created and PAM sequences were predicted. In the cases of CRISPR-Cas loci for which no suitable protospacer sequences could be identified (*S. mutans* UA159, *C. jejuni* NCTC 11168, *P. multocida* Pm70, *F. novicida* U112), closely related strains of the same species were selected (Supplementary Table S2). The spacer contents of the type II CRISPR arrays in selected strains were analyzed (http://crispr.u-psud.fr/Server/). The spacer sequences were then used to select cognate protospacer sequences as described above.

[00491] Protein sequence analysis

[00492] Position-Specific Iterated (PSI)-BLAST program (45) was used to retrieve orthologs of the Cas9 family in the NCBI nr database. Sequences shorter than 800 amino acids were discarded. The BLASTClust program (46) set up with a length coverage cutoff of 0.8 and a score coverage threshold (bit score divided by alignment length) of 0.8 was used to cluster the remaining sequences (Supplementary Table S2). This procedure produced 82 clusters. In the case of sequences reported in this study, one or several representatives from each cluster were selected and aligned using the MUSCLE program (47) with default parameters, followed by a manual correction on the basis of local alignments obtained using PSI-BLAST (45) and HHpred programs (48). The confidently aligned blocks (Supplementary Figure S2) with 285 informative positions were used for maximum likelihood tree reconstruction using the FastTree program (49) with the default parameters: JTT evolutionary model, discrete gamma model with 20 rate categories. The same program was used to calculate the bootstrap values. Cas1 sequences were

selected from the corresponding *cas* operons (Supplementary Table S2). A few incomplete sequences were substituted by other Cas1 sequences from the same Cas9 cluster (Supplementary Table S2). Several Cas1 proteins from subtypes I-A, B, C and E were included as an outgroup. Cas1 sequences were aligned using the same approach described above and 252 informative positions (Supplementary Figure S3) were used for maximum likelihood tree reconstruction using the FastTree program. RNase III multiple sequence alignment was prepared using the MUSCLE program.

[00493] RNA sequence and structure analysis

[00494] RNA duplex secondary structures were predicted using RNAcofold of the Vienna RNA package (50,51) and RNAhybrid (http://bibiserv.techfak.uni-bielefeld.de/rnahybrid/). The structure predictions were then visualized using VARNA (52).

Supplementary Table S1. Strains, plasmids and primers used in the study.

Strain	Relevant characteristics	Source
Streptococcus pyogenes		
WT		
EC904	SF370 (M1 serotype) (WT)	ATCC 700294
Δ*cas9*		
EC1788	EC904Δ*cas9*	(16)
Δ*rnc*		
EC1636	EC904Δ*rnc*	(16)
Δ*cas9* in SF370 + *cas9* complementations *in trans*		
EC2121	EC1788 + pEC714 (P*cas9*(Spy)-*cas9*(Spy)-CtHis)	This study
EC2127	EC1788 + pEC710 (171 tracrRNA-P*cas9*(Spy)-CtHis)	This study
EC2150	EC1788 + pEC553 (P*cas9*(Spy)-*cas9*-HH983AA(Spy)-CtHis)	This study
EC2151	EC1788 + pEC554 (P*cas9*(Spy)-*cas9*-D10A(Spy)-CtHis)	This study
EC2152	EC1788 + pEC555 (P*cas9*(Spy)-*cas9*-H840A(Spy)-CtHis)	This study
EC2153	EC1788 + pEC556 (P*cas9*(Spy)-*cas9*-N854A(Spy)-	This study

	CtHis	
EC2154	EC1788 + pEC557 (P*cas9*(Spy)-*cas9*-N863A(Spy)-CtHis)	This study
EC2155	EC1788 + pEC558 (P*cas9*(Spy)-*cas9*-D986A(Spy)-CtHis)	This study
EC2156	EC1788 + pEC559 (P*cas9*(Spy)-*cas9*-E762A(Spy)-CtHis)	This study
EC2118	EC1788 + pEC518 (P*cas9*(Spy)-*cas9*(Cje)-CtHis)	This study
EC2128	EC1788 + pEC538 (P*cas9*(Spy)-*cas9*(Fno)-CtHis)	This study
EC2199	EC1788 + pEC544 (P*cas9*(Spy)-*cas9*(Nme)-CtHis)	This study
EC2119	EC1788 + pEC520 (P*cas9*(Spy)-*cas9*(Pmu)-CtHis)	This study
EC2111	EC1788 + pEC519 (P*cas9*(Spy)-*cas9*(Smu)-CtHis)	This study
EC2112	EC1788 + pEC521 (P*cas9*(Spy)-*cas9*(Sth*)-CtHis)	This study
EC2120	EC1788 + pEC522 (P*cas9*(Spy)-*cas9*(Sth**)-CtHis)	This study
Δ*rnc* in SF370 + *rnc* complementations *in trans*		
EC2076	EC1636 + pEC484 (P*rnc*(Spy)-*rnc*(Spy))	This study
EC2084	EC1636 + pEC505 (P*rnc*(Spy)-*rnc*-catalytically inactive(Spy))	This study
EC2083	EC1636 + pEC504 (P*rnc*(Spy)-*rnc*-RNA binding inactive(Spy))	This study
EC2078	EC1636 + pEC486 (P*rnc*(Spy)-*rnc*(Cje))	This study
EC2080	EC1636 + pEC492 (P*rnc*(Spy)-*rnc*(Eco))	This study
EC2126	EC1636 + pEC537 (P*rnc*(Spy)-*rnc*(Fno))	This study
EC2085	EC1636 + pEC506 (P*rnc*(Spy)-*rnc*(Nme))	This study
EC2077	EC1636 + pEC485 (P*rnc*(Spy)-*rnc*(Pmu))	This study
EC2086	EC1636 + pEC507 (P*rnc*(Spy)-*rnc*(Sau))	This study
EC2082	EC1636 + pEC494 (P*rnc*(Spy)-*rnc*(Smu))	This study
EC2131	EC1636 + pEC534 (P*rnc*(Spy)-*rnc*(Sth))	This study
Campylobacter jejuni		
EC437	NCTC 11168; ATCC 700819 (WT), CIP 107370	Pasteur Institute

Francisella novicida		
EC1041	U112 (WT)	Anders Sjöstedt
Neisseria meningitidis		
EC438	CIP 107858	Pasteur Institute
Pasteurella multocida		
EC439	Pm70 (WT), ATCC BAA-1113	Pasteur Institute
Staphylococcus aureus		
EC36	COL (WT)	Lab strain collection
Streptococcus mutans		
EC1293	UA159 (WT)	(16)
Streptococcus thermophilus		
EC810	LMD-9 (WT)	(16)
E. coli		
RDN204	TOP10, host for cloning	Invitrogen
EC1265	Rosetta	Novagen

[a] Cje: *Campylobacter jejuni* NCTC 11168; Eco: *Escherichia coli* TOP10; Fno: *Francisella novicida* U112; Nme: *Neisseria meningitidis* A Z2491; Pmu: *Pasteurella multocida* Pm70; Sau: *Staphylococcus aureus* COL; Smu: *Streptococcus mutans* UA159; Spy: *Streptococcus pyogenes* SF370; Sth: *Streptococcus thermophilus* LMD-9.

Plasmid	**Relevant characteristics**	**Source**
Vectors for *S. pyogenes*		
pEC85	*repDEG*-pAMβ1, pJH1-*aphIII*, ColE1	Bernhard Roppenser
Plasmids for *cas9* domain functional and co-evolution analysis in *S. pyogenes* SF370		
pEC268	pEC85Ω171 tracrRNA (171 nt form)	(16)
pEC309	pEC85Ω P*cas9*(Spy)-*cas9*(Spy)	(16)
pEC368	pEC85Ω171 tracrRNA-P*cas9*(Spy)-*cas9*(Spy)	(16)
pEC710	pEC85Ω171 tracrRNA-P*cas9*(Spy)-CtHis	This study
pEC714	pEC710Ω*cas9*(Spy)	This study

pEC553	pEC710Ω*cas9*-HH983AA(Spy)-CtHis	This study
pEC615	pEC553Ω*speM*	This study
pEC554	pEC710Ω*cas9*-D10A(Spy)-CtHis	This study
pEC659	pEC554ΩspeM	This study
pEC555	pEC710Ω*cas9*-H840A(Spy)-CtHis	This study
pEC660	pEC555ΩspeM	This study
pEC556	pEC710Ω*cas9*-N854A(Spy)-CtHis	This study
pEC661	pEC556ΩspeM	This study
pEC557	pEC710Ω*cas9*-N863A(Spy)-CtHis	This study
pEC618	pEC557Ω*speM*	This study
pEC558	pEC710Ω*cas9*-D986A(Spy)-CtHis	This study
pEC662	pEC558ΩspeM	This study
pEC559	pEC710Ω*cas9*-E762A(Spy)-CtHis	This study
pEC619	pEC559ΩspeM	This study
pEC518	pEC710Ω*cas9*(Cje)-CtHis	This study
pEC538	pEC710Ω*cas9*(Fno)-CtHis	This study
pEC544	pEC710Ω*cas9*(Nme)-CtHis	This study
pEC520	pEC710Ωcas9(Pmu)-CtHis	This study
pEC519	pEC710Ω*cas9*(Smu)-CtHis	This study
pEC521	pEC710Ωcas9(Sth*)-CtHis	This study
pEC522	pEC710Ωcas9(Sth**)-CtHis	This study

Plasmids for *rnc* co-evolution analysis in *S. pyogenes* SF370

pEC483	pEC85ΩP*rnc*(Spy)	This study
pEC484	pEC85ΩP*rnc*(Spy)-*rnc*(Spy)	This study
pEC505	pEC85ΩP*rnc*(Spy)-*rnc*-catalytically inactive(Spy)	This study
pEC504	pEC85ΩP*rnc*(Spy)-*rnc*-RNA binding inactive(Spy)	This study
pEC486	pEC85ΩP*rnc*(Spy)-*rnc*(Cje)	This study

pEC492	pEC85ΩP*rnc*(Spy)-*rnc*(Eco)	This study
pEC537	pEC85ΩP*rnc*(Spy)-rnc(Fno)	This study
pEC506	pEC85ΩP*rnc*(Spy)-rnc(Nme)	This study
pEC485	pEC85ΩP*rnc*(Spy)-*rnc*(Pmu)	This study
pEC507	pEC85ΩP*rnc*(Spy)-*rnc*(Sau)	This study
pEC494	pEC85ΩP*rnc*(Spy)-*rnc*(Smu)	This study
pEC534	pEC85ΩP*rnc*(Spy)-*rnc*(Sth)	This study
Plasmids for protospacer study *in vitro*		
pEC287	pEC85ΩP*speM-speM* (10 bp downstream protospacer: GGGTATTGGG)	Lab plasmid collection
pEC691	pEC287 (10 bp downstream protospacer: TGGTATTGGG)	This study
pEC692	pEC287 (10 bp downstream protospacer: TGGTGTTGGG)	This study
pEC693	pEC287 (10 bp downstream protospacer: GGGTGATTGG)	This study
pEC694	pEC287 (10 bp downstream protospacer: GGAGAATGGG)	This study
pEC696	pEC287 (10 bp downstream protospacer: GGGTCATAGG)	This study
pEC697	pEC287 (10 bp downstream protospacer: AGAAACAGGG)	This study
pEC698	pEC287 (10 bp downstream protospacer: AGAACCAGGG)	This study
pEC701	pEC287 (10 bp downstream protospacer: GTTTGATTGG)	This study
pEC706	pEC287 (10 bp downstream protospacer: GGAAAATGGG)	This study
Plasmids for Cas9 overexpression		
pEC225	pET16b	Novagen
pEC621	pEC225 inserted with cassette harboring NotI,	This study

	SacI, SalI site	
pEC626	pEC621Ω*cas9*(Spy)	This study
pEC627	pEC621Ω*cas9*-D10A(Spy)	This study
pEC628	pEC621Ω*cas9*-E762A(Spy)	This study
pEC629	pEC621Ωcas9-H840A(Spy)	This study
pEC630	pEC621Ωcas9-N854A(Spy)	This study
pEC631	pEC621Ωcas9-HH983AA(Spy)	This study
pEC632	pEC621Ω*cas9*(Cje)	This study
pEC633	pEC621Ω*cas9*(Pmu)	This study
pEC634	pEC621Ω*cas9*(Nme)	This study
pEC635	pEC621Ω*cas9*(Smu)	This study
pEC638	pEC621Ω*cas9*-N863A(Spy)	This study
pEC639	pEC621Ω*cas9*-D986A(Spy)	This study
pEC640	pEC621Ω*cas9*(Sth*)	This study
pEC641	pEC621Ω*cas9*(Sth**)	This study
pEC657	pEC621Ω*cas9*(Fno)	This study

Purpose	Primer	Sequence 5'-3'[a]	F/R[b]	Usage[c]
tracrRNA expression in *S. pyogenes* SF370				
tracrRNA	OLEC1014	*GGACTAGCCTTATTTTAACTTG*	R	NB (3' probe)
crRNA (CRISPR01 (type II-A) expression in *S. pyogenes* SF370				
crRNA	OLEC1049	*GGACCATTCAAAACAGCATAGCTCTAAAAC*	R	NB (repeat)
Loading controls for Northern blots				
5S rRNA	OLEC288	*CTAAGCGACTACCTTATCTCA*	R	NB
His-tagged *cas9* constructs (pEC85-based)				

pEC710	OLEC2151	GCAGGAATTCATCAGTGATGGTGATGGTGATGCCCGGGTTTGTCGACCT*CCTAAAATAAAAAGTTTAAATTAAATC*	F	Cloning
	OLEC2066	GGTGGTCTGCAG*GTTTGCAGTCAGAGTAGAATAGAAG*	R	
pEC714	OLEC2096	ATGCAGGTCGAC*ATGGATAAGAAATACTCAATAGGC*	F	Expression *cas9*(Spy)
	OLEC2097	ATGCAGCCCGGG*GTCACCTCCTAGCTGACTCAAATC*	R	
speM	OLEC2867	ATGCAGCCTGCAGG*GTGACAGAGAGAAACTTGATTCAAC*	F	Cloning of *speM* in other plasmids
	OLEC2868	ATGCAGCCTGCAGG*CTTCGTTTAAGTAAACATCAAAGTG*	R	
pEC518	OLEC2104	ATGCAGGTCGAC*GTGGCAAGAATTTTGGCATTTG*	F	Cloning *cas9*(Cje)
	OLEC2105	ATGCAGCCCGGG*TTTTTTAAAATCTTCTCTTTGTC*	R	
pEC538	OLEC2840	ATTAGTCGAC*ATGAATTTCAAAATATTGCCAATAG*	F	Cloning *cas9*(Fno)
	OLEC2841	ATTACCCGGG*ATTATTAGATGTTTCATTATAAATAC*	R	
pEC544	OLEC2092	ATGCAGGTCGAC*ATGGCTGCCTTCAAACCTAATCC*	F	Cloning *cas9*(Nme)
	OLEC2093	ATGCAGCCCGGG*ACGGACAGGCGGGCGTTTTTTCAG*	R	
pEC520	OLEC2100	ATGCAGGTCGAC*ATGCAAACAACAAATTTAAGTTA*	F	Cloning *cas9*(Pmu)
	OLEC2101	ATGCAGCCCGGG*ACGCACAGGTTGTCTTTGCTGAG*	R	
pEC519	OLEC2090	ATGCAGGTCGAC*ATGAAAAAACCTTACTCTATTGGAC*	F	Cloning *cas9*(Smu)
	OLEC2091	ATGCAGCCCGGG*GTCTCCTCCTAACTTATTGAGATC*	R	
pEC521	OLEC2098	ATGCAGGTCGAC*ATGACTAAGCCATACTCAATTGG*	F	Cloning *cas9*(Sth*)

	OLEC2099	ATGCAGCCCGGGACCCTCTCCTAGTTTGGCAAGGTC	R	
pEC522	OLEC2102	ATGCAGGTCGACATGAGTGACTTAGTTTTAGGACTTG	F	Cloning *cas9*(Sth**)
	OLEC2103	ATGCAGCCCGGGAAAATCTAGCTTAGGCTTATCACC	R	
pEC553	OLEC2229	GTACGTGAGATTAACAATTACGCTGCTGCCCATGATGCGTATCTA	F	Mutagenesis *cas9*-HH983AA(Spy)
	OLEC2230	TAGATACGCATCATGGGCAGCAGCGTAATTGTTAATCTCACGTAC	R	
pEC554	OLEC2128	GAAATACTCAATAGGCTTAGCTATCGGCACAAATAGCGTCG	F	Mutagenesis *cas9*-D10A(Spy)
	OLEC2129	CGACGCTATTTGTGCCGATAGCTAAGCCTATTGAGTATTTC	R	
pEC555	OLEC2223	TTTAAGTGATTATGATGTCGATGCCATTGTTCCACAAAGTTTCCT	F	Mutagenesis *cas9*-H840A(Spy)
	OLEC2224	AGGAAACTTTGTGGAACAATGGCATCGACATCATAATCACTTAAA	R	
pEC556	OLEC2225	CCTTAAAGACGATTCAATAGACGCTAAGGTCTTAACGCGTTCTGA	F	Mutagenesis *cas9*-N854A(Spy)
	OLEC2226	TCAGAACGCGTTAAGACCTTAGCGTCTATTGAATCGTCTTTAAGG	R	
pEC557	OLEC2227	GGTCTTAACGCGTTCTGATAAAGCTCGTGGTAAATCGGATAACGT	F	Mutagenesis *cas9*-N863A(Spy)
	OLEC2228	ACGTTATCCGATTTACCACGAGCTTTATCAGAACGCGTTAAGACC	R	
pEC558	OLEC2231	GTAACAATTACCATCATGCCCATGCTGCGTATCTAAATGCCGTCG	F	Mutagenesis *cas9*-D986A(Spy)
	OLEC2232	CGACGGCATTTAGATACGCAGCATGGGCATGATGGTAATTGTTAC	R	
pEC559	OLEC2221	CAGAAAATATCGTTATTGCAATGGCACGTGAAAATCAGACA	F	Mutagenesis *cas9*-E762A(Spy)
	OLEC2222	TGTCTGATTTTCACGTGCCATTGCAATAACGATATTTTCTG	R	

***rnc* constructs (pEC85-based)**

pEC483	OLEC2149	ATGCAGGCATGCCCTGTAGTTTTGGCTTGTCTGATC	F	Cloning in pEC85
	OLEC3274	ATGCAGAGCTCCATGG*AAAATCCCTTTCATATTTGTCAGTAGACC*	R	
pEC484	OLEC2109	ATGCAGCCATGG*AACAGCTTGAAGAGTTACTCTCAAC*	F	Cloning *rnc*(Spy), SEQ
	OLEC1668	CTTTTAAAAACATCTAAACCTCAC	R	
pEC504	OLEC2109	ATGCAGCCATGG*AACAGCTTGAAGAGTTACTCTCAAC*	F	Cloning of *rnc* RNA binding inactive(Spy)
	OLEC2656	ATGCAGGAATTC*CTACCCTTTTTCCACCTGAGGAATC*	R	
pEC505	OLEC2142	GAACGCTTGGAATTTTTAGGAGCCGCTGTTCTACAATTGATTATT	F	Mutagenesis of *rnc* catalytically inactive(Spy)
	OLEC2143	AATAATCAATTGTAGAACAGCGGCTCCTAAAAATTCCAAGCGTTC	R	
pEC486	OLEC2116	ATGCAGCCATGGAAAACATTGAAAAGCTAGAGCAGAG	F	Cloning *rnc*(Cje), SEQ
	OLEC2117	ATGCAGGAATTCCTATAAAGCTCCTAATTTCTCAAG	R	
pEC492	OLEC2124	ATGCAGCCATGGACCCCATCGTAATTAATCGGCTTC	F	Cloning *rnc*(Eco), SEQ
	OLEC2125	ATGCAGGAATTCTCATTCCAGCTCCAGTTTTTTCAACG	R	
pEC537	OLEC2842	ATTACCATGG*TTCCTGAATATTCACGATTTTATAAC*	F	Cloning *rnc*(Fno), SEQ
	OLEC2843	ATTGAATTC*CTATTTTTTTTCATGTAAGCCTTGTTGTG*	R	
pEC506	OLEC2118	ATGCAGCCATGGAAGACGATGTTTTGAAACAGCAGG	F	Cloning *rnc*(Nme), SEQ
	OLEC2119	ATGCAGGAATTCTCATTTCTTTTTCTTCTTCAGCGGC	R	
pEC485	OLEC211	ATGCAGCCATGGCTCAAAATTTAGAACGTTTACA	F	Cloning *rnc*(Pmu),

	4	ACG		SEQ
	OLEC2115	ATGCAGGAATTCTCATTTCATTTCCAATAATTGT	R	
pEC507	OLEC2126	ATGCAGCCATGGCTAAACAAAAGAAAAGTGAGATAG	F	Cloning *rnc*(Sau), SEQ
	OLEC2127	ATGCAGGAATTCCTATTTAATTTGTTTTAATTGCTTATAGG	R	
pEC494	OLEC2110	ATGCAGCCATGGAAACATTAGAAAAAAAACTGGCAG	F	Cloning *rnc*(Smu), SEQ
	OLEC2111	ATGCAGGAATTCTTAAGAACCTCGTTGAAGTTTTTC	R	
pEC534	OLEC2849	ATTACCATGGATCAACTTGAACAAAAACTTGAACAGGACTTTGG	F	Cloning *rnc*(Sth), SEQ
	OLEC2850	ATTAGAATTCTTAATTACCTAGTTGTTCAAGGGCAGACTTCGC	R	
Cas9 overexpression (pEC621 based)				
pEC621	OLEC2978	TAGCGGCCGCGAGCTCGTCGACGC	F	Cassette inserting NotI, SacI, SalI site in pEC225
	OLEC2979	TAGCGTCGACGAGCTCGCGGCCGC	R	
pEC626, 627, 628, 629, 630, 631, 638, 639	OLEC2097	ATGCAGGTCGAC*ATGGATAAGAAATACTCAATAGGC*	F	Cloning *cas9*(Spy and all mutants)
	OLEC2983	AGCTAGCGGCCGC*TCAGTCACCTCCTAGCTGACTCAAATC*	R	
pEC632	OLEC2104	ATGCAGGTCGAC*GTGGCAAGAATTTTGGCATTTG*	F	Cloning *cas9*(Cje)
	OLEC2986	ATGCAGCGGCCGC*TCATTTTTTAAAATCTTCTCTTTGTC*	R	
pEC633	OLEC2100	ATGCAGGTCGAC*ATGCAAACAACAAATTTAAGTTA*	F	Cloning *cas9*(Pmu)
	OLEC2173	ATGACGCGGCCGC*TTAACGCACAGGTTGTCTTTGCTG*	R	
pEC634	OLEC209	ATGCAGGTCGAC*ATGGCTGCCTTCAAACCTAATC*	F	Cloning *cas9*(Nme)

	2	*C*		
	OLEC2982	ATGACGCGGCCGC*TTAACGGACAGGCGGGCGTTTTTTCAG*	R	
pEC635	OLEC2090	ATGCAGGTCGAC*ATGAAAAAACCTTACTCTATTGGAC*	F	Cloning *cas9*(Smu)
	OLEC2981	ATGACGCGGCCGC*TTAGTCTCCTCCTAACTTATTGAG*	R	
pEC640	OLEC2098	ATGCAGGTCGAC*ATGACTAAGCCATACTCAATTGG*	F	Cloning *cas9*(Sth*)
	OLEC2984	ATGACGCGGCCGC*TTAACCCTCTCCTAGTTTGGCAAG*	R	
pEC641	OLEC2102	ATGCAGGTCGAC*ATGAGTGACTTAGTTTTAGGACTTG*	F	Cloning *cas9*(Sth**)
	OLEC2985	ATGACGCGGCCGC*TTAAAAATCTAGCTTAGGCTTATCAC*	R	
pEC657	OLEC2840	ATTAGTCGAC*ATGAATTTCAAAATATTGCCAATAG*	F	Cloning *cas9*(Fno)
	OLEC2987	ATGCAGCGGCCGC*CTAATTATTAGATGTTTCATTATAAATAC*	R	
Mutagenesis 10 bp downstream of *speM* protospacer				
pEC691	OLEC3140	*CAACCACTAATTTCTAGAAAAATCTTCG*	R	Mutagenesis on pEC287
	OLEC3141	*CAATTTGTAAAAAA*TGGTATTGGGGAATTC	F	
pEC692	OLEC3140	*CAACCACTAATTTCTAGAAAAATCTTCG*	R	Mutagenesis on pEC287
	OLEC3142	*CAATTTGTAAAAAA*TGGTGTTGGGGAATTC	F	
pEC693	OLEC3140	*CAACCACTAATTTCTAGAAAAATCTTCG*	R	Mutagenesis on pEC287
	OLEC3144	*CAATTTGTAAAAAAGGGTGATTGGGAATTC*	F	
pEC694	OLEC3140	*CAACCACTAATTTCTAGAAAAATCTTCG*	R	Mutagenesis on pEC287

	OLEC3143	*CAATTTGTAAAAAAGGAGAATGGGGAATTC*	F	
pEC696	OLEC3194	*CAACCACTAATTTTTAGAAAAATCTTCG*	R	Mutagenesis on pEC693
	OLEC3197	*CAATTTGTAAAAAAGGGTCATAGGGAATTC*	F	
pEC697	OLEC3194	*CAACCACTAATTTTTAGAAAAATCTTCG*	R	Mutagenesis on pEC694
	OLEC3198	*CAATTTGTAAAAAAAGAAACAGGGGAATTC*	F	
pEC698	OLEC3194	*CAACCACTAATTTTTAGAAAAATCTTCG*	R	Mutagenesis on pEC694
	OLEC3199	*CAATTTGTAAAAAAAGAACCAGGGGAATTC*	F	
pEC701	OLEC3194	*CAACCACTAATTTTTAGAAAAATCTTCG*	R	Mutagenesis on pEC693
	OLEC3204	*CAATTTGTAAAAAAGTTTGATTGGGAATTC*	F	
pEC706	OLEC3194	*CAACCACTAATTTTTAGAAAAATCTTCG*	R	Mutagenesis on pEC694
	OLEC3208	*CAATTTGTAAAAAAGGAAAATGGGGAATTC*	F	
***In vitro* tracrRNA and crRNA of *Streptococcus pyogenes* SF370 (*speM* spacer underlined)**				
T7-tracrRNA	OLEC1521	GAAATT**AATACGACTCACTATAG***AAAACAGCATAGCAAGTTAAAATAA*	F	T7-tracrRNA 5'
	OLEC1522	*AAAAAAAGCACCGACTCGGTGCCAC*	R	T7-tracrRNA 3'
T7-crRNA (template)	OLEC2177	GAAATT**AATACGACTCACTA**TA**GG**ATAACTCAATTTGTAAAAAAGTTTTAGAGCTATGCTGTTTTG	F	crRNA speM 5'
	OLEC2179	CAAAACAGCATAGCTCTAAAACTTTTTTACAAATTGAGTTAT**CCTATAGTGAGTCGTATTA**ATTTC	R	crRNA speM 3'
***In vitro* tracrRNA and crRNA of *Neisseria meningitidis* A Z2491 (*speM* spacer underlined)**				
T7-tracrRNA	OLEC3083	**GAAA**TT**AATACGACTCACTA**TA**GG**GAGAGCGAAATGAGAACCGTTGCTACAATAAGGCGTCTGAAAAGATGTGCCGCAACGCTCTGCCCCTTAAAGCTTCT	F	T7-tracrRNA 5'

(template)		GCTTTAAGGGGCATCGTTTATT		
	OLEC3084	AATAAACGATGCCCCTTAAAGCAGAAGC**TTT**AAG GGGCAGAGCGTTGCGGCACATCTTTTCAGACGC CTTATTG**T**AGCAACGGTTCTCATTTCGCTC**T**CC**CT ATAGTGAGTCGTATTAATTTC**	R	T7-tracrRNA 3'
T7-crRNA (template)	OLEC2209	G**AAATTAATACGACTCACTATAG**ATGATAACTCA ATTTGTAAAAAAGTTGTAGCTCCCTTTCTCATTT	F	crRNA speM 5'
	OLEC2214	AAATGAGAAAGGGAGCTACAACTTTTTTACAAA**TT** GAGTTATCAT**CTATAGTGAGTCGTATTAATTTC**	R	crRNA speM 3'

***In vitro* tracrRNA and crRNA of *Streptococcus mutans* UA159 (*speM* spacer underlined)**

T7-tracrRNA	OLEC3098	**GAAATTAATACGACTCACTATAG***GAAACAACACA GCAAGTTAAAATAAG*	F	T7-tracrRNA 5'
	OLEC3099	*AAATAAAAAAGCACCGAATCGG*	R	T7-tracrRNA 3'
T7-crRNA (template)	OLEC3085	**GAAATTAATACGACTCACTATAG**GATAACTCAAT TTGTAAAAAAGTTTTAGAGCTGTGTTGT	F	crRNA speM 5'
	OLEC3086	ACAACACAGCTCTAAAACTTTTTTACAAATTGAGT TATC**CTATAGTGAGTCGTATTAATTTC**	R	crRNA speM 3'

***In vitro* tracrRNA and crRNA of *Campylobacter jejuni* NCTC 11168 (*speM* spacer underlined)**

T7-tracrRNA (template)	OLEC3128	**GAAATTAATACGACTCACTATAG**GAAGGGACTA AAATAAAGAGTTTGCGGGACTCTGCGGGGTTACA ATCCCCTAAAACCGC	F	T7-tracrRNA 5'
	OLEC3129	GCGGTTTTAGGGGATTGTAACCCCGCAGAGTCC CGCAAACTCTTTATTTTAGTCCCTTCC**TATAGTGA GTCGTATTAATTTC**	R	T7-tracrRNA 3'
T7-crRNA (template)	OLEC3087	**GAAATTAATACGACTCACTA**TAGGATAACTCAAT TTGTAAAAAAGTTTTAGTCCCT	F	crRNA speM 5'
	OLEC3088	AGGGACTAAAACTTTTTTACAAATTGAGTTATC**CT ATAGTGAGTCGTATTAATTTC**	R	crRNA speM 3'

***In vitro* tracrRNA and crRNAs of *Francisella novicida* U112 (*speM* spacer underlined)**

T7-tracrRNA	OLEC3102	**GAAATTAATACGACTCACTATAG***GGTACCAAATA ATTAATGCTCTG*	F	T7-tracrRNA 5'
	OLEC310	*GTTATTCAGACGTGTCAAACAG*	R	T7-tracrRNA 3'

	3			
T7-crRNA (template)	OLEC3089	**GAAATTAATACGACTCACTATAG**GATAACTCAATTTGTAAAAAAGTTTCAGTTGCTGAATTATTTGGTAAAC	F	crRNA speM 5'
	OLEC3090	GTTTACCAAATAATTCAGCAACTGAAACTTTTTTACAAATTGAGTTATC**CTATAGTGAGTCGTATTAAT TTC**	R	crRNA speM 3'

***In vitro* tracrRNA and crRNAs of *Streptococcus thermophilus** LMD-9 (*speM* spacer underlined)**

T7-tracrRNA	OLEC3104	**GAAATTAATACGACTCACTATAG***GAACAACACAGCGAGTTAAAATAAGG*	F	T7-tracrRNA 5'
	OLEC3105	*AAAAAAAACACCGAATCGGTG*	R	T7-tracrRNA 3'
T7-crRNA (template)	OLEC3085	**GAAATTAATACGACTCACTATAG**GATAACTCAATTTGTAAAAAAGTTTTAGAGCTGTGTTGT	F	crRNA speM 5'
	OLEC3086	ACAACACAGCTCTAAAACTTTTTTACAAATTGAGTTATC**CTATAGTGAGTCGTATTAATTTC**	R	crRNA speM 3'

In vitro* tracrRNA and crRNAs of *Streptococcus thermophilus LMD-9 (*speM* spacer underlined)**

T7-tracrRNA	OLEC3106	**GAAATTAATACGACTCACTATAG***GCTTAAATCTTGCAGAAGCTACAAAG*	F	T7-tracrRNA 5'
	OLEC3107	*AAATAACGAAAACACCCTGCC*	R	T7-tracrRNA 3'
T7-crRNA (template)	OLEC3091	**GAAATTAATACGACTCACTATAG**GATAACTCAATTTGTAAAAAAGTTTTTGTACTCTCAAGATTTA	F	crRNA speM 5'
	OLEC3092	TAAATCTTGAGAGTACAAAAACTTTTTTACAAATTGAGTTATC**CTATAGTGAGTCGTATTAATTTC**	R	crRNA speM 3'

***In vitro* tracrRNA and crRNAs of *Pasteurella multocida* Pm70 (*speM* spacer underlined)**

T7-tracrRNA	OLEC3108	**GAAATTAATACGACTCACTATAG***GCTGCGAAATGAGAGACGTTGCTAC*	F	T7-tracrRNA 5'
	OLEC3109	*AAAAACGATGCCCCTTGCAATTAAG*	R	T7-tracrRNA 3'
T7-crRNA (template)	OLEC3093	**GAAATTAATACGACTCACTATAG**GATAACTCAATTTGTAAAAAAGTTGTAGTTCCCTCTCTCATTTCGC	F	crRNA speM 5'
	OLEC3094	GCGAAATGAGAGAGGGAACTACAACTTTTTTACAAATTGAGTTATC**CTATAGTGAGTCGTATTAATTTC**	R	crRNA speM 3'

Primers for sequencing analysis

***cas9 Streptococcus mutans* UA159**

cas9(Smu)	OLEC2792	*ATGAAAAAACCTTACTCTATTGGA*	F	SEQ
	OLEC2793	*GATTTTAAAAAGCATTTTGAATTA*	F	SEQ
	OLEC2794	*TACTTGCCAAATCAAAAAGTTCTT*	F	SEQ
	OLEC2795	*ATTATGGGACATCAACCTGAAAAT*	F	SEQ
	OLEC2796	*TACCCACAATTGGAACCTGAATTT*	F	SEQ

***cas9 Neisseria meningitidis* A Z2491**

cas9(Nme)	OLEC2797	*ATGGCTGCCTTCAAACCTAATCCA*	F	SEQ
	OLEC2798	*GTTCAAAAAATGTTGGGGCATTGC*	F	SEQ
	OLEC2799	*ATCCATATTGAAACTGCAAGGGAA*	F	SEQ
	OLEC2800	*AACGCGTTTGACGGTAAAACCATA*	F	SEQ

***cas9 Streptococcus thermophilus** LMD-9*

cas9(Sth*)	OLEC2807	*ATGACTAAGCCATACTCAATTGGA*	F	SEQ
	OLEC2808	*GATTTTAGGAAATGTTTTAATTTA*	F	SEQ
	OLEC2809	*TATTTGCCAGAAGAGAAGGTACTT*	F	SEQ
	OLEC2810	*GTAATGGGAGGAAGAAAACCCGAG*	F	SEQ
	OLEC281	*GCAAGTGCTTTACTTAAGAAATAC*	F	SEQ

	1			
	OLEC2812	*TTACTTTATCATGCTAAGAGAATA*	F	SEQ
cas9 Streptococcus thermophilus LMD-9**				
cas9(Sth**)	OLEC2817	*ATGAGTGACTTAGTTTTAGGACTT*	F	SEQ
	OLEC2818	*ATTTTTGGAATTCTAATTGGGAAA*	F	SEQ
	OLEC2819	*GGAGACTTTGACAATATTGTCATC*	F	SEQ
	OLEC2820	*TTGAATTTGTGGAAAAAACAAAAG*	F	SEQ
	OLEC2821	*CAGGAAAAATACAATGACATTAAG*	F	SEQ
***cas9 Pasteurella multocida* Pm70**				
cas9(Pmu)	OLEC2813	*ATGCAAACAACAAATTTAAGTTAT*	F	SEQ
	OLEC2814	*ACGCATGAAAAAAATGAGTTTAAA*	F	SEQ
	OLEC2815	*CTTGGGAAATCTTTTAAAGAACGT*	F	SEQ
	OLEC2816	*TATGAAATGGTGGATCAAGAAAGC*	F	SEQ
***cas9 Campylobacter jejuni* NCTC 11168**				
cas9(Cje)	OLEC2822	*GTGGCAAGAATTTTGGCATTTGAT*	F	SEQ
	OLEC2823	*GATGAAAAAAGAGCGCCAAAAAAT*	F	SEQ
	OLEC2824	*AACTACAAGGCCAAAAAAGACGCC*	F	SEQ
	OLEC2825	*AACAAAAGGAAGTTTTTTGAGCCT*	F	SEQ
***cas9 Francisella novicida* U112**				

cas9(Fno)	OLEC2869	*ATGAATTTCAAAATATTGCCAATA*	F	SEQ
	OLEC2870	*TTAGATACTCTTTTAACTGATGAT*	F	SEQ
	OLEC2871	*TTAAAAGTCTTAAAGTCAAGTAAA*	F	SEQ
	OLEC2872	*GGTTCAGAAGATAAAAAAGGTAAT*	F	SEQ
	OLEC2873	*AGAATTTTCTGCCTACGTGATCTT*	F	SEQ
	OLEC2874	*CCAATACTAATCCATAAAGAACTA*	F	SEQ
	OLEC2875	*ACATCAAAAAATATTTTTTGGCTG*	F	SEQ

[a] *italic*, sequence annealing to the template; <u>underlined</u>, restriction site; **bold**, T7 promoter.
[b] F, forward primer; R, reverse primer.
[c] NB, probe for Northern blot; SEQ, sequencing

Example 1

[00495] Diversity of Cas9 orthologs.

[00496] To investigate the evolution and diversity of dual-RNA:Cas9 systems, publicly available genomes were subjected to multiple rounds of BLAST search using previously retrieved Cas9 sequences as queries (15). Cas9 orthologs were identified in 653 bacterial strains representing 347 species (Supplementary Table S2). After removing incomplete or highly similar sequences, we selected 83 diverse, representative Cas9 orthologs for multiple sequence alignment and phylogenetic tree reconstruction (Figure 1A, Supplementary Table S2, Supplementary Figures S2 and S4, see Materials and Methods). The Cas9 tree topology largely agrees with the phylogeny of the corresponding Cas1 proteins (Supplementary Table S2, Supplementary Figures S3 and S4) and fully supports the previously described classification of type II CRISPR-Cas into three subtypes, II-A (specified by *csn2*), II-B (characterized by long and most diverged *cas9* variants (formerly *csx12*) and *cas4*), and II-C (three-*cas* gene operon) (15).

Supplementary Table S2. List of bacterial strains with identified Cas9 orthologs.

Cluster[a]	Strain[b]	Cas9 length (aa)	Cas9 GI	Cas1GI[c]	Subtype[d]
1	*Dolosigranulum pigrum* ATCC 51524	1332	375088882		Type II-A
	Enterococcus faecalis ATCC 29200	1337	229548613		
	Enterococcus faecalis ATCC 4200	1337	256617555		
	Enterococcus faecalis D6	1337	257086028		
	Enterococcus faecalis E1Sol	1337	257080914		
	Enterococcus faecalis OG1RF	1337	384512368		
	Enterococcus faecalis TX0470	1337	312900261		
	Enterococcus faecalis TX4244	1337	422695652		
	Enterococcus faecium 1,141,733	1339	257888853		
	Enterococcus faecium 1,231,408	1340	257893735		
	Enterococcus faecium E1133	1339	430847551		
	Enterococcus faecium E3083	1340	431757680		
	Enterococcus faecium PC4.1	1340	293379700		
	Enterococcus faecium TX1330	1340	227550972		
	Enterococcus faecium TX1337RF	1340	424765774		
	Enterococcus hirae ATCC 9790	1336	392988474		
	Enterococcus italicus DSM 15952	1330	315641599		
	Lactobacillus animalis KCTC 3501	1314	335357451		
	Listeria innocua ATCC 33091	1337	423101383		
	Listeria innocua Clip11262	1334	16801805		
	Listeria innocua FSL S4-378	1103	422414122		
	Listeria ivanovii FSL F6-596	953	315305353		
	Listeria monocytogenes 10403S	1334	386044902		
	Listeria monocytogenes FSL J1-175	1099	255520581		

Cluster[a]	Strain[b]	Cas9 length (aa)	Cas9 GI	Cas1GI[c]	Subtype[d]
	Listeria monocytogenes FSL J1-194	1334	254825045		
	Listeria monocytogenes FSL J1-208	1334	422810631		
	Listeria monocytogenes FSL N3-165	1334	254829042		
	Listeria monocytogenes FSL R2-503	1334	254854201		
	Listeria monocytogenes str. 1/2a F6854	1334	47097148		
	Streptococcus agalactiae 2603V/R	1370	22537057		
	Streptococcus agalactiae 515	1377	77413160		
	Streptococcus agalactiae A909	1370	76788458		
	Streptococcus agalactiae ATCC 13813	1378	339301617		
	Streptococcus agalactiae CJB111	1370	77411010		
	Streptococcus agalactiae COH1	1370	77407964		
	Streptococcus agalactiae FSL S3-026	1370	417005168		
	Streptococcus agalactiae GB00112	1370	421147428		
	Streptococcus agalactiae H36B	1370	77405721		
	Streptococcus agalactiae NEM316	1377	25010965		
	Streptococcus agalactiae SA20-06	1370	410594450		
	Streptococcus agalactiae STIR-CD-17	1370	421532069		
	Streptococcus anginosus F0211	1345	315223162		
	Streptococcus anginosus SK1138	1386	421490579		
	Streptococcus anginosus SK52 = DSM 20563	1396	335031483		
	Streptococcus bovis ATCC 700338	1373	306833855		
	Streptococcus canis FSL Z3-227	1375	392329410		
	Streptococcus constellatus subsp. *constellatus* SK53	1345	418965022		
	Streptococcus dysgalactiae subsp. *equisimilis* AC-2713	1371	410494913		
	Streptococcus dysgalactiae subsp. *equisimilis* ATCC 12394	1371	386317166		
	Streptococcus dysgalactiae subsp. *equisimilis* GGS_124	1371	251782637		
	Streptococcus dysgalactiae subsp. *equisimilis* RE378	1371	408401787		
	Streptococcus equi subsp. *zooepidemicus* MGCS10565	1348	195978435		
	Streptococcus equinus ATCC 9812	1377	320547102		
	Streptococcus gallolyticus subsp. *gallolyticus* ATCC BAA-2069	1370	325978669		
	Streptococcus gallolyticus subsp. *gallolyticus* TX20005	1370	306831733		
	Streptococcus gallolyticus UCN34	1371	288905639		
	Streptococcus infantarius subsp. *infantarius* CJ18	1375	379705580		
	Streptococcus iniae 9117	1368	406658208		
	Streptococcus macacae NCTC 11558	1338	357636406		
	Streptococcus mitis SK321	1392	307710946		
	Streptococcus mutans 11SSST2	1345	449165720		
	Streptococcus mutans 11SSST2	1345	449951835		
1	*Streptococcus mutans* 11VS1	1345	449976542		
(continued)	*Streptococcus mutans* 14D	1345	450149988		Type II-A
	Streptococcus mutans 15VF2	1355	449170557		
	Streptococcus mutans 15VF2	1355	449965974		
	Streptococcus mutans 1SM1	1345	449158457		

Cluster[a]	Strain[b]	Cas9 length (aa)	Cas9 GI	Cas1GI[c]	Subtype[d]
	Streptococcus mutans 1SM1	1345	449920643		
	Streptococcus mutans 24	1350	449247589		
	Streptococcus mutans 24	1350	450180942		
	Streptococcus mutans 2VS1	1345	449174812		
	Streptococcus mutans 2VS1	1345	449968746		
	Streptococcus mutans 3SN1	1345	449162653		
	Streptococcus mutans 3SN1	1345	449931425		
	Streptococcus mutans 4SM1	1345	449159838		
	Streptococcus mutans 4SM1	1345	449927152		
	Streptococcus mutans 4VF1	1345	449167132		
	Streptococcus mutans 4VF1	1345	449961027		
	Streptococcus mutans 5SM3	1345	449176693		
	Streptococcus mutans 5SM3	1345	449980571		
	Streptococcus mutans 66-2A	1359	449240165		
	Streptococcus mutans 66-2A	1359	450160342		
	Streptococcus mutans 8ID3	1345	449154769		
	Streptococcus mutans 8ID3	1345	449872064		
	Streptococcus mutans A19	1345	449187668		
	Streptococcus mutans A19	1345	450013175		
	Streptococcus mutans B	1345	450166294		
	Streptococcus mutans G123	1345	450029806		
	Streptococcus mutans GS-5	1345	397650022		
	Streptococcus mutans LJ23	1345	387785882		
	Streptococcus mutans M21	1345	449194333		
	Streptococcus mutans M21	1345	450036249		
	Streptococcus mutans M230	1345	449260994		
	Streptococcus mutans M230	1345	449903532		
	Streptococcus mutans M2A	1345	449209586		
	Streptococcus mutans M2A	1345	450074072		
	Streptococcus mutans N29	1345	449182997		
	Streptococcus mutans N29	1345	450003067		
	Streptococcus mutans N3209	1345	449210660		
	Streptococcus mutans N3209	1345	450077860		
	Streptococcus mutans N66	1345	449212466		
	Streptococcus mutans N66	1345	450083993		
	Streptococcus mutans NFSM1	1350	449202104		
	Streptococcus mutans NFSM1	1350	450051112		
	Streptococcus mutans NLML1	1345	450140393		
	Streptococcus mutans NLML4	1338	449202681		
	Streptococcus mutans NLML4	1338	450059882		
	Streptococcus mutans NLML9	1345	449209148		
	Streptococcus mutans NLML9	1345	450066176		
	Streptococcus mutans NMT4863	1355	449186850		
	Streptococcus mutans NMT4863	1355	450007078		
	Streptococcus mutans NN2025	1345	290580220		
	Streptococcus mutans NV1996	1345	450086338		
	Streptococcus mutans NVAB	1345	449181424		
	Streptococcus mutans NVAB	1345	449990810		
	Streptococcus mutans R221	1345	449258042		
	Streptococcus mutans R221	1345	449899675		
	Streptococcus mutans S1B	1345	449251227		
	Streptococcus mutans S1B	1345	449877120		

Cluster[a]	Strain[b]	Cas9 length (aa)	Cas9 GI	Cas1GI[c]	Subtype[d]
1 (continued)	*Streptococcus mutans* SF1	1345	450098705		Type II-A
	Streptococcus mutans SF14	1345	449221374		
	Streptococcus mutans SF14	1345	450107816		
	Streptococcus mutans SM1	1345	449245264		
	Streptococcus mutans SM1	1345	450176410		
	Streptococcus mutans SM4	1345	449246010		
	Streptococcus mutans SM4	1345	450170248		
	Streptococcus mutans SM6	1345	449223000		
	Streptococcus mutans SM6	1345	450112022		
	Streptococcus mutans ST6	1350	449227252		
	Streptococcus mutans ST6	1350	450123011		
	***Streptococcus mutans* UA159**	**1345**	**24379809**	**24379808**	
	Streptococcus mutans W6	1345	450094364		
	Streptococcus oralis SK304	1373	421488030		
	Streptococcus oralis SK610	1371	419782534		
	Streptococcus pseudoporcinus LQ 940-04	1374	416852857		
	***Streptococcus pyogenes* SF370 (M1 GAS)**	**1368**	**13622193**	**13622194**	
	Streptococcus pyogenes MGAS10270	1368	94543903		
	Streptococcus pyogenes MGAS10750	1371	94994317		
	Streptococcus pyogenes MGAS15252	1367	383479946		
	Streptococcus pyogenes MGAS2096	1368	94992340		
	Streptococcus pyogenes MGAS315	1368	21910213		
	Streptococcus pyogenes MGAS5005	1368	71910582		
	Streptococcus pyogenes MGAS6180	1368	71903413		
	Streptococcus pyogenes MGAS9429	1368	94988516		
	Streptococcus pyogenes NZ131	1368	209559356		
	Streptococcus pyogenes SSI-1	1368	28896088		
	Streptococcus ratti FA-1 = DSM 20564	1370	400290495		
	Streptococcus salivarius K12	1385	421452908		
	Streptococcus sanguinis SK115	1377	422848603		
	Streptococcus sanguinis SK330	1392	422860049		
	Streptococcus sanguinis SK353	1370	422821159		
	Streptococcus sp. C300	1377	322375978		
	Streptococcus sp. F0441	1371	414157437		
	Streptococcus sp. M334	1375	322378004		
	Streptococcus sp. oral taxon 56 str. F0418	1371	339640839		
	Streptococcus suis ST1	1381	389856936		
	Streptococcus thermophilus	1388	343794781		
	***Streptococcus thermophilus* LMD-9**	**1388**	**116628213**	**116628212**	
	Streptococcus thermophilus MN-ZLW-002	1388	387910220		
	Streptococcus thermophilus ND03	1388	386087120		
2	*Campylobacter coli* 1098	984	419564797		Type II-C
	Campylobacter coli 111-3	984	419536531		
	Campylobacter coli 132-6	987	419572019		
	Campylobacter coli 151-9	984	419603415		
	Campylobacter coli 1909	984	419576091		
	Campylobacter coli 1957	965	419581876		
	Campylobacter coli 2692	984	419553162		
	Campylobacter coli 59-2	984	419578074		
	Campylobacter coli 67-8	965	419587721		
	Campylobacter coli 80352	965	419558307		
	Campylobacter coli 80352	987	419559505		

Cluster[a]	Strain[b]	Cas9 length (aa)	Cas9 GI	Cas1GI[c]	Subtype[d]
	Campylobacter jejuni subsp. *doylei* 269.97	984	153952471		
	Campylobacter jejuni subsp. *jejuni* 110-21	987	419676124		
	Campylobacter jejuni subsp. *jejuni* 129-258	987	419619138		
	Campylobacter jejuni subsp. *jejuni* 1336	987	283956897		
	Campylobacter jejuni subsp. *jejuni* 140-16	984	419681578		
	Campylobacter jejuni subsp. *jejuni* 1577	984	419685099		
	Campylobacter jejuni subsp. *jejuni* 1854	987	419689467		
	Campylobacter jejuni subsp. *jejuni* 1997-10	984	419666522		
	Campylobacter jejuni subsp. *jejuni* 2008-1025	987	419650041		
	Campylobacter jejuni subsp. *jejuni* 2008-872	984	419654778		
	Campylobacter jejuni subsp. *jejuni* 2008-979	987	419660762		
	Campylobacter jejuni subsp. *jejuni* 2008-988	965	419656328		
	Campylobacter jejuni subsp. *jejuni* 2008-988	984	419655317		
	Campylobacter jejuni subsp. *jejuni* 260.94	961	86152042		
	Campylobacter jejuni subsp. *jejuni* 414	985	283953849		
	Campylobacter jejuni subsp. *jejuni* 51037	984	419674189		
	Campylobacter jejuni subsp. *jejuni* 51494	984	419619463		
	Campylobacter jejuni subsp. *jejuni* 53161	987	419647275		
	Campylobacter jejuni subsp. *jejuni* 60004	984	419629136		
	Campylobacter jejuni subsp. *jejuni* 81116	984	157415744		
	Campylobacter jejuni subsp. *jejuni* 84-25	984	88596565		
	Campylobacter jejuni subsp. *jejuni* 87459	984	419680124		
	Campylobacter jejuni subsp. *jejuni* ATCC 33560	984	419643715		
	Campylobacter jejuni subsp. *jejuni* CF93-6	987	86149266		
	Campylobacter jejuni subsp. *jejuni* CG8486	984	148925683		
	Campylobacter jejuni subsp. *jejuni* HB93-13	984	86152450		
2 (continued)	*Campylobacter jejuni* subsp. *jejuni* LMG 23210	987	419696801		**Type II-C**
	Campylobacter jejuni subsp. *jejuni* LMG 23211	984	419697443		
	Campylobacter jejuni subsp. *jejuni* LMG 23263	984	419628620		
	Campylobacter jejuni subsp. *jejuni* LMG 23264	984	419632476		
	Campylobacter jejuni subsp. *jejuni* LMG 23269	987	419634246		
	Campylobacter jejuni subsp. *jejuni* LMG 23357	987	419641132		
	***Campylobacter jejuni* subsp. *jejuni* NCTC 11168**	**984**	**218563121**	**218563120**	
	Campylobacter jejuni subsp. *jejuni* NW	983	424845990		
	Campylobacter jejuni subsp. *jejuni* PT14	987	407942868		
	Campylobacter lari	1003	345468028		
	Helicobacter canadensis MIT 98-5491	1007	253828136		
	Helicobacter cinaedi ATCC BAA-847	1023	396079277		
	Helicobacter cinaedi CCUG 18818	1023	313144862		
	Helicobacter cinaedi PAGU611	1023	386762035		
	Catellicoccus marimammalium M35/04/3	1140	424780480		
	Lactobacillus farciminis KCTC 3681	1126	336394701		
	Listeriaceae bacterium TTU M1-001	1087	381184145		
	Streptococcus anginosus 1_2_62CV	1125	319939170		
	Streptococcus gallolyticus UCN34	1130	288905632		
3	*Streptococcus gordonii* str. Challis substr. CH1	1136	157150687		**Type II-A**
	Streptococcus infantarius ATCC BAA-102	1129	171779984		
	Streptococcus macedonicus ACA-DC 198	1130	374338350		
	Streptococcus mitis ATCC 6249	1134	306829274		
	Streptococcus mutans NLML5	1128	449203378		
	Streptococcus mutans NLML5	1128	450064617		

Cluster[a]	Strain[b]	Cas9 length (aa)	Cas9 GI	Cas1GI[c]	Subtype[d]
3 (continued)	*Streptococcus mutans* NLML8	1125	449151037		Type II-A
	Streptococcus mutans NLML8	1125	450133520		
	Streptococcus mutans ST1	1134	449228751		
	Streptococcus mutans ST1	1134	450114718		
	Streptococcus mutans U2A	1125	449232458		
	Streptococcus mutans U2A	1125	450125471		
	Streptococcus oralis SK1074	1121	418974877		
	Streptococcus oralis SK313	1134	417940002		
	Streptococcus parasanguinis F0449	1140	419799964		
	Streptococcus pasteurianus ATCC 43144	1130	336064611		
	Streptococcus salivarius JIM8777	1127	387783792		
	Streptococcus salivarius PS4	1135	419707401		
	Streptococcus sp. BS35b	1026	401684660		
	Streptococcus sp. C150	1139	322372617		
	Streptococcus sp. GMD6S	1121	406576934		
	Streptococcus suis 89/1591	1122	223932525		
	Streptococcus suis D9	1122	386584496		
	Streptococcus suis ST3	1122	330833104		
	Streptococcus thermophilus CNRZ1066	1128	55822627		
	Streptococcus thermophilus JIM 8232	1121	386344353		
	***Streptococcus thermophilus* LMD-9**	**1121**	**116627542**	**116627543**	
	Streptococcus thermophilus LMG 18311	1122	55820735		
	Streptococcus thermophilus MN-ZLW-002	1121	387909441		
	Streptococcus thermophilus MTCC 5460	1122	445374534		
	Streptococcus thermophilus ND03	1121	386086348		
	Streptococcus vestibularis ATCC 49124	1128	322517104		
4	*Actinobacillus minor* NM305	1056	240949037		Type II-C
	Actinobacillus pleuropneumoniae serovar 10 str. D13039	1054	307256472		
	Actinobacillus succinogenes 130Z	1062	152978060		
	Actinobacillus suis H91-0380	1054	407692091		
	Haemophilus parainfluenzae ATCC 33392	1054	325578067		
	Haemophilus parainfluenzae CCUG 13788	1052	359298684		
	Haemophilus parainfluenzae T3T1	1052	345430422		
	Haemophilus sputorum HK 2154	1052	402304649		
	Kingella kingae PYKK081	1060	381401699		
	Neisseria bacilliformis ATCC BAA-1200	1077	329117879		
	Neisseria cinerea ATCC 14685	1082	261378287		
	Neisseria flavescens SK114	1081	241759613		
	Neisseria lactamica 020-06	1082	313669044		
	Neisseria meningitidis 053442	1082	161869390		
	Neisseria meningitidis 2007056	1082	433531983		
	Neisseria meningitidis 63049	1082	433514137		
	Neisseria meningitidis 8013	1082	385324780		
	Neisseria meningitidis 92045	1082	421559784		
	Neisseria meningitidis 93003	1081	421538794		
	Neisseria meningitidis 93004	1081	421541126		
	Neisseria meningitidis 96023	1082	433518260		
	Neisseria meningitidis 98008	1081	421555531		
	Neisseria meningitidis alpha14	1082	254804356		
	Neisseria meningitidis alpha275	1082	254672046		
	Neisseria meningitidis ATCC 13091	1082	304388355		

Cluster[a]	Strain[b]	Cas9 length (aa)	Cas9 GI	Cas1GI[c]	Subtype[d]
	Neisseria meningitidis N1568	1081	416164244		
	Neisseria meningitidis NM140	1081	421545139		
	Neisseria meningitidis NM220	1082	418291220		
	Neisseria meningitidis NM233	1082	418288950		
	Neisseria meningitidis WUE 2594	1082	385337435		
	***Neisseria meningitidis* Z2491**	**1082**	**218767588**	**218767587**	
	Neisseria sp. oral taxon 14 str. F0314	1089	298369677		
	Neisseria wadsworthii 9715	1097	350570326		
	Pasteurella multocida subsp. *gallicida* X73	1058	425063822		
	Pasteurella multocida subsp. *multocida* str. P52VAC	1056	421263876		
	***Pasteurella multocida* subsp. *multocida* str. Pm70**	**1056**	**15602992**	**15602991**	
	Simonsiella muelleri ATCC 29453	1063	404379108		
5	*Lactobacillus brevis* subsp. *gravesensis* ATCC 27305	1377	227509761		Type II-A
	Lactobacillus buchneri CD034	1371	406027703		
	Lactobacillus buchneri NRRL B-30929	1371	331702228		
	Lactobacillus casei BL23	1361	191639137		
	Lactobacillus casei Lc-10	1361	418010298		
	Lactobacillus casei M36	1363	417996992		
	Lactobacillus casei str. Zhang	1361	301067199		
	Lactobacillus casei T71499	1360	417999832		
	Lactobacillus casei UCD174	1366	418002962		
	Lactobacillus casei W56	1389	409997999		
	Lactobacillus coryniformis subsp. *coryniformis* KCTC 3167	1354	333394446		
	Lactobacillus curvatus CRL 705	1368	354808135		
	Lactobacillus fermentum 28-3-CHN	1313	260662220		
	Lactobacillus fermentum ATCC 14931	1381	227514633		
	Lactobacillus florum 2F	1327	408790128		
	Lactobacillus gasseri JV-V03	1391	300361537		
	Lactobacillus hominis CRBIP 24.179	1386	395244248		
	Lactobacillus jensenii 269-3	1391	238854567		
	Lactobacillus jensenii 27-2-CHN	1395	256852176		
	Lactobacillus johnsonii DPC 6026	1375	385826041		
	Lactobacillus mucosae LM1	1382	377831443		
	Lactobacillus paracasei subsp. *paracasei* 8700:2	1362	239630053		
5 (continued)	*Lactobacillus pentosus* IG1	1382	339637353		Type II-A
	Lactobacillus pentosus KCA1	1361	392947436		
	Lactobacillus pentosus MP-10	1358	334881121		
	Lactobacillus plantarum ZJ316	1358	448819853		
	***Lactobacillus rhamnosus* GG**	**1363**	**258509199**	**258509198**	
	Lactobacillus rhamnosus HN001	1361	199597394		
	Lactobacillus rhamnosus R0011	1361	418072660		
	Lactobacillus ruminis ATCC 25644	1375	323340068		
	Lactobacillus salivarius SMXD51	1339	418960525		
	Lactobacillus sanfranciscensis TMW 1.1304	1331	347534532		
	Lactobacillus sp. 66c	1419	408410332		
	Pediococcus acidilactici DSM 20284	1364	304386254		
	Pediococcus acidilactici MA18/5M	1366	418068659		
	Psychroflexus torquis ATCC 700755	1509	408489713		

Cluster[a]	Strain[b]	Cas9 length (aa)	Cas9 GI	Cas1GI[c]	Subtype[d]
6	*Anaerophaga* sp. HS1	1552	371776944		Type II-C
	Anaerophaga thermohalophila DSM 12881	1515	346224232		
	Bacteroides coprophilus DSM 18228	1509	224026357		
	Bacteroides coprosuis DSM 18011	1504	333031006		
	Bacteroides dorei DSM 17855	1504	212694363		
	Bacteroides eggerthii 1_2_48FAA	1509	317474201		
	Bacteroides faecis 27-5	1526	380696107		
	Bacteroides fluxus YIT 12057	1509	329965125		
	Bacteroides nordii CL02T12C05	1512	393788929		
	***Bacteroides* sp. 20_3**	**1517**	**301311869**	**301311870**	
	Bacteroides sp. D2	1510	383115507		
	Bacteroides uniformis CL03T00C23	1508	423303159		
	Bacteroides vulgatus CL09T03C04	1504	423312075		
	Capnocytophaga gingivalis ATCC 33624	1436	228473057		
	Capnocytophaga sp. CM59	1437	402830627		
	Capnocytophaga sp. oral taxon 324 str. F0483	1471	429756885		
	Capnocytophaga sp. oral taxon 326 str. F0382	1450	429752492		
	Capnocytophaga sp. oral taxon 412 str. F0487	1450	393778597		
	Chryseobacterium sp. CF314	1419	399023756		
	Fibrobacter succinogenes subsp. *succinogenes* S85	1512	261414553		
	Flavobacteriaceae bacterium S85	1516	372210605		
	Flavobacterium columnare ATCC 49512	1459	365960762		
	Fluviicola taffensis DSM 16823	1458	327405121		
	Mucilaginibacter paludis DSM 18603	1473	373954054		
	Myroides odoratus DSM 2801	1466	374597806		
	Ornithobacterium rhinotracheale DSM 15997	1535	392391493		Type II-C
	Prevotella bivia JCVIHMP010	1485	282858617		
	Prevotella buccae ATCC 33574	1457	315607525		
	Prevotella nigrescens ATCC 33563	1506	340351024		
	Prevotella sp. MSX73	1483	402307189		
	Prevotella timonensis CRIS 5C-B1	1487	282881485		
	Prevotella veroralis F0319	1496	260592128		
	Sphingobacterium spiritivorum ATCC 33861	1426	300771242		
	Weeksella virosa DSM 16922	1440	325955459		
7	*Bacteroides fragilis* 638R	1436	375360193		Type II-C
	***Bacteroides fragilis* NCTC 9343**	**1436**	**60683389**	**60683388**	
	Bacteroides sp. 2_1_16	1436	265767599		
	Bacteroides sp. 3_1_19	1424	298377533		
	Bacteroides sp. D2	1436	383110723		
	Bacteroidetes oral taxon 274 str. F0058	1434	298373376		
	Belliella baltica DSM 15883	1352	390944707		
	Bergeyella zoohelcum CCUG 30536	1430	406673990		
	Capnocytophaga canimorsus Cc5	1430	340622236		
	Capnocytophaga ochracea DSM 7271	1426	256819408		
	Capnocytophaga sp. oral taxon 329 str. F0087	1435	332882466		
	Capnocytophaga sp. oral taxon 335 str. F0486	1426	420149252		
	Capnocytophaga sp. oral taxon 380 str. F0488	1432	429748017		
	Capnocytophaga sputigena Capno	1426	213962376		
7 (continued)	*Flavobacterium psychrophilum* JIP02/86	1354	150025575		Type II-C
	Galbibacter sp. ck-I2-15	1391	408370397		
	Indibacter alkaliphilus LW1	1354	404451234		

Cluster[a]	Strain[b]	Cas9 length (aa)	Cas9 GI	Cas1GI[c]	Subtype[d]
	Joostella marina DSM 19592	1397	386818981		
	Kordia algicida OT-1	1391	163754820		
	Marinilabilia sp. AK2	1345	410030899		
	Myroides injenensis M09-0166	1401	399927444		
	Niabella soli DSM 19437	1426	374372722		
	Parabacteroides johnsonii DSM 18315	1443	218258638		
	Parabacteroides sp. D13	1424	256840409		
	Prevotella histicola F0411	1375	357042839		
	Prevotella intermedia 17	1380	387132277		
	Prevotella nigrescens F0103	1380	445119230		
	Prevotella oralis ATCC 33269	1391	323344874		
	Prevotella sp. oral taxon 306 str. F0472	1375	383811446		
	Riemerella anatipestifer RA-CH-1	1405	407451859		
	Riemerella anatipestifer RA-GD	1400	386321727		
	Zunongwangia profunda SM-A87	1388	295136244		
8	*Actinomyces coleocanis* DSM 15436	1105	227494853		Type II-C
	Actinomyces georgiae F0490	1113	420151340		
	Actinomyces naeslundii str. Howell 279	1101	400293272		
	Actinomyces sp. ICM47	1144	396585058		
	Actinomyces sp. oral taxon 175 str. F0384	1095	343523232		
	Actinomyces sp. oral taxon 181 str. F0379	1103	429758968		
	Actinomyces sp. oral taxon 848 str. F0332	1120	269219760		
	Actinomyces turicensis ACS-279-V-Col4	1114	405979650		
	Bifidobacterium dentium Bd1	1138	283456135		
	***Bifidobacterium longum* DJO10A**	**1187**	**189440764**	189440765	
	Bifidobacterium longum subsp. *longum* 2-2B	1124	419852381		
	Bifidobacterium longum subsp. *longum* KACC 91563	1138	384200944		
	Bifidobacterium sp. 12_1_47BFAA	1151	317482066	**317482065**	
	Corynebacterium accolens ATCC 49725	1099	227502575		
	Corynebacterium accolens ATCC 49726	1099	306835141		
	Corynebacterium diphtheriae 241	1084	375289763		
	Corynebacterium diphtheriae 31A	1084	376283539		
	Corynebacterium diphtheriae BH8	1084	376286566		
	Corynebacterium diphtheriae bv. *intermedius* str. NCTC 5011	1084	419861895		
	Corynebacterium diphtheriae C7 (beta)	1084	376289243		
	Corynebacterium diphtheriae HC02	1084	376292154		
	Corynebacterium diphtheriae NCTC 13129	1084	38232678		
	Corynebacterium diphtheriae VA01	1084	376256051		
	Corynebacterium matruchotii ATCC 14266	1089	305681510		
	Corynebacterium matruchotii ATCC 33806	1069	225021644		
	Gardnerella vaginalis 1500E	1186	415717744		
	Gardnerella vaginalis 284V	1186	415703177		
	Gardnerella vaginalis 5-1	1186	298252606		
	Mobiluncus curtisii subsp. *holmesii* ATCC 35242	1123	315656340		
	Mobiluncus mulieris 28-1	1091	269977848		
	Mobiluncus mulieris FB024-16	1091	307700167		
	Scardovia inopinata F0304	1178	294790575		
9	*Bacillus cereus* BAG4X12-1	1068	423439645		Type II-C
	Bacillus cereus BAG4X2-1	1078	423445130		
	Bacillus cereus Rock1-15	1069	229113166		

Cluster[a]	Strain[b]	Cas9 length (aa)	Cas9 GI	Cas1GI[c]	Subtype[d]
	***Bacillus smithii* 7_3_47FAA**	**1088**	**365156657**	365156658	
	Bacillus thuringiensis serovar *finitimus* YBT-020	1069	384183447		
	Brevibacillus laterosporus GI-9	1092	421874297	**421874296**	
	Clostridium perfringens C str. JGS1495	1065	169343975		
	Clostridium perfringens D str. JGS1721	1065	182624245		
	Sporolactobacillus vineae DSM 21990 = SL153	1084	404330915		
	Gemella haemolysans ATCC 10379	1392	241889924		
10	*Gemella morbillorum* M424	1385	317495358		**Type II-A**
	Megasphaera sp. UPII 135-E	1352	342218215		
	***Veillonella atypica* ACS-134-V-Col7a**	**1398**	**303229466**	**303229394**	
10 (continued)	*Veillonella parvula* ATCC 17745	1398	282849530		**Type II-A**
	Veillonella sp. 6_1_27	1395	294792465		
	Veillonella sp. oral taxon 780 str. F0422	1120	342213964		
	Treponema denticola AL-2	1395	449103686		
	Treponema denticola ASLM	1395	449106292		
	***Treponema denticola* ATCC 35405**	**1395**	**42525843**	**42525844**	
11	*Treponema denticola* H1-T	1395	449118593		**Type II-A**
	Treponema denticola H-22	1395	449117322		
	Treponema denticola OTK	1395	449125136		
	Treponema denticola SP37	1395	449130155		
	***Mycoplasma canis* PG 14**	**1233**	**384393286**	**384393287**	
	Mycoplasma canis PG 14	1233	419703974		
	Mycoplasma canis UF31	1233	384937953		
12	*Mycoplasma canis* UF33	1233	419704625		**Type II-A**
	Mycoplasma canis UFG1	1233	419705269		
	Mycoplasma canis UFG4	1233	419705920		
	Mycoplasma cynos C142	1239	433625054		
	Enterococcus faecalis Fly1	1150	257084992		
	Enterococcus faecalis R508	1150	424761124		
	Enterococcus faecalis T11	1150	257419486		
13	***Enterococcus faecalis* TX0012**	**1150**	**315149830**	**315149831**	**Type II-A**
	Enterococcus faecalis TX0012	1150	422729710		
	Enterococcus faecalis TX1342	1150	422701955		
	Facklamia hominis CCUG 36813	1142	406671118		
	Gluconacetobacter diazotrophicus PAl 5	1003	209542524		
	Gluconacetobacter diazotrophicus PAl 5	1050	162147907		
	Methylocystis sp. ATCC 49242	1080	323139312		
14	***Methylosinus trichosporium* OB3b**	**1082**	**296446027**	**296446028**	**Type II-C**
	Rhodopseudomonas palustris BisB18	1066	90425961		
	Rhodopseudomonas palustris BisB5	1064	91975509		
	Tistrella mobilis KA081020-065	1049	389874754		
	Francisella cf. *novicida* 3523	1646	387824704		
	Francisella cf. *novicida* Fx1	1629	385792694		
	Francisella novicida FTG	1629	208779141		
15	*Francisella novicida* GA99-3548	1629	254374175		**Type II-B**
	***Francisella novicida* U112**	**1629**	**118497352**	**118497353**	
	Francisella tularensis subsp. *novicida* GA99-3549	1629	254372717		
	Acidovorax avenae subsp. *avenae* ATCC 19860	1045	326315085		
16	*Alicycliphilus denitrificans* BC	1029	319760940		**Type II-C**
	***Alicycliphilus denitrificans* K601**	**1029**	**330822845**	**330822846**	
	gamma proteobacterium HdN1	1025	304313029		

Cluster[a]	Strain[b]	Cas9 length (aa)	Cas9 GI	Cas1GI[c]	Subtype[d]
	Nitrosomonas sp. AL212	1044	325983496		
	Verminephrobacter eiseniae EF01-2	1068	121608211		
17	*Mycoplasma gallisepticum* NC95_13295-2-2P	1269	401767318		Type II-A
	Mycoplasma gallisepticum NY01_2001.047-5-1P	1224	401768851		
	***Mycoplasma gallisepticum* str. F**	**1269**	**284931710**	**284931711**	
	Mycoplasma gallisepticum str. F	1269	385326554		
	Mycoplasma gallisepticum str. R(low)	1270	294660600		
18	*Prevotella buccalis* ATCC 35310	1218	282878504		Type II-C
	Prevotella ruminicola 23	1204	294674019		
	Prevotella stercorea DSM 18206	1216	359406728		
	Prevotella tannerae ATCC 51259	1234	258648111		
	***Prevotella timonensis* CRIS 5C-B1**	**1218**	**282880052**	**282880053**	
19	*Phascolarctobacterium succinatutens* YIT 12067	1087	323142435		Type II-C
	Roseburia intestinalis L1-82	1140	257413184		
	Roseburia intestinalis M50/1	1128	291537230		
	***Roseburia inulinivorans* DSM 16841**	**1152**	**225377804**	**225377803**	
	Subdoligranulum sp. 4_3_54A2FAA	1084	365132400		
20	***Coriobacterium glomerans* PW2**	**1384**	**328956315**	**328956316**	Type II-A
	Eggerthella sp. YY7918	1380	339445983		
20 (continued)	*Gordonibacter pamelaeae* 7-10-1-b	1371	295106015		Type II-A
	Olsenella uli DSM 7084	1399	302336020		
21	***Fusobacterium nucleatum* subsp. *vincentii* ATCC 49256**	**1374**	**34762592**	34762593	Type II-A
	Fusobacterium sp. 1_1_41FAA	1367	294782278		
	Fusobacterium sp. 3_1_27	1367	294785695		
	Fusobacterium sp. 3_1_36A2	1367	256845019	**256845020**	
22	*Finegoldia magna* ACS-171-V-Col3	1347	302380288		Type II-A
	***Finegoldia magna* ATCC 29328**	**1348**	**169823755**	**169823756**	
	Finegoldia magna SY403409CC001050417	1348	417926052		
	Helcococcus kunzii ATCC 51366	1338	375092427		
23	*Prevotella denticola* CRIS 18C-A	1422	325859619		Type II-C
	Prevotella micans F0438	1425	373501184		
	***Prevotella* sp. C561**	**1424**	**345885718**	**345885719**	
24	*Leuconostoc gelidum* KCTC 3527	1355	333398273		Type II-A
	***Oenococcus kitaharae* DSM 17330**	**1389**	**366983953**	**366983954**	
	Oenococcus kitaharae DSM 17330	1389	372325145		
25	*Anaerococcus tetradius* ATCC 35098	1361	227501312		Type II-A
	Lactobacillus iners LactinV 11V1-d	1369	309803917		
	***Peptoniphilus duerdenii* ATCC BAA-1640**	**1364**	**304438954**	**304438953**	
26	***Coprococcus catus* GD/7**	**1338**	**291520705**	**291520706**	Type II-A
	Dorea longicatena DSM 13814	1340	153855454		
	Ruminococcus lactaris ATCC 29176	1341	197301447		
27	***Staphylococcus pseudintermedius* ED99**	**1334**	**323463801**	**323463802**	Type II-A
	Staphylococcus pseudintermedius ED99	1334	386318630		
	Staphylococcus simulans ACS-120-V-Sch1	1112	414160476		
28	***Dinoroseobacter shibae* DFL 12**	**1079**	**159042956**	**159042957**	Type II-C
	Sphingobium sp. AP49	1110	398385143		
	Sphingomonas sp. S17	1090	332188827		
29	***Flavobacterium branchiophilum* FL-15**	**1473**	**347536497**	no *cas1*	Type II-C
	Flavobacterium columnare ATCC 49512	1535	365959402		

Cluster[a]	Strain[b]	Cas9 length (aa)	Cas9 GI	Cas1GI[c]	Subtype[d]
30	***Bifidobacterium bifidum* S17**	**1420**	**310286728**	**310286727**	**Type II-A**
	Scardovia wiggsiae F0424	1471	423349694		
31	*Burkholderiales* bacterium 1_1_47	1428	303257695		**Type II-B**
	***Parasutterella excrementihominis* YIT 11859**	**1428**	**331001027**	**331001028**	
32	***Streptococcus sanguinis* SK49**	**1421**	**422884106**	**422884107**	**Type II-A**
	Streptococcus sp. oral taxon 71 str. 73H25AP	1420	306826314		
33	*Eubacterium* sp. AS15	1391	402309258		**Type II-A**
	***Eubacterium yurii* subsp. *margaretiae* ATCC 43715**	**1391**	**306821691**	**306821690**	
34	*Legionella pneumophila* 130b	1372	307608922		**Type II-B**
	***Legionella pneumophila* str. Paris**	**1372**	**54296138**	**54296139**	
35	*Acidaminococcus intestini* RyC-MR95	1358	352684361		**Type II-A**
	***Acidaminococcus* sp. D21**	**1358**	**227824983**	**227824982**	
36	***Lactobacillus farciminis* KCTC 3681**	**1356**	**336394882**	**336394883**	**Type II-A**
	Lactobacillus versmoldensis KCTC 3814	1289	365906066		
37	*Mycoplasma synoviae* 53	1304	144575181		**Type II-A**
	***Mycoplasma synoviae* 53**	**1314**	**71894592**	**71894593**	
38	***Elusimicrobium minutum* Pei191**	**1195**	**187250660**	**187250661**	**Type II-C**
	uncultured Termite group 1 bacterium phylotype Rs-D17	1032	189485059		
39	*Clostridium spiroforme* DSM 1552	1116	169349750		**Type II-A**
	***Eubacterium dolichum* DSM 3991**	**1096**	**160915782**	**160915783**	
40	***Eubacterium rectale* ATCC 33656**	**1114**	**238924075**	**238924076**	**Type II-A**
	Eubacterium ventriosum ATCC 27560	1107	154482474		
41	*Staphylococcus aureus* subsp. aureus	1053	403411236		**Type II-A**
	***Staphylococcus lugdunensis* M23590**	**1054**	**315659848**	**315659847**	
42	***Ignavibacterium album* JCM 16511**	**1688**	**385811609**	**385811610**	**Type II-C**
43	***Odoribacter laneus* YIT 12061**	**1498**	**374384763**	**374384762**	**Type II-C**
44	***Caenispirillum salinarum* AK4**	**1442**	**427429481**	**427429479**	**Type II-C**
45	***Sutterella wadsworthensis* 3_1_45B**	**1422**	**319941583**	**319941582**	**Type II-B**
46	***Bergeyella zoohelcum* ATCC 43767**	**1415**	**423317190**	**423317188**	**Type II-C**
47	***Wolinella succinogenes* DSM 1740**	**1409**	**34557932**	**34557933**	**Type II-B**
48	***gamma proteobacterium* HTCC5015**	**1397**	**254447899**	no *cas1*	**Type II-B**
49	***Filifactor alocis* ATCC 35896**	**1365**	**374307738**	**374307737**	**Type II-A**
50	***Planococcus antarcticus* DSM 14505**	**1333**	**389815359**	**389815358**	**Type II-A**
51	***Catenibacterium mitsuokai* DSM 15897**	**1329**	**224543312**	**224543313**	**Type II-A**
52	***Solobacterium moorei* F0204**	**1327**	**320528778**	**320528779**	**Type II-A**
53	***Fructobacillus fructosus* KCTC 3544**	**1323**	**339625081**	**339625080**	**Type II-A**
54	***Mycoplasma ovipneumoniae* SC01**	**1265**	**363542550**	**363542551**	**Type II-A**
54	*Streptobacillus moniliformis* DSM 12112	1259	269123826		
55	***Mycoplasma mobile* 163K**	**1236**	**47458868**	**47458867**	**Type II-A**
56	***Porphyromonas* sp. oral taxon 279 str. F0450**	**1197**	**402847315**	**402847305**	**Type II-C**
57	***Actinomyces* sp. oral taxon 180 str. F0310**	**1181**	**315605738**	**315605739**	**Type II-C**
58	***Sphaerochaeta globus* str. Buddy**	**1179**	**325972003**	**325972002**	**Type II-C**
59	***Rhodospirillum rubrum* ATCC 11170**	**1173**	**83591793**	**83591790**	**Type II-C**
60	***Azospirillum* sp. B510**	**1168**	**288957741**	**288957738**	**Type II-C**
61	***Nitrobacter hamburgensis* X14**	**1166**	**92109262**	no *cas1*	**Type II-C**
62	***Ruminococcus albus* 8**	**1156**	**325677756**	**325677757**	**Type II-C**
63	***Barnesiella intestinihominis* YIT 11860**	**1153**	**404487228**	**404487227**	**Type II-C**
64	***Alicyclobacillus hesperidum* URH17-3-68**	**1146**	**403744858**	**403744859**	**Type II-C**

Cluster[a]	Strain[b]	Cas9 length (aa)	Cas9 GI	Cas1GI[c]	Subtype[d]
65	***Acidothermus cellulolyticus* 11B**	**1138**	**117929158**	**117929157**	**Type II-C**
66	***Nitratifractor salsuginis* DSM 16511**	**1132**	**319957206**	**319957207**	**Type II-C**
67	***Acidovorax ebreus* TPSY**	**1131**	**222109285**	**222109284**	**Type II-C**
67	*Francisella tularensis* subsp. *tularensis* WY96-3418	1125	134302318		
68	***Lactobacillus coryniformis* subsp. *torquens* KCTC 3535**	**1119**	**336393381**	**336393380**	**Type II-C**
69	***Alcanivorax* sp. W11-5**	**1113**	**407803669**	**407803668**	**Type II-C**
70	***Akkermansia muciniphila* ATCC BAA-835**	**1101**	**187736489**	**187736488**	**Type II-C**
71	***Ilyobacter polytropus* DSM 2926**	**1092**	**310780384**	**310780383**	**Type II-C**
72	***Bradyrhizobium* sp. BTAi1**	**1064**	**148255343**	no *cas1*	**Type II-C**
73	***Ralstonia syzygii* R24**	**1062**	**344171927**	**344171926**	**Type II-C**
74	***Treponema* sp. JC4**	**1062**	**384109266**	**384109265**	**Type II-C**
75	***Wolinella succinogenes* DSM 1740**	**1059**	**34557790**	**34557789**	**Type II-C**
76	***Rhodovulum* sp. PH10**	**1059**	**402849997**	**402849996**	**Type II-C**
77	***Aminomonas paucivorans* DSM 12260**	**1052**	**312879015**	**312879014**	**Type II-C**
77	*Bacteroides* sp. 3_1_33FAA	1055	265750948		
78	***Parvibaculum lavamentivorans* DS-1**	**1037**	**154250555**	**154250554**	**Type II-C**
79	***Candidatus Puniceispirillum* marinum IMCC1322**	**1035**	**294086111**	**294086112**	**Type II-C**
80	*Blastopirellula marina* DSM 3645	1027	87307579		
80	***Helicobacter mustelae* 12198**	**1024**	**291276265**	**291276264**	**Type II-C**
81	***Clostridium cellulolyticum* H10**	**1021**	**220930482**	**220930481**	**Type II-C**
82	*Lactobacillus crispatus* FB077-07	857	423321767		
82	**uncultured delta proteobacterium HF0070_07E19**	**1011**	**297182908**	no *cas1*	**Type II-C**
	Acetobacter aceti NBRC 14818	240	340779894		
	Acetobacter aceti NBRC 14818	376	340779669		
	Acetobacter aceti NBRC 14818	400	340779439		
	Actinobacillus ureae ATCC 25976	239	322514756		
	Actinobacillus ureae ATCC 25976	400	322514772		
	Bacillus cereus BAG2X1-3	333	423408783		
	Bacteroides cellulosilyticus DSM 14838	206	224535831		
	Bacteroides cellulosilyticus DSM 14838	1219	224535832		
	Bacteroides coprosuis DSM 18011	349	333031028		
	Bacteroides oleiciplenus YIT 12058	653	427387687		
	Bacteroides oleiciplenus YIT 12058	779	427387686		
	Bacteroides sp. 9_1_42FAA	1055	237710146		
	Bacteroides uniformis CL03T12C37	286	423308124		
	Bacteroides uniformis CL03T12C37	1210	423308121		
	Bifidobacterium bifidum IPLA 20015	1281	421736922		
	Bifidobacterium dentium ATCC 27678	1121	171742822		
	Bifidobacterium longum subsp. *longum* 1-6B	182	419848319		
	Bifidobacterium longum subsp. *longum* 1-6B	354	419847807		
	Bifidobacterium longum subsp. *longum* 1-6B	441	419848320		
	Bifidobacterium longum subsp. *longum* 44B	166	419856168		
	Bifidobacterium longum subsp. *longum* 44B	967	419856216		
	Butyrivibrio fibrisolvens 16/4	103	291518094		
	Butyrivibrio fibrisolvens 16/4	177	291518096		
	Butyrivibrio fibrisolvens 16/4	765	291518097		
	Campylobacter coli 2685	933	419548338		

Cluster[a]	Strain[b]	Cas9 length (aa)	Cas9 GI	Cas1GI[c]	Subtype[d]
	Campylobacter jejuni subsp. *jejuni* 2008-894	666	419652996		
	Campylobacter jejuni subsp. *jejuni* 305	190	317510779		
	Campylobacter jejuni subsp. *jejuni* 305	759	317510780		
	Campylobacter jejuni subsp. *jejuni* 327	462	415747744		
	Campylobacter jejuni subsp. *jejuni* 327	512	415747743		
	Campylobacter jejuni subsp. *jejuni* CG8421	721	205356639		
	Campylobacter jejuni subsp. *jejuni* M1	861	384442103		
	candidate division TM7 single-cell isolate TM7c	372	167957190		
	Capnocytophaga ochracea F0287	303	315224863		
	Capnocytophaga ochracea F0287	1117	315224862		
	Coprococcus comes ATCC 27758	686	226325213		
	Diplosphaera colitermitum TAV2	210	225164109		
	Enterococcus faecalis TX1467	921	422867931		
	Enterococcus faecalis TX4248	936	307270261		
	Enterococcus faecium E2620	892	431752788		
	Enterococcus sp. 7L76	116	295113136		
	Francisella tularensis subsp. *holarctica* 257	878	254367943		
	Francisella tularensis subsp. *holarctica* FSC022	158	254369498		
	Francisella tularensis subsp. *holarctica* FSC022	244	254369502		
	Francisella tularensis subsp. *holarctica* FSC022	292	254369497		
	Francisella tularensis subsp. *holarctica* FSC022	393	254369499		
	Francisella tularensis subsp. *holarctica* FSC022	501	254369496		
	Francisella tularensis subsp. *holarctica* LVS	158	89256630		
	Francisella tularensis subsp. *holarctica* LVS	393	89256631		
	Francisella tularensis subsp. *holarctica* URFT1	53	290953529		
	Francisella tularensis subsp. *holarctica* URFT1	285	290953528		
	Francisella tularensis subsp. *tularensis* SCHU S4	1123	56707712		
	Gemella haemolysans M341	1258	329766883		
	Haemophilus pittmaniae HK 85	121	343519651		
	Haemophilus pittmaniae HK 85	203	343519677		
	Haemophilus pittmaniae HK 85	650	343519679		
	Helicobacter hepaticus ATCC 51449	131	32266975		
	Helicobacter pullorum MIT 98-5489	344	242308998		
	Helicobacter pullorum MIT 98-5489	702	242309214		
	Kingella kingae ATCC 23330	1000	333374624		
	Lactobacillus buchneri ATCC 11577	1239	227512703		
	Lactobacillus casei 21/1	234	417984225		
	Lactobacillus casei 21/1	1128	417984226		
	Lactobacillus casei CRF28	566	417994652		
	Lactobacillus casei CRF28	700	417993346		
	Lactobacillus casei UW1	315	418005912		
	Lactobacillus casei UW1	330	418005913		
	Lactobacillus casei UW1	412	418005908		
	Lactobacillus casei UW4	236	418008739		
	Lactobacillus casei UW4	330	418008740		
	Lactobacillus crispatus 214-1	534	293381764		
	Lactobacillus crispatus CTV-05	298	312978192		
	Lactobacillus crispatus FB049-03	206	423318602		
	Lactobacillus crispatus FB049-03	347	423318603		
	Lactobacillus crispatus FB049-03	857	423318600		
	Lactobacillus crispatus JV-V01	278	227878395		

Cluster[a]	Strain[b]	Cas9 length (aa)	Cas9 GI	Cas1GI[c]	Subtype[d]
	Lactobacillus crispatus JV-V01	544	227878705		
	Lactobacillus crispatus MV-1A-US	277	256850790		
	Lactobacillus crispatus MV-1A-US	538	256850346		
	Lactobacillus crispatus MV-3A-US	279	262048056		
	Lactobacillus delbrueckii subsp. *bulgaricus* 2038	544	385815564		
	Lactobacillus delbrueckii subsp. *bulgaricus* 2038	669	385815562		
	Lactobacillus iners LactinV 09V1-c	255	309804524		
	Lactobacillus iners LactinV 09V1-c	343	309804534		
	Lactobacillus iners LactinV 09V1-c	447	309804536		
	Lactobacillus iners SPIN 2503V10-D	270	309809475		
	Lactobacillus iners SPIN 2503V10-D	667	309809480		
	Lactobacillus ruminis ATCC 25644	1352	417973941		
	Lactobacillus salivarius ACS-116-V-Col5a	629	301299400		
	Lactobacillus salivarius CECT 5713	897	385839899		
	Lactobacillus salivarius UCC118	1149	90961083		
	Leptospira inadai serovar Lyme str. 10	125	398345609		
	Leptospira inadai serovar Lyme str. 10	418	398341884		
	Leptospira inadai serovar Lyme str. 10	907	398345610		
	Leuconostoc pseudomesenteroides 4882	468	399517481		
	Leuconostoc pseudomesenteroides 4882	883	399517482		
	Listeria ivanovii FSL F6-596	232	315301622		
	Listeria ivanovii FSL F6-596	849	315301624		
	Listeria monocytogenes FSL F2-208	782	422410878		
	Listeria monocytogenes FSL J1-208	300	255024093		
	Listeria seeligeri FSL N1-067	874	313631816		
	Listeria seeligeri FSL N1-067	874	422420175		
	Maritimibacter alkaliphilus HTCC2654	997	84685065		
	Mycoplasma iowae 695	226	350547050		
	Mycoplasma iowae 695	933	350546886		
	Neisseria lactamica ATCC 23970	408	269215119		
	Neisseria lactamica ATCC 23970	666	269215120		
	Neisseria lactamica Y92-1009	241	422110930		
	Neisseria lactamica Y92-1009	828	422110931		
	Neisseria meningitidis NM3001	67	421568320		
	Neisseria meningitidis NM3001	976	421568319		
	Neisseria mucosa C102	220	319639577		
	Neisseria sp. oral taxon 20 str. F0370	392	429743981		
	Neisseria sp. oral taxon 20 str. F0370	701	429743980		
	Neisseria subflava NJ9703	587	284799897		
	Nitritalea halalkaliphila LW7	79	390445315		
	Nitrobacter hamburgensis X14	641	92118334		
	Oribacterium sinus F0268	653	227873236		
	Parabacteroides merdae ATCC 43184	103	154493351		
	Parabacteroides merdae CL03T12C32	84	423346601		
	Parabacteroides merdae CL09T00C40	82	423723156		
	Pasteurella bettyae CCUG 2042	398	387770127		
	Pasteurella bettyae CCUG 2042	610	387770112		
	Pasteurella multocida subsp. *multocida* str. Anand1_buffalo	199	421253447		
	Pasteurella multocida subsp. *multocida* str. Anand1_cattle	53	421259752		
	Pasteurella multocida subsp. *multocida* str.	63	421259756		

Cluster[a]	Strain[b]	Cas9 length (aa)	Cas9 GI	Cas1GI[c]	Subtype[d]
	Anand1_cattle				
	Pasteurella multocida subsp. *multocida* str. Anand1_cattle	134	421259749		
	Pediococcus acidilactici 7_4	1229	270290729		
	Pediococcus lolii NGRI 0510Q	270	427443367		
	Pediococcus lolii NGRI 0510Q	1016	427441502		
	Peptoniphilus sp. oral taxon 386 str. F0131	1341	299144352		
	Porphyromonas catoniae F0037	211	429741290		
	Porphyromonas catoniae F0037	1009	429741242		
	Prevotella denticola F0289	1218	327314511		
	Prevotella disiens FB035-09AN	443	303235616		
	Prevotella disiens FB035-09AN	795	303237415		
	Prevotella melaninogenica D18	1354	288802595		
	Prevotella multiformis DSM 16608	129	325268382		
	Prevotella multiformis DSM 16608	535	325268383		
	Prevotella oulorum F0390	691	345881543		
	Prevotella oulorum F0390	774	345881542		
	Prevotella saccharolytica F0055	242	429739781		
	Prevotella sp. oral taxon 317 str. F0108	593	288929745		
	Prevotella sp. oral taxon 317 str. F0108	1174	288930149		
	Prevotella sp. oral taxon 472 str. F0295	241	260910968		
	Prevotella sp. oral taxon 472 str. F0295	992	260910970		
	Pseudoramibacter alactolyticus ATCC 23263	586	315926102		
	Pseudoramibacter alactolyticus ATCC 23263	770	315926103		
	Rhizobium etli GR56	103	218671711		
	Riemerella anatipestifer ATCC 11845 = DSM 15868	1145	383485594		
	Sphingobacterium spiritivorum ATCC 33300	116	227540450		
	Sphingobacterium spiritivorum ATCC 33300	1306	227540451		
	Staphylococcus massiliensis S46	475	425737243		
	Staphylococcus massiliensis S46	581	425737242		
	Staphylococcus simulans ACS-120-V-Sch1	1112	410878248		
	Streptococcus agalactiae 18RS21	773	76799343		
	Streptococcus downei F0415	994	312866154		
	Streptococcus dysgalactiae subsp. *equisimilis* SK1249	538	417753185		
	Streptococcus dysgalactiae subsp. *equisimilis* SK1250	1155	417926916		
	Streptococcus mutans SA38	1229	449253007		
	Streptococcus mutans SA38	1229	449880497		
	Streptococcus oralis SK255	550	417794716		
	Streptococcus oralis SK255	670	417793840		
	Streptococcus pseudoporcinus SPIN 20026	1326	313890160		
	Streptococcus pyogenes M49 591	1052	56808315		
	Streptococcus sanguinis VMC66	1167	323351495		
	Streptococcus sp. BS35b	93	401683465		
	Streptococcus sp. GMD4S	206	419816637		
	Streptococcus sp. GMD4S	317	419819606		
	Streptococcus thermophilus CNCM I-1630	302	418027683		
	Streptococcus thermophilus CNCM I-1630	595	418027684		
	Streptococcus thermophilus MTCC 5461	39	445389093		
	Streptococcus vestibularis F0396	97	312863468		

Cluster[a]	Strain[b]	Cas9 length (aa)	Cas9 GI	Cas1GI[c]	Subtype[d]
	Streptococcus vestibularis F0396	1038	312863582		
	Sutterella parvirubra YIT 11816	406	378822098		
	Sutterella parvirubra YIT 11816	951	378821855		
	Sutterella wadsworthensis 2_1_59BFAA	389	422348538		
	Tannerella sp. 6_1_58FAA_CT1	976	365118488		
	Treponema denticola ATCC 33520	631	449107910		
	Treponema denticola ATCC 33520	769	449107911		
	Treponema denticola F0402	357	422340642		
	Treponema denticola F0402	370	422340641		
	Treponema denticola F0402	631	422340640		
	Treponema phagedenis F0421	591	320536383		
	Treponema phagedenis F0421	738	320536384		
	Treponema vincentii ATCC 35580	281	257456747		
	Treponema vincentii ATCC 35580	992	257456748		
	uncultured bacterium	600	406975829		
	uncultured bacterium	1017	406999582		
	uncultured bacterium T3_7_42578	675	411001094		
	uncultured Termite group 1 bacterium phylotype Rs-D17	166	189485058		
	uncultured Termite group 1 bacterium phylotype Rs-D17	1032	189485225		
	Verminephrobacter aporrectodeae subsp. *tuberculatae* At4	983	347820874		

[a]Cas9 sequences are grouped according to the BLASTclust clustering program. Truncated sequences were not selected for the analysis and are listed at bottom of the table without any cluster number (see Materials and Methods).
[b]Bacterial strains harboring *cas9* gene orthologue are listed; GI, GenInfo Identifier. Bold, cluster representatives chosen for the alignment and tree reconstruction. Grey, discarded, incomplete Cas9 sequences (see Materials and Methods). Note, that the incomplete sequences were all confirmed to be truncated Cas9 orthologues due to the presence of conserved motifs and similarity to the other Cas9 orthologues.
[c]Cas1 GenInfo Identifier of the representative sequences chosen for the alignment and tree reconstruction are given. Grey, discarded, incomplete sequences. When possible, alternative Cas1 sequence from the same cluster as the discarded Cas1 sequence was selected (clusters 8, 9 and 21, in bold).
[d]Type II CRISPR subtype of the CRISPR loci of the Cas9 cluster as inferred from the representative Cas1 and Cas9 trees topology.

[00497] Analysis of the composition of *cas* genes, transcription direction of the CRISPR arrays with respect to that of the *cas* operon, and location and orientation of tracrRNAs resulted in the division of subtypes into groups with distinct locus characteristics, especially within the subtype II-A (Figure 1, clusters marked with different colors) (15). We selected Cas9 enzymes representative of the major type II groups. Cas9 orthologs of *S. pyogenes, S. thermophilus** (CRISPR3) and *S. mutans* were chosen for type II-A systems associated with shorter, ~220 amino acid Csn2 variants (Csn2a). Cas9 of *S. thermophilus*** (CRISPR1) represents a distinct group of type II-A sequences associated with longer, ~350 amino acid

version of Csn2 orthologs (Csn2b). Cas9 of *F. novicida* was selected for type II-B. The closely related Cas9 orthologs of *P. multocida* and *N. meningitidis* and the distinct, short Cas9 of *C. jejuni* were chosen for type II-C (Figure 1B). Expression of associated tracrRNAs and crRNAs in *S. pyogenes*, *S. mutans*, *F. novicida, N. meningitidis* and *C. jejuni* was already validated by deep RNA sequencing (15,16). The RNAs in *S. thermophilus* and *P. multocida* were predicted bioinformatically based on the sequences from related species within the same type II group. Figure 1B shows the organization of the eight selected type II CRISPR-Cas loci and highlights our previous findings demonstrating that the type II loci architectures are highly variable among subtypes, yet conserved within each group (15). These variations are in good agreement with the clustering derived from the Cas9 and Cas1 phylogenetic trees (Figure 1A, Supplementary Figure S4).

[00498] Thus, to evaluate dual-RNA:Cas9 diversity, the bioinformatics analysis of type II CRISPR-Cas systems from available genomes identified Cas9 orthologs in a plethora of bacterial species that belong to 12 phyla and were isolated from diverse environments (Supplementary Tables S2 and S4). Most of the strains that harbor type II CRISPR-Cas systems (and accordingly Cas9) are pathogens and commensals of vertebrates. A majority of these strains were isolated from gastrointestinal tracts and feces of mammals, fish and birds, but also from wounds, abscesses and spinocereberal fluid of septicaemia patients. Strains were also isolated from invertebrates and environmental samples, including fresh and sea water, plant material, soil and food, the latter comprising species used in fermentation processes. Cas9 is also present in species from extreme environments such as deep sea sediments, hot springs and Antarctic ice, further demonstrating the wide spread of type II CRISPR-Cas systems in bacteria. A comparison of the taxonomy and habitats of representative strains with the phylogenetic clustering of Cas9 sequences shows little correlation (Supplementary Figure S11). In particular, clusters of Cas9 genes were identified from taxonomically distant bacteria that were isolated from similar habitats. Examples include diverse Firmicutes, Molicutes, Spirochaete and Fusobacteria, that were all isolated from gastrointestinal tracts of mammals, and members of different Proteobacteria, Firmicutes and Fusobacteria families mostly found in environmental samples (Supplementary Figure S11, clusters 1 and 3). A few exceptions involve grouping of Cas9 genes from closely related species isolated from diverse habitats such as Actinobacteria isolated from human and dog specimens but also from hot springs (Supplementary Figure S11, clusters 2, 4 and 5). This complex distribution of Cas9 across bacterial genomes indicates that evolution of dual-RNA:Cas9 systems in bacteria occurs both vertically and horizontally (55).

Example 2

[00499] Bacterial RNases III are interchangeable in dual-RNA maturation.

[00500] As described in *S. pyogenes* and *S. thermophilus*, RNase III plays an essential role in the biogenesis of dual-RNA:Cas9 systems by co-processing tracrRNA and pre-crRNA at the level of antirepeat:repeat duplexes (16,17). The interchangeability of *S. pyogenes* RNase III with RNases III from selected bacterial species was analyzed in the co-processing of *S. pyogenes* tracrRNA:pre-crRNA, including strains that lack type II CRISPR-Cas (*S. aureus* COL, *E. coli* TOP10). Northern blot analysis showed that all RNases III studied can co-process the RNA duplex (Figure 2, Supplementary Figure S5), indicating that there is no species-specificity for tracrRNA:pre-crRNA cleavage by RNase III. Multiple sequence alignment of RNase III orthologs demonstrates conservation of the catalytic aspartate residue and the dsRNA binding domain (Figure 2, Supplementary Figure S6) that are both required for RNA co-processing (Figure 2, Supplementary Figure S5). These data imply that the conservation of tracrRNA:pre-crRNA co-processing by bacterial RNase III provides a degree of flexibility allowing the functionality of dual-RNA:Cas9 systems in multiple species upon horizontal transfer.

[00501] Thus, to investigate the basis for the horizontal dissemination of CRISPR-Cas modules among bacteria, the specificity of RNase III utilized by type II CRISPR-Cas for dual-RNA maturation was analyzed. Complementation analysis shows that RNase III from a variety of species, including bacteria that lack type II CRISPR-Cas, can process *S. pyogenes* tracrRNA:pre-crRNA, suggesting that type II CRISPR-Cas systems can exploit any double-stranded RNA cleavage activity. This finding is consistent with the observation of *S. pyogenes* dual-RNA maturation in human cells which is apparently mediated by host RNases (2).

Example 3

[00502] Cas9 HNH and split RuvC domains are the catalytic moieties for DNA interference.

[00503] Comparison of Cas9 sequences revealed high diversity in amino acid composition and length (984 amino acid for *C. jejuni* to 1648 amino acids for *F. novicida*), especially in the linker sequence between the highly conserved N-terminal RuvC and central RuvC-HNH-RuvC regions and in the C-terminal extension (Supplementary Figure S2). Several studies demonstrated the importance of the nuclease motifs for dsDNA cleavage activity by mutating one aspartate in the N-terminal motif of the RuvC domain and one or several residues in the predicted catalytic motif of the HNH domain of the Cas9 enzyme (14,22,23). To investigate the relevance of all catalytic motifs for tracrRNA:pre-crRNA processing and/or DNA interference, alanine substitutions of selected residues were created (Figure 3A). In addition to the already published catalytic amino acids, we created Cas9 point mutants of conserved amino acid residues in the central RuvC motifs (14) (Figure 3A, Supplementary Figure S2). Northern blot analysis of *S. pyogenes cas9* deletion mutant complemented with each of the *cas9* point mutants revealed the presence of mature tracrRNA and crRNA forms, demonstrating that none of the catalytic motifs is involved in dual-RNA maturation by RNase III. This is in agreement with previous data showing that

RNase III is the enzyme that specifically cleaves tracrRNA:pre-crRNA duplex (16). Cas9 seems to have a stabilizing function on dual-RNA. We show that the catalytic motifs are not involved in RNA duplex stabilization (Figure 3B, Supplementary Figure S7).

[00504] To investigate the involvement of the conserved motifs of Cas9 in DNA interference *in vivo*, a previously described plasmid-based read-out system was used that mimics infection with invading protospacer-containing DNA elements (16). Transformation assays were done in *S. pyogenes* WT or a *cas9* deletion mutant using plasmids containing the *speM* protospacer gene (complementary to the second spacer of *S. pyogenes* SF370 type II CRISPR array (16)) and WT or mutant *cas9* (Figure 3C). In this assay, Cas9 expressed following plasmid delivery in bacterial cells catalyzes its own vector cleavage, when active. Control experiments showed that the *speM* protospacer-containing plasmid was not tolerated in WT *S. pyogenes*, demonstrating activity of WT CRISPR-Cas. Similarly, a plasmid containing the *speM* protospacer and encoding WT Cas9 could not be maintained in the *cas9* deletion mutant, demonstrating that Cas9 is able to cleave the plasmid from which it is expressed. Except for Cas9 N854A, all plasmids encoding Cas9 mutants were tolerated in the *cas9* deletion strain, indicating abrogation of Cas9 interference activity for these variants.

[00505] The *in vivo* DNA targeting data were confirmed with *in vitro* DNA cleavage assays. Purified WT and mutant Cas9 proteins were incubated with tracrRNA:crRNA targeting *speM* and subjected to cleavage of plasmid DNA containing the *speM* protospacer. WT and N854A Cas9 show dsDNA cleavage activity, whereas the other Cas9 mutants cleave only one strand of the dsDNA substrate, yielding nicked open circular plasmid DNA (Figure 3D). This corroborates the results obtained *in vivo* showing the importance of the conserved nuclease motifs for DNA interference by Cas9. In addition to the previously published data demonstrating the importance of the N-terminal RuvC motif and the catalytic motif of HNH, we thus defined new catalytic residues in the central RuvC motifs.

[00506] Dual-RNA and Cas9 sequences have widely evolved in bacteria (15). However, despite the high sequence variability among Cas9 sequences, certain motifs are conserved. In addition to the previously identified central HNH and N-terminal RuvC catalytic motifs (20,21,44,56), we show that the two middle RuvC motifs are required for interference activity *in vivo* and *in vitro*. In agreement with previous findings, deactivation of either one of the catalytic motifs (RuvC or HNH) results in nicking activity of Cas9 originating from the other motif (2,8,24,25). None of the mutations introduced in these conserved motifs affected the role of Cas9 in tracrRNA:pre-crRNA maturation by RNase III *in vivo*.

Example 4

[00507] Only Cas9 from closely related CRISPR-Cas systems can substitute for *S. pyogenes* Cas9 in tracrRNA-directed pre-crRNA maturation by RNase III.

[00508] Beside the conservation of the HNH and split RuvC domains involved in DNA cleavage (14,15), the length of Cas9 orthologs and the amino acid sequences of Cas9 are highly variable among the different groups of type II CRISPR-Cas systems (Figure 4A, Supplementary Figure S2). Hence, whether this variability plays a role in the specificity of Cas9 with regard to tracrRNA:pre-crRNA duplex and mature crRNA stabilization was investigated. A *S. pyogenes cas9* deletion mutant was complemented with Cas9 from selected bacterial species representative of the various type II groups and analyzed tracrRNA:pre-crRNA processing by Northern blot. Cas9 proteins from *S. mutans* and *S. thermophilus** can substitute for the stabilizing role of *S. pyogenes* Cas9 in RNA processing by RNase III (Figure 4B, Supplementary Figure S8). By contrast, Cas9 from *S. thermophilus***, *C. jejuni*, *N. meningitidis*, *P. multocida* and *F. novicida* could not complement the lack of RNA processing in the *cas9* mutant of *S. pyogenes*. In these strains, the 75-nt processed form of tracrRNA is observed as a very weak signal of background level of dual-RNA processed by RNase III in the absence of Cas9. Overall, only Cas9 from closely related systems of *S. pyogenes* in the type II-A cluster can substitute endogenous Cas9 role in dual-RNA stabilization and subsequent maturation by RNase III.

[00509] Thus, substitution of orthologs from the selected species for the endogenous *S. pyogenes* Cas9 shows that only Cas9 proteins from the *S. pyogenes* subcluster are capable of assisting tracrRNA:pre-crRNA processing by RNase III. This result indicates that the less-conserved inter-motif regions, which are the basis for the Cas9 subgrouping, could be responsible for Cas9 specificity for certain dual-RNAs.

Example 5

[00510] Cas9 orthologs require their specific PAM sequence for DNA cleavage activity.

[00511] In *S. pyogenes* and *S. thermophilus** types II-A, PAMs were identified as NGG and NGGNG, respectively. In these two species, mutating the PAM abrogates DNA interference by dual-RNA:Cas9 (14,22,23). To identify the functional PAMs for Cas9 from bacterial species other than *S. pyogenes* and *S. thermophilus*, potential protospacers matching spacer sequences in the selected CRISPR arrays were searched using BLAST. For *S. mutans* UA159, *C. jejuni* NCTC 11168, *P. multocida* Pm70 and *F. novicida* U112, potential protospacers were identified. Therefore, strains that harbor a closely related variant of Cas9 (Supplementary Table S2) were searched and their spacer sequences analyzed following the same approach (Supplementary Table S3). The identified 10 nt sequences located directly downstream of the protospacer sequence were aligned and the most common nucleotides that could represent PAM sequences were delineated. Based on the data visualized as a logo plot (Figure 5A), plasmid DNA substrates were designed containing the *speM* protospacer followed by different adjacent sequences either comprising the predicted PAM or not (Figure 5B). The Cas9 orthologous proteins were purified (Supplementary Figure S1) and dual-RNA orthologs were designed based on deep RNA sequencing data

(15), with the spacer sequence of crRNA targeting *speM*. To determine the protospacer-adjacent sequences critical for efficient DNA targeting, the purified Cas9 orthologs and their cognate dual-RNAs were used in DNA cleavage assays with different plasmid substrates (Figure 5C, Supplementary Figure S9). The previously published PAMs for Cas9 from *S. pyogenes* (NGG), *S. mutans* (NGG), *S. thermophilus** (NGGNG) and *N. meningitidis* (NNNNGATT) (27,28,53,54) were confirmed by multiple sequence alignments and *in vitro* cleavage assay, validating our approach. However, dual-RNA guided Cas9 from *S. thermophilus** could efficiently cleave target DNA in the presence of only NGG instead of NGGNG (Supplementary Figure S9). This is in contrast to data obtained *in vivo*, where mutation of the third G abrogates interference by Cas9 of *S. thermophilus** (23). For *S. thermophilus****, the PAM was published as NNAGAAW (27), which differs by one base from the sequence that we derived (NNAAAAW). *In vitro* cleavage assays with these two sequences demonstrate that the DNA substrate with the "NNAAAAW" PAM is cleaved more efficiently by Cas9 of *S. thermophilus**** compared to the "NNAGAAW" PAM (Supplementary Figure S9).

[00512] Using the same approach, the PAM activity of the most common protospacer-downstream sequences for *C. jejuni*, *F. novicida* and *P. multocida* were validated by *in vitro* cleavage assays, resulting in the most probable PAM sequences being NNNNACA (*C. jejuni*), GNNNCNNA (*P. multocida*) and NG (*F. novicida*) (Figure 5C, Supplementary Figure S9). Analysis of the protospacer-adjacent sequence from *C. jejuni* shows the same frequency of C and A ("NNNNCCA" or "NNNNACA") at position 5 downstream of the protospacer (Supplementary Table S3). Hence, both substrates were tested for cleavage activity by *C. jejuni* dual-RNA:Cas9. Only the DNA target containing A at this position was cleaved efficiently (Supplementary Figure S9). This result could be explained by the origin of the protospacer, with the "NNNNCCA" PAM being mostly found in genomic DNA or prophages of *Campylobacter* strains. In this case, the mutated PAM sequence on the chromosomally located protospacer prevents self-targeting. The *P. multocida* PAM requires further verification given that the multiple sequence alignment was derived from only two protospacer sequences. Thus, a series of specific PAMs that enable dsDNA cleavage by dual-RNA:Cas9 complexes from different bacterial species *in vitro* were identified. For gene editing purposes, it is contemplated that a range of potential motifs be analyzed to select those PAMs that would allow efficient targeting with limited off-site effect.

Supplementary Table S3. Overview of type II CRISPR-Cas spacer sequences from selected bacterial strains with BLAST candidate protospacers and their downstream sequence.

Strain[a]	Number of spacers	CRISPR Spacer[b]	Spacer sequence	Blast candidate[c]	% Identity[d]	10 bp downstream protospacer[e]
Streptococcus pyogenes SF370 (Accession: NC_002737)	6	1	TGCGCTGGTTGATTTCTTCTTGC GCTTTTT	*S. pyogenes* MGAS1882 (MGAS1882_1116), MGAS8232 (spyM18_0769), MGAS10394 (M6_Spy0995, M6_Spy1349), SSI-1 (SPs0926), ΦP9 **endopeptidase gene**	100	TGGCTTT TTC
				S. pyogenes MGAS2096 (MGAS2096_Spy1450), A20 (A20_1472c), M1 476 (M1GAS476_1503), MGAS9429 (MGAS9429_Spy1426), MGAS5005 (M5005_Spy1424) **endopeptidase gene**	97	TGGCTTT TTC
				S. pyogenes M1 GAS (SPy_0700), MGAS2096 (MGAS2096_Spy0592) **endopeptidase gene**	97	TTGCTTT TTC
				S. pyogenes MGAS6180 (M28_Spy1234); NIH1 (NIH1.1_43), SSI-1 (SPs0647), MGAS315 (SpyM3_0930, SpyM3_1215) **phage related gene**	100	TGACTTT TTC
		2	TTATATGAACATAACTCAATTTG TAAAAAA	**gene for pyrogenic exotoxin M** (*speM*) of several Streptococci strains	100	GGGTATT GGG
		3	AGGAATATCCGCAATAATTAATT GCGCTCT	*S. pyogenes* MGAS8232 (spyM18_0742), MGAS10750 (MGAS10750_Spy0588), MGAS10270 (MGAS10270_Spy0563) **adenine specific methylase gene**	100	TGGTATG TTG
				S. pyogenes Manfredo (SpyM50653) **adenine specific methylase gene**	97	TGGTATG TTG
		4	AGTGCCGAGGAAAAATTAGGTGC GCTTGGC	*S. pyogenes* Alab49 (SPYALAB49_001176), MGAS10750 (MGAS10750_Spy1285), MGAS9429 (MGAS9429_Spy0843), MGAS10394 (M6_Spy1203), SSI-1 (SPs0763), MGAS315 (SpyM3_1101), ΦH4489A (hylP) **hyaluronoglucosaminidase gene**	100	TGGCGCA TTA
				S. pyogenes MGAS8232 (spyM18_1254), NZ131 (Spy49_0785) **hyaluronoglucosaminidase gene**	97	TGGCGCA TTA
		5	TAAATTTGTTTAGCAGGTAAACC GTGCTTT	*S. pyogenes* MGAS10750 (MGAS10750_Spy0839), MGAS10270 (MGAS10270_Spy0546, MGAS10270_Spy0804), SSI-1 (SPs0517, SPs0888), MGAS1882 (MGAS1882_1156), MGAS8232, NZ131(Spy49_1511c), MGAS315 (SpyM3_0965, SpyM3_1347) **phage protein gene or intergenic region**	100	TCGTTAT ATC

Strain[a]	Number of spacers	CRISPR Spacer[b]	Spacer sequence	Blast candidate[c]	% Identity[d]	10 bp downstream protospacer[e]
Streptococcus mutans UA159 (Accession: NC_004350)	5	3	CTAACTATGATGACACAACAGCTTTTAGCG	ΦM102 (orf13) **putative tail protein gene**	100	TTGAAATTTC
Streptococcus mutans LJ23 (Accession: NC_017768)	8	2	TGAAGTGCAAGCTTACGTGACTGACTCGCG	ΦM102 (orf15) **putative minor structural protein**	90	AGGTATGCAG
Streptococcus mutans GS-5 (Accession: NC_018089)	21	3	TAATAGCAATCGTGACGGACGTATTGATTT	ΦM102 (orf15) **putative minor structural protein**	97	AGGTGAAATT
		5	GTTGAGTGCAACAGCTAGCTAATAGCTTTT	ΦM102	100	AGGCTGGCAC
		16	AGGCATTTTCTGATTGAGATTTTCGATATT	ΦM102 (orf3) **putative large terminase gene**	93	TGGAAAGATG
		18	TATAGCTAATATGTGTATACTGACAGCGCA	ΦM102 (orf7) **putative DNA packaging protein gene**	100	AGAAAGATTG
Streptococcus mutans NN2025 (Accession: NC_013928)	69	2	GATTGTGCCCGCTAGTAAACCGCCTCGCGC	ΦM102 (orf20) **putative endolysin gene**	93	TGGAGATTTG
		6	GATTGTATCAGTAATCGAACTTCTGCTTAT	ΦM102 (orf38, orf39) **hypothetical protein gene**	93	GGGATTTGAC
		8	TGGTCCAAAGTGCAGAGCCAAAGAAAAACA	ΦM102 (orf11) **putative major tail protein gene**	97	AAGCGGTCTT
		9	ATTGTCAATCGCCGTTCTGCGCTTGCGACG	ΦM102 (orf17) **hypothetical protein gene**	90	CGGTTTTGAA
		17	GCTTGAATATAATTGTGTATCCGCCAATGA	ΦM102 (orf21) **putative replisome organizer gene**	93	CGAATTACGA
		23	AAAAAGAAACGCCTTTTGATTTGACCAATC	ΦM102 (orf14) **putative receptor-binding protein gene**	90	AAGAGCAAGA
		29	AGTTATTAATATCTATGACAGTCTCAAAGA	ΦM102 (orf14) **putative receptor-binding protein gene**	93	GGGCTACAGA
		37	TTCTGGCTGTCTTTCAGAGTGATAAGCGCA	ΦM102 (orf2) **putative small terminase gene**	100	GGATTTTTCA
		40	TGCAAGTTATCTTGCTATGTGGACGAATTG	ΦM102 (orf9) **hypothetical protein gene**	93	GGGAACAATC
		43	GCAATTTAGTTTTATTCCGTGGGAGCAGCA	ΦM102 (orf3) **putative large terminase gene**	93	AGGCTCCATT

Strain[a]	Number of spacers	CRISPR Spacer[b]	Spacer sequence	Blast candidate[c]	% Identity[d]	10 bp downstream protospacer[e]
		48	AGAGTATAGCCAGTGTTTTCAAGGCCTTTA	ΦM102 (orf12) **putative tape measure protein gene**	93	GTGGTGACAA
		49	CGCAACAATGACTATTAATATCAACGGTGG	ΦM102 (orf15) **putative minor structural protein gene**	93	CGGGAGCAAT
		56	AATCGCTTCTTTGCTAACCACAATTTGTGC	ΦM102 (orf26) **putative RecT family single-strand annealing protein gene**	93	AGGCGCAGAG
		60	AAATGCTCTTGAAGAACCTGATAGATGACA	ΦM102 (orf3) **putative large terminase gene**	93	GAGACGAAAA
		66	TGCAAAAGATGGCCTCGAGCAATTATCGCA	ΦM102 (orf33) **hypothetical protein gene**	100	TGGATTAAGC
Streptococcus thermophilus LMD-9 CASS4 locus (Accession: NC_008532)	8	2	TCAATGAGTGGTATCCAAGACGAAAACTTA	*Streptococcus thermophilus* plasmid pSt106 **putative resolvase gene**	100	TGGCAAGTTT
		3	CCTTGTCGTGGCTCTCCATACGCCCATATA	*Streptococcus thermophilus* plasmid pND103	100	AGGCGGCGGGT
		4	TGTTTGGGAAACCGCAGTAGCCATGATTAA	Φ7201 (orf33)	100	AGGTGTCGCT
		5	ACAGAGTACAATATTGTCCTCATTGGAGACAC	Φ TP-J34 (orf11) **hypothetical protein gene**	94	TGGCGGTAGGA
		6	CTCATATTCGTTAGTTGCTTTTGTCATAAA	ΦSfi19 (orf1626) **minor tail protein gene**	100	TGGTGCTAAT
				ΦYMC 2011 (SsaI_phage00063) **putative minor tail protein gene**	90	GGGTGCTAAC
				Φ7201 (orf33)	90	TGGTGCTAGA
Streptococcus thermophilus LMD-9 CASS4a locus (Accession: NC_008532)	16	2	CTTCACCTCAAATCTTAGAGCTGGACTAAA	Φ7201 (orf39)	100	GTAGAATAGA
		3	ATGTCTGAAAAATAACCGACCATCATTACT	Φ TP-J34 (orf49), ΦSfi11 (orf669) **putative minor structural protein gene**	93	CCAGAATGTC
		4	GAAGCTCATCATGTTAAGGCTAAAACCTAT	ΦALQ13.2 (orf35) **helicase gene**	90	CTAAAAATTA
				ΦSfi11 (orf443), ΦSFi18 (orf443), ΦSfi21 (orf443), ΦSfi19 (orf443), ΦO1205 (orf10) **putative helicase gene**	90	CTCAAAATTA
		5	TAGTCTAAATAGATTTCTTGCACCATTGTA	Φ1033, Φ 1042 **nonfunctional host specificity protein gene**	97	ATAAAATTCA

Strain[a]	Number of spacers	CRISPR Spacer[b]	Spacer sequence	Blast candidate[c]	% identity[d]	10 bp downstream protospacer[e]
				ΦDT1.1 (orf18), ΦDT1.2 (orf18), ΦDT1.3 (orf18), ΦDT1.4 (orf18), ΦDT1.5 (orf18), ΦMD4 (orf18) **host specificity protein gene**	93	ATAAAAT TCA
		6	ATTCGTGAAAAAATATCGTGAAA TAGGCAA	pSt08 plasmid	97	CCGAAAA ATA
		7	TCTAGGCTCATCTAAAGATAAAT CAGTAGC	ΦALQ13.2 (orf25), Φ858 (orf30), ΦST3 (orf253) **endonuclease gene**	90	TGAAAAA TTA
				ΦJ1 (orf253), ΦS3b (orf253) **endonuclease gene**	90	TGAAAGA TTA
		13	AACTACCAAGCAAATCAGCAATC AATAAGT	ΦSfi11	100	TCAAGAA TAT
				ΦYMC-2011 (SsaI_phage00051) **predicted clp-protease gene**	93	TTAAGAA CAT
				ΦSfi21 (orf221) **clp-protease gene**	90	TCAAGAA TAT
		16	AACAGTTACTATTAATCACGATT CCAACGG	Φ858 (orf22)	93	AAAAAAA ACT
				Φ2972 (orf21) **structural protein gene**	93	AAAAAAA ACT
				ΦAbc2 (orf17) **tail protein gene**	93	TAAAAAG ACT
Campylobacter jejuni subsp. jejuni NCTC 11168 (Accession: NC_002163)	5	1-5		no significant BLAST hits		
Campylobacter jejuni subsp. jejuni CF93-6 (Accession: AANJ00000000)	5	3	TCATCATCACTTAAAACCTTAAA TTTACC	*C. jejuni* RM1221 (CJE1445) **hypothetical protein gene**	93	ATAACGC AAG

Strain[a]	Number of spacers	CRISPR Spacer[b]	Spacer sequence	Blast candidate[c]	% Identity[d]	10 bp downstream protospacer[e]
Campylobacter jejuni subsp. jejuni HB93-13c_jejuni_subsp_jejunihb_13_42 (Accession: AANQ00000000)	9	1	GCATTGCTTTACTACATAGCCAGTCGTGTA	*C. jejuni subsp. doylei* 269.97 (JJD26997_1148) **conserved hypothetical protein gene**	100	TCACACACGC
		2	TTATTTTTGTCGCTAATTGCACCTAAAGAC	*C. jejuni subsp. doylei* 269.97 (JJD26997_0867) **putative primase gene**	97	TCCAACACCAT
Campylobacter jejuni subsp. jejuni NW genomic scaffold Mich_State_Univ:Contig3 (Accession: JH376989 REGION: 13521..15062)	5	5	GGGACACGAGGAATCCTGTCTGAATCCGGG	*C. jejuni subsp. jejuni* PT14 (A911_r08426, A911_r08428, A911_r08430), NCTC 11168-BN148 (BN148_r02, BN148_r05, BN148_r08), S3 (CJS3_1811, CJS3_1817, CJS3_1830), ICDCCJ07001 (ICDCCJ07001_29, ICDCCJ07001_396, ICDCCJ07001_718), M1 (CJM1_0031, CJM1_0413, CJM1_0727), IA3902 (CJSA_Cj23SA, CJSA_Cj23SB, CJSA_Cj23SAC), BABS091400, 81116 (C8J_Cj23SA, C8J_Cj23SB, C8J_Cj23SC), 81-176 (CJJ81176_1714, CJJ81176_1727, CJJ81176_1707), NCTC 11168; *C. jejuni* DSM 4688, UNSW091300, strain 100, RP0001, 102-27 (rrlC, rrlB, rrlA), 69-30 (rrlC, rrlB, rrlA), 140-16 (rrlC, rrlB, rrlA), 110-21 (rrlC, rrlB, rrlA), RM1221 (CJE_Cj23SA, CJE_Cj23SB, CJE_Cj23SC), TGH9011_ATCC43431 (rrl); *C. coli* 59-2 (rrlC, rrlB, rrlA); *C. jejuni subsp. doylei* 269.97 (JJD26997_0040, JJD26997_1264, JJD26997_1520) **23S rRNA gene**	100	TCGACCACGA
Campylobacter jejuni subsp. doylei 269.97 (Accession: NC_009707)	5	2	CTAAGCAATCTTATTTTACCATCTTTTTTA	*C. jejuni* strain TGH 9011 (Tgh093)	97	TAAAACACTT
				C. jejuni RM1221 (CJE1099) **hypothetical protein gene**	93	TAAAACACTT
Campylobacter jejuni subsp. jejuni 1336 (Accession: NZ_CM000854 NZ_ADGL01000000)	2	1	TTACTGATATTAAAATTAACTCCATAATTT	*C. jejuni* 00-3477 (cje0227), *C. jejuni subsp. jejuni* S3 (CJS3_0723), ΦCGC-2007 **prophage related genes**	100	GCTGCCATTA
				C. jejuni NCTC 13255 (putative CJIE1-2-like prophage), 99-7046 (putative CJIE1-3-like prophage), 00-2425 (putative CJIE1 prophage), RM1221 (CJE0227)*C. jejuni subsp. jejuni* ICDCCJ07001 (ICDCCJ07001_691) **major tail sheath protein**	93	GCTGCCATTA

Strain[a]	Number of spacers	CRISPR Spacer[b]	Spacer sequence	Blast candidate[c]	% Identity[d]	10 bp downstream protospacer[e]
		2	ATAAAGCTAATGCAAAAGTTGAAAACAAA	*C. jejuni* NCTC 13255 (putative CJIE1-2-like prophage), 99-7046 (putative CJIE1-3-like prophage), 00-3477 (putative CJIE1-4 Mu-like prophage), 00-2425 (putative CJIE1 prophage), RM1221 (CJE0238),*C. jejuni subsp. jejuni* S3 (CJS3_0704), ICDCCJ07001, *C. hyoilei* **hypothetical protein gene**	100	AGAGCTA TAA
Campylobacter jejuni subsp. jejuni 414 (Accession: NZ_CM000855 NZ_ADGM01000000)	33	2	TTTATCTGCATCCATAATGGCAATGAGTGA	*C. jejuni subsp. jejuni* PT14 (A911_03310), NCTC 11168-BN148 (BN148_0680c), S3 (CJS3_0675), ICDCCJ07001 (ICDCCJ07001_619), M1 (CJM1_0650), IA3902 (CJSA_0644), 81116 (C8J_0632), 81-176 (CJJ81176_0703), NCTC 11168 (Cj0680c), P694a (Cj0680c), P569a (Cj0680c), P179a (Cj0680c), H73020 (Cj0680c), H704a (Cj0680c), *C. jejuni* RM1221 (CJE0778), *C. jejuni subsp. doylei* 269.97 (JJD26997_1327) **excinuclease ABC subunit B gene**	97	AACTTCA GGC
Neisseria meningitidis serogroup A strain Z2491 (Accession: NC_003116)	16	2	CTTCTGCCTTTTTACAAGCTCGCTTTCTTT	*N. gonorrhoeae* (NGU65994, PivNG), FA 1090 (NGO1137, NGO1164, NGO1262) **invertase related genes, phage associated protein genes**	97	CGCCGAC CGG
		3	TTTGGTAAAGGTTTCTGTTGCGACCCGAAT	*N.meningitidis* NZ-05/33 (NMBNZ0533_1722), M04-240196 (NMBNZ0533_1722), M01-240149 (NMBH4476_1701), H44/76 (NMBH4476_1701) **hypothetical proteins upstream of transposase gene**	100	GTCTGAT TTT
				N. lactamica isolate 3207487 (plasmid pNL3.2), *N. lactamica* (plasmid pNL9)	97	GTCTGAT TTT
				N. gonorrhoeae TCDC-NG08107, NCCP11945 **intergenic region (putative phage proteins)**	93	GGCTGTT TTT
				N. gonorrhoeae NCCP11945 (NGK_1948, NGK_1990, NGK_2023) **hypothetical protein genes**	93	GTCTGAT TTT
				N. gonorrhoeae **intergenic region PivNG**	93	GTCTGAT TTT
				N. gonorrhoeae FA 1090 **numerous intergenic regions in prophages**	93	GTCTGAT TTT
		7	AAATTCGTTTCAGATAGCAAACGCAGTAGT	*N. gonorrhoeae* TCDC-NG08107, *N. gonorrhoeae* NCCP11945 **intergenic region (putative phage proteins)**	97	GGACGAT TTC
		12	GGGTAGCCAGTGCTAAAACCGCACCCGCTT	*N. lactamica* plasmid pNL9	93	TGCCGGC ATA
				N. meningitidis plasmid pJS-B	100	TACGAAC ATT
		13	CCAAATAGAAATACATACGCCGAGTAATTA	*N. lactamica* plasmid pNL9	93	AGCTGCT TTG

Strain[a]	Number of spacers	CRISPR Spacer[b]	Spacer sequence	Blast candidate[c]	% Identity[d]	10 bp downstream protospacer[e]
				N. meningitidis plasmid pJS-B	97	AGCCGCT TTG
		14	TTTCTTTTTGTAATTGTTCTGCC TTTTTTA	*N. lactamica* plasmid pNL9	100	ATTGGAT TTT
				N. meningitidis plasmid pJS-B	100	ATTGGAT TTT
		15	TACCCACGGCGGAAACCATTGCC ACAAAAC	*N. meningitidis* strain alpha522 draft genome (NMALPHA522_0671), H44/76 (NMBH4476_0684), 053442 (NMCC_0153), *N. meningitidis* serogroup C FAM18 (NMC1864) **hypothetical protein gene**	100	CCATGAT TAC
				N.meningitidis M04-240196 (NMBM04240196_0048, NMBM04240196_0749) **putative membrane protein gene**	100	CCATGAT TAC
Pasteurella multocida str. Pm70 (Accession: NC_002663)	5	1-5		no significant BLAST hits		
Pasteurella multocida subsp. gallicida X73 (Accession: CM001580 AMBP01000000)	20	9	AAAGAATACACCCTTATTCCAAA AAGTTTG	*P. multocida* 1.8 kb plasmid	100	GCGAGAG ATG
		15	GTCTGAACAGTATTAACACTTCC TGTTTCT	*P. multocida subsp. multocida* str. HN06(PMCN06_2098) **hypothetical protein gene**	97	GGATCGT ACT
Francisella tularensis subsp. novicida U112 (Accession: NC_008601)	13	1-13		no significant BLAST hits		
Francisella novicida FTG	22	15	ATCTCAAAAGCAGCTCTTTCGCG TGTAATATCGTT	*F. cf. novicida* 3523 (FN3523_1002) **phage protein gene**	91	TGCATTA GAT

Strain[a]	Number of spacers	CRISPR Spacer[b]	Spacer sequence	Blast candidate[c]	% Identity[d]	10 bp downstream protospacer[e]
FTG scaffold 1 genomic scaffold (Accession: NZ_DS995363 NZ_ABXZ01000000)		19	CTATCTAAGAGAACTTACAAGAC AAGAGAAAATACT	*F. cf. novicida* 3523 (FN3523_0993) **hypothetical protein gene**	94	TGGTGGT AAA
Francisella tularensis subsp. novicida GA99-3548 supercont1.3 (Accession: DS264589 ABAH01000000)	10	2	AGCCCTATCAGAAATATATGCAA GTTTGAATATAG	*F. cf. novicida* 3523 (FN3523_1009) **phage-related baseplate assembly protein gene**	89	AGGGTGT AGC
		3	AGATAACTCTTATATTGATTTGT ATATTGAAGATA	*F. cf. novicida* 3523 (FN3523_1006) **hypothetical protein gene**	94	TGATTAG CAT
		4	CGCAAAAAAGGCGAATTTGAGCA GAAAATTTGGGC	*F. cf. novicida* 3523 (FN3523_0999) **hypothetical protein gene**	91	TGGTATT GAT

[a]Selected strains used in this study. No potential protospacers were found for *Streptococcus mutans* UA159, *Campylobacter jejuni* subsp. *jejuni* NCTC 11168, *Pasteurella multocida* str. Pm70 and *Francisella tularensis* subsp. *novicida* U112. Therefore, closely related strains were analyzed for the presence of type II CRISPR-Cas arrays. Spacer sequences from selected arrays were then used to search for protospacer candidates.

[b]Numbering of spacers starts from the leader proximal end based on RNAseq data (15). Spacers with no significant protospacer BLAST hit are not listed in the table.

[c]A BLAST candidate was considered a potential protospacer when the identity to the spacer was ≥ 90% and when the protospacer originated either from phage, plasmid or genomic DNA related to the analyzed species. For each identified protospacer, the strain name, the protospacer-containing gene locus and the potential function of the gene are given.

[d]Percentage identity between spacer and protospacer sequence. e10 nt sequence located directly 3' of the protospacer sequence. The identified sequences for each bacterial species were aligned using GeneDoc (http://www.nrbsc.org/gfx/genedoc/). The degree of conservation is indicated with a color code (black: 100%, dark grey: ≥ 80%, light grey: ≥ 60%). These sequences were used to create the logo plot represented in Figure 5.

Supplementary Table S4 – Cas9 is present in bacteria from 12 different phyla and diverse habitats

Strain[a]	Class	Isolation/habitat[b]
Actinobacteria		
Actinobacteridae		
Acidothermus cellulolyticus 11B	*Acidothermaceae*	extremophile (hot water spring)
Actinomyces coleocanis	*Actinomycetaceae*	dog genital tract

Strain[a]	Class	Isolation/habitat[b]
Actinomyces georgiae F0490	*Actinomycetaceae*	oral cavity
Actinomyces naeslundii str. Howell 279	*Actinomycetaceae*	oral cavity
Actinomyces sp. ICM47	*Actinomycetaceae*	ND
Actinomyces sp. oral taxon 175 str. F0384	*Actinomycetaceae*	oral cavity
Actinomyces sp. oral taxon 180 str. F0310	*Actinomycetaceae*	oral cavity
Actinomyces sp. oral taxon 181 str. F0379	*Actinomycetaceae*	oral cavity
Actinomyces sp. oral taxon 848 str. F0332	*Actinomycetaceae*	oral cavity
Actinomyces turicensis ACS-279-V-Col4	*Actinomycetaceae*	genital tract
Bifidobacterium bifidum S17	*Bifidobacteriaceae*	gastrointestinal tract/feces
Bifidobacterium dentium Bd1	*Bifidobacteriaceae*	oral cavity
Bifidobacterium longum DJO10A	*Bifidobacteriaceae*	gastrointestinal tract/feces
Bifidobacterium sp. 12_1_47BFAA	*Bifidobacteriaceae*	gastrointestinal tract/feces
Corynebacterium accolens ATCC 49726	*Corynebacterineae*	wound
Corynebacterium diphteriae NCTC 13129	*Corynebacterineae*	oral cavity
Corynebacterium matruchotii ATCC 14266	*Corynebacterineae*	oral cavity
Gardnerella vaginalis 5-1	*Bifidobacteriaceae*	genital tract
Mobiluncus curtisii ATCC 35242	*Actinomycetaceae*	genital tract
Mobiluncus mulieris 28-1	*Actinomycetaceae*	genital tract
Scardovia inopinata F0304	*Bifidobacteriaceae*	oral cavity
Scardovia wiggsiae F0424	*Bifidobacteriaceae*	oral cavity
Coriobacteridae		
Coriobacterium glomerans PW2	*Coriobacteriaceae*	invertebrate (red soldier bug)
Eggerthella sp. YY7918	*Coriobacteriaceae*	gastrointestinal tract/feces
Gordonibacter pamelaeae 7-10-1-b	*Coriobacteriaceae*	gastrointestinal tract/feces
Olsenella uli DSM 7084	*Coriobacteriaceae*	oral cavity
Bacteroidetes		
Bacteroidia		
Anaerophaga sp. HS1	*Marinilabiliaceae*	extremophile (hot water spring)
Anaerophaga thermohalophila DSM 12881	*Marinilabiliaceae*	environmental sample (oil residue)
Bacteroides cellulosilyticus DSM 14838	*Bacteroidaceae*	gastrointestinal tract/feces
Bacteroides coprophilus DSM 18228	*Bacteroidaceae*	gastrointestinal tract/feces
Bacteroides coprosuis DSM 18011	*Bacteroidaceae*	pig feces
Bacteroides dorei DSM 17855	*Bacteroidaceae*	gastrointestinal tract/feces
Bacteroides eggerthii 1_2_48FAA	*Bacteroidaceae*	gastrointestinal tract/feces
Bacteroides faecis MAJ27	*Bacteroidaceae*	gastrointestinal tract/feces
Bacteroides fluxus YIT 12057	*Bacteroidaceae*	gastrointestinal tract/feces
Bacteroides fragilis NCTC9343	*Bacteroidaceae*	gastrointestinal tract/feces
Bacteroides nordii CL02T12C05	*Bacteroidaceae*	gastrointestinal tract/feces
Bacteroides oleiciplenus YIT 12058	*Bacteroidaceae*	gastrointestinal tract/feces
Bacteroides sp. 2_1_16	*Bacteroidaceae*	gastrointestinal tract/feces
Bacteroides sp. 203	*Bacteroidaceae*	gastrointestinal tract/feces
Bacteroides sp. 3_1_19	*Bacteroidaceae*	gastrointestinal tract/feces
Bacteroides sp. 3_1_33FAA	*Bacteroidaceae*	gastrointestinal tract/feces
Bacteroides sp. 9_1_42FAA	*Bacteroidaceae*	gastrointestinal tract/feces
Bacteroides sp. D2	*Bacteroidaceae*	gastrointestinal tract/feces
Bacteroides uniformis CL03T00C23	*Bacteroidaceae*	gastrointestinal tract/feces
Bacteroides vulgatus CL09T03C04	*Bacteroidaceae*	gastrointestinal tract/feces
Bacteroidetes oral taxon 274 str. F0058	*Bacteroidaceae*	oral cavity
Barnesiella intestinihominis YIT 11860	*Bacteroidaceae*	gastrointestinal tract/feces
Bacteroidia (continued)		

Strain[a]	Class	Isolation/habitat[b]
Marinilabilia sp. AK2	*Marinilabiliaceae*	extremophile (solar saltern)
Odoribacter laneus YIT 12061	*Porphyromonadaceae*	gastrointestinal tract/feces
Parabacteroides johnsonii DSM 18315	*Bacteroidaceae*	gastrointestinal tract/feces
Parabacteroides sp. D13	*Bacteroidaceae*	gastrointestinal tract/feces
Porphyromonas catoniae F0037	*Porphyromonadaceae*	oral cavity
Porphyromonas sp. oral taxon 279 str. F0450	*Porphyromonadaceae*	oral cavity
Prevotella bivia JCVIHMP010	*Prevotellaceae*	genital tract
Prevotella buccae ATCC 33574	*Prevotellaceae*	oral cavity
Prevotella buccalis ATCC 35310	*Prevotellaceae*	oral cavity
Prevotella denticola F0289	*Prevotellaceae*	oral cavity
Prevotella disiens FB035-09AN	*Prevotellaceae*	oral cavity
Prevotella histicola F0411	*Prevotellaceae*	oral cavity
Prevotella intermedia 17	*Prevotellaceae*	oral cavity
Prevotella melaninogenica D18	*Prevotellaceae*	oral cavity/rumen
Prevotella micans F0438	*Prevotellaceae*	oral cavity
Prevotella multiformis DSM 16608	*Prevotellaceae*	oral cavity
Prevotella nigrescens ATCC 33563	*Prevotellaceae*	oral cavity
Prevotella oralis ATCC 33269	*Prevotellaceae*	oral cavity
Prevotella oulorum F0390	*Prevotellaceae*	oral cavity
Prevotella ruminicola 23	*Prevotellaceae*	rumen
Prevotella saccharolytica F0055	*Prevotellaceae*	oral cavity
Prevotella sp. C561	*Prevotellaceae*	oral cavity
Prevotella sp. MSX73	*Prevotellaceae*	oral cavity
Prevotella sp. oral taxon 306 str. F0472	*Prevotellaceae*	oral cavity
Prevotella sp. oral taxon 317 str. F0108	*Prevotellaceae*	oral cavity
Prevotella sp. oral taxon 472 str. F0295	*Prevotellaceae*	oral cavity
Prevotella stercorea DSM 18206	*Prevotellaceae*	gastrointestinal tract/feces
Prevotella tannerae ATCC 51259	*Prevotellaceae*	oral cavity
Prevotella timonensis CRIS 5C-B1	*Prevotellaceae*	wound (breast abscess)
Prevotella veroralis F0319	*Prevotellaceae*	oral cavity
Tannerella sp. 6_1_58FAA_CT1	*Porphyromonadaceae*	gastrointestinal tract/feces
	Cytophagia	
Belliella baltica DSM 15883	*Cyclobacteriaceae*	environmental sample (groundwater)
Indibacter alkaliphilus LW1	*Cyclobacteriaceae*	extremophile (soda lake)
Nitritalea halalkaliphila LW7	*Cyclobacteriaceae*	extremophile (saline soda lake)
	Flavobacteria	
Bergeyella zoohelcum ATCC 43767	*Flavobacteriaceae*	oral cavity
Capnocytophaga canimorsus Cc5	*Flavobacteriaceae*	dog and cat oral cavity/zoonotic infections
Capnocytophaga gingivalis ATCC 33624	*Flavobacteriaceae*	oral cavity
Capnocytophaga ochracea DSM 7271	*Flavobacteriaceae*	oral cavity
Capnocytophaga sp. CM59	*Flavobacteriaceae*	oral cavity
Capnocytophaga sp. oral taxon 324 str. F0483	*Flavobacteriaceae*	oral cavity
Capnocytophaga sp. oral taxon 326 str. F0382	*Flavobacteriaceae*	oral cavity
Capnocytophaga sp. oral taxon 329 str. F0087	*Flavobacteriaceae*	oral cavity
Capnocytophaga sp. oral taxon 335 str. F0486	*Flavobacteriaceae*	oral cavity
Capnocytophaga sp. oral taxon 380 str. F0488	*Flavobacteriaceae*	oral cavity
Capnocytophaga sp. oral taxon 412 str. F0487	*Flavobacteriaceae*	oral cavity
Capnocytophaga sputigena ATCC 33612	*Flavobacteriaceae*	oral cavity
Chryseobacterium sp. CF314	*Flavobacteriaceae*	vegetation
Flavobacteriaceae bacterium S85	*Flavobacteriaceae*	environmental sample (seawater)
Flavobacterium branchiophilum FL-15	*Flavobacteriaceae*	fish pathogen

Strain[a]	Class	Isolation/habitat[b]
Flavobacterium columnare ATCC 49512	*Flavobacteriaceae*	fish pathogen
Flavobacterium psychrophilum JIP02/86	*Flavobacteriaceae*	fish pathogen
Fluviicola taffensis DSM 16823	*Cryomorphaceae*	environmental sample (fresh water)
Galbibacter sp. ck-I2-15	*Flavobacteriaceae*	extremophile (deep sea sediment)
Joostella marina DSM 19592	*Flavobacteriaceae*	environmental sample (seawater)
Kordia algicida OT-1	*Flavobacteriaceae*	environmental sample (seawater)
Myroides injenensis M09-0166	*Flavobacteriaceae*	human clinical specimens
Myroides odoratus DSM 2801	*Flavobacteriaceae*	fish
	Flavobacteria (continued)	
Ornithobacterium rhinotracheale DSM 15997	*Flavobacteriaceae*	bird respiratory tract
Psychroflexus torquis ATCC 700755	*Flavobacteriaceae*	extremophile (antarctic ice)
Riemerella anatipestifer ATCC 11845 = DSM 15868	*Flavobacteriaceae*	bird
Weeksella virosa DSM 16922	*Flavobacteriaceae*	genital tract/urine
Zunongwangia profunda SM-A87	*Flavobacteriaceae*	extremophile (deep sea sediment)
	Sphingobacteria	
Mucilaginibacter paludis DSM 18603	*Sphingobacteriaceae*	food (fermented)
Niabella soli DSM 19437	*Chitinophagaceae*	environmental sample (soil)
Sphingobacterium spiritivorum ATCC 33861	*Sphingobacteriaceae*	human clinical specimens
	Firmicutes	
	Bacilli	
Alicycliphilus denitrificans	*Alicyclobacillaceae*	environmental sample (sewage)
Alicyclobacillus hesperidum URH17-3-68	*Alicyclobacillaceae*	extremophile (hot water spring)
Bacillus cereus Rock1-15	*Bacillaceae*	environmental sample (soil)
Bacillus smithii 7 3 47FAA	*Bacillaceae*	human clinical specimens
Bacillus thuringiensis serovar *finitimus* YBT-020	*Bacillaceae*	environmental sample (soil)
Brevibacillus laterosporus GI-9	*Paenibacillaceae*	environmental sample (soil)
Catellicoccus marimammalium M35/04/3	*Enterococcaceae*	grey seal gastrointestinal tract
Dolosigranulum pigrum ATCC 51524	*Carnobacteriaceae*	human clinical specimens
Enterococcus faecalis TX0012	*Enterococcaceae*	gastrointestinal tract/feces
Enterococcus faecium 1231408	*Enterococcaceae*	gastrointestinal tract/feces
Enterococcus hirae ATCC 9790	*Enterococcaceae*	gastrointestinal tract/feces
Enterococcus italicus DSM 15952	*Enterococcaceae*	food (fermented)
Enterococcus sp. 7L76	*Enterococcaceae*	gastrointestinal tract/feces
Facklamia hominis CCUG 36813	*Aerococcaceae*	burbuncle (human)
Fructobacillus fructosus KCTC 3544	*Leuconostocaceae*	vegetation
Gemella haemolysans ATCC 10379	*Streptococcaceae*	oral cavity
Gemella moribillum M424	*Streptococcaceae*	gastrointestinal tract/feces
Lactobacillus animalis KCTC 3501	*Lactobacillaceae*	food (fermented)
Lactobacillus brevis subsp. *gravesensis* ATCC 27305	*Lactobacillaceae*	food (fermented)
Lactobacillus buchneri ATCC 11577	*Lactobacillaceae*	food (fermented)
Lactobacillus casei str. Zhang	*Lactobacillaceae*	gastrointestinal tract/feces
Lactobacillus coryniformis subsp. *coryniformis* KCTC 3167	*Lactobacillaceae*	food (fermented)
Lactobacillus coryniformis subsp. *torquens* KCTC 3535	*Lactobacillaceae*	food (fermented)
Lactobacillus crispatus FB049-03	*Lactobacillaceae*	genital tract
Lactobacillus curvatus CRL 705	*Lactobacillaceae*	food (fermented)
Lactobacillus delbrueckii subsp. bulgaricus 2038	*Lactobacillaceae*	food (fermented)

Strain[a]	Class	Isolation/habitat[b]
Lactobacillus farciminis KCTC 3681	*Lactobacillaceae*	food (fermented)
Lactobacillus fermentum ATCC 14931	*Lactobacillaceae*	food (fermented)
Lactobacillus florum 2F	*Lactobacillaceae*	vegetation
Lactobacillus gasseri JV-V03	*Lactobacillaceae*	oral cavity
Lactobacillus hominis CRBIP 24.179	*Lactobacillaceae*	gastrointestinal tract/feces
Lactobacillus iners LactinV 11V1-d	*Lactobacillaceae*	genital tract/urine
Lactobacillus jensenii 269-3	*Lactobacillaceae*	genital tract/blood
Lactobacillus johnsonii DPC 6026	*Lactobacillaceae*	pig gastrointestinal tract
Lactobacillus mucosae LM1	*Lactobacillaceae*	wild pig gastrointestinal tract
Lactobacillus paracasei subsp. *paracasei* 8700:2	*Lactobacillaceae*	food (fermented)
Lactobacillus pentosus IG1	*Lactobacillaceae*	food (fermented)
Lactobacillus plantarum ZJ316	*Lactobacillaceae*	gastrointestinal tract/feces
Lactobacillus rhamnosus GG	*Lactobacillaceae*	gastrointestinal tract/feces
Lactobacillus ruminis ATCC 25644	*Lactobacillaceae*	rumen
Lactobacillus salivarius UCC118	*Lactobacillaceae*	oral cavity
Lactobacillus sanfranciscensis TMW 1-1304	*Lactobacillaceae*	food (fermented)
Lactobacillus sp. 66c	*Lactobacillaceae*	ND
Lactobacillus versmoldensis KCTC 3814	*Lactobacillaceae*	food (fermented)
Leuconostoc gelidum KCTC 3527	*Leuconostocaceae*	food (fermented)
Leuconostoc pseudomesenteroides 4882	*Leuconostocaceae*	food (fermented)
Bacilli (continued)		
Listeria innocua Clip11262	*Listeriaceae*	environmental sample (soil)
Listeria ivanovii FSL F6-596	*Listeriaceae*	animal and human/environmental samples
Listeria monocytogenes str. 1/2a F6854	*Listeriaceae*	animal and human/environmental samples
Listeria seeligeri FSL N1-067	*Listeriaceae*	animal and human/environmental samples
Listeriaceae bacterium TTU M1-001	*Listeriaceae*	environmental sample (soil)
Oenococcus kitaharae DSM 17330	*Leuconostocaceae*	food (fermented)
Pediococcus acidilactici DSM 20284	*Lactobacillaceae*	vegetation
Pediococcus lolii NGRI 0510Q	*Lactobacillaceae*	vegetation (fermented)
Planococcus antarcticus DSM 14505	*Planococcaceae*	extremophile (antarctic)
Sporolactobacillus vineae DSM 21990 = SL153	*Sporolactobacillaceae*	environmental sample (soil)
Staphylococcus aureus subsp. *aureus*	*Staphylococcaceeae*	human clinical specimens
Staphylococcus lugdunensis M23590	*Staphylococcaceeae*	human clinical specimens
Staphylococcus massiliensis S46	*Staphylococcaceeae*	skin
Staphylococcus pseudintermedius ED99	*Staphylococcaceeae*	dog skin
Staphylococcus simulans ACS-120-V-Sch1	*Staphylococcaceeae*	genital tract
Streptococcus agalactiae 2603V/R	*Streptococcaceae*	gastrointestinal tract/feces
Streptococcus anginosus F0211	*Streptococcaceae*	oral cavity
Streptococcus bovis ATCC 700338	*Streptococcaceae*	rumen/zoonotic infections
Streptococcus canis FSL Z3-227	*Streptococcaceae*	food (fermented)
Streptococcus constellatus subsp. *constellatus* SK53	*Streptococcaceae*	human clinical specimens
Streptococcus downei F0415	*Streptococcaceae*	monkey oral cavity
Streptococcus dysgalactiae DSM 12112	*Streptococcaceae*	various animals/zoonotic infections
Streptococcus equi subsp. *zooepidemicus* MGCS10565	*Streptococcaceae*	horse respiratory tract
Streptococcus equinus ATCC 9812	*Streptococcaceae*	ruminants alimentary tract
Streptococcus gallolyticus UCN34	*Streptococcaceae*	ruminants alimentary tract
Streptococcus gordonii str. Challis substr. CH1	*Streptococcaceae*	oral cavity
Streptococcus infantarius ATCC BAA-102	*Streptococcaceae*	gastrointestinal tract/feces

Strain[a]	Class	Isolation/habitat[b]
Streptococcus iniae 9117	*Streptococcaceae*	fish/human pathogen
Streptococcus macacae NCTC 11558	*Streptococcaceae*	monkey oral cavity
Streptococcus macedonicus ACA-DC 198	*Streptococcaceae*	food (fermented)
Streptococcus mitis ATCC 6249	*Streptococcaceae*	oral cavity
Streptococcus mutans UA159	*Streptococcaceae*	oral cavity
Streptococcus oralis SK1074	*Streptococcaceae*	oral cavity
Streptococcus parasanguinis F0449	*Streptococcaceae*	oral cavity
Streptococcus pasteurianus ATCC 43144	*Streptococcaceae*	blood
Streptococcus pseudoporcinus SPIN 20026	*Streptococcaceae*	genital tract
Streptococcus pyogenes SF370	*Streptococcaceae*	oral cavity/wounds
Streptococcus ratti FA-1 = DSM 20564	*Streptococcaceae*	rat oral cavity
Streptococcus salivarius JIM8777	*Streptococcaceae*	oral cavity
Streptococcus sanguinis VMC66	*Streptococcaceae*	oral cavity
Streptococcus sp. BS35b	*Streptococcaceae*	oral cavity
Streptococcus sp. C150	*Streptococcaceae*	oral cavity (expectorated sputum)
Streptococcus sp. C300	*Streptococcaceae*	oral cavity (expectorated sputum)
Streptococcus sp. F0441	*Streptococcaceae*	oral cavity
Streptococcus sp. GMD4S	*Streptococcaceae*	oral cavity
Streptococcus sp. GMD6S	*Streptococcaceae*	oral cavity
Streptococcus sp. M334	*Streptococcaceae*	oral cavity (expectorated sputum)
Streptococcus sp. oral taxon 056 str. F0418	*Streptococcaceae*	oral cavity
Streptococcus sp. oral taxon 071 str. 73H25AP	*Streptococcaceae*	oral cavity
Streptococcus suis 89/1591	*Streptococcaceae*	pig
Streptococcus thermophilus LMD-9	*Streptococcaceae*	food (fermented)
Streptococcus vestibularis ATCC 49124	*Streptococcaceae*	oral cavity
	Clostridia	
Acidaminococcus intestini RyC-MR95	*Acidaminococcaceae*	wound/abscess
Acidaminococcus sp. D21	*Acidaminococcaceae*	gastrointestinal tract/feces
Aminomonas paucivorans DSM 12260	*Syntrophoomonadaceae*	environmental sample (sewage)
Anaerococcus tetradius ATCC 35098	*Peptostreptococcaceae*	human clinical specimens
Butyrivibrio fibrisolvens 16/4	*Lachnospiraceae*	rumen
Catenibacterium mitsuokai DSM 15897	*Lachnospiraceae*	gastrointestinal tract/feces
Clostridium cellulolyticum H10	*Clostridiaceae*	vegetation (composted)
	Clostridia (continued)	
Clostridium perfringens D str. JGS1721	*Clostridiaceae*	environmental sample (vegetation/marine sediment)
Clostridium spiroforme DSM 1552	*Clostridiaceae*	gastrointestinal tract/feces
Coprococcus catus GD/7	*Lachnospiraceae*	gastrointestinal tract/feces
Coprococcus comes ATCC 27758	*Lachnospiraceae*	gastrointestinal tract/feces
Dorea longicatena DSM 13814	*Clostridiaceae*	gastrointestinal tract/feces
Eubacterium dolichum DSM 3991	*Eubacteriaceae*	gastrointestinal tract/feces
Eubacterium rectale ATCC 33656	*Eubacteriaceae*	gastrointestinal tract/feces
Eubacterium sp. AS15	*Eubacteriaceae*	oral cavity
Eubacterium ventriosum ATCC 27560	*Eubacteriaceae*	gastrointestinal tract/feces
Eubacterium yurii subsp. margaretiae ATCC 43715	*Peptostreptococcaceae*	oral cavity
Filifactor alocis ATCC 35896	*Peptostreptococcaceae*	cat and human oral cavity
Finegoldia magna ATCC 29328	*Peptostreptococcaceae*	oral cavity
Helcococcus kunzii ATCC 51366	*Clostridiales Family XI*	wound
Oribacterium sinus F0268	*Lachnospiraceae*	human clinical specimens
Peptoniphilus duerdenii ATCC BAA-1640	*Peptostreptococcaceae*	wound
Peptoniphilus sp. oral taxon 386 str. F0131	*Peptostreptococcaceae*	oral cavity

Strain[a]	Class	Isolation/habitat[b]
Phascolarctobacterium sp. YIT 12067	*Acidaminococcaceae*	gastrointestinal tract/feces
Phascolarctobacterium succinatutens YIT 12067	*Acidaminococcaceae*	gastrointestinal tract/feces
Pseudoramibacter alactolyticus ATCC 23263	*Clostridiaceae*	oral cavity
Roseburia intestinalis L1-82	*Lachnospiraceae*	gastrointestinal tract/feces
Roseburia inulinivorans DSM 16841	*Lachnospiraceae*	gastrointestinal tract/feces
Ruminococcus albus 8	*Ruminococcaceae*	gastrointestinal tract/feces
Ruminococcus lactaris ATCC 29176	*Ruminococcaceae*	gastrointestinal tract/feces
Subdoligranulum sp. 4_3_54A2FAA	*Ruminococcaceae*	gastrointestinal tract/feces
	Negativicutes	
Megasphaera sp. UPII 135-E	*Veillonellaceae*	rumen
Veillonella atypica ACS-134-V-Col7a	*Veillonellaceae*	oral cavity
Veillonella parvula ATCC17745	*Veillonellaceae*	gastrointestinal/genital tract
Veillonella sp. 6_1_27	*Veillonellaceae*	gastrointestinal tract/feces
Veillonella sp. oral taxon 780 str. F0422	*Veillonellaceae*	oral cavity
	Proteobacteria	
	Alphaproteobacteria	
Acetobacter aceti NBRC 14818	*Acetobacteraceae*	environmental sample
Azospirillum sp. B510	*Rhodospirillaceae*	vegetation
Bradyrhizobium sp. BTAi1	*Bradyrhizobiaceae*	vegetation
Caenispirillum salinarum AK4	*Rhodospirillaceae*	extremophile (solar saltern)
Dinoroseobacter shibae DFL 12	*Rhodobacteraceae*	environmental sample (seawater)
Gluconacetobacter diazotrophicus PAl5	*Acetobacteriaceae*	vegetation
Maritimibacter alkaliphilus ATCC2654	*Rhodobacteraceae*	environmental sample (seawater)
Methylocystis sp. ATCC 49242	*Methylocystaceae*	environmental sample (sewage, fresh water)
Methylosinus trichosporium OB3b	*Methylocystaceae*	environmental sample (soil, fresh water)
Nitrobacter hamburgensis X14	*Bradyrhizobiaceae*	environmental sample (soil)
Parvibaculum lavamentivorans DS-1	*Phyllobacteriaceae*	environmental sample (sewage)
Puniceispirillum marinum IMCC1322	*SAR16 clade*	environmental sample (seawater)
Rhodopseudomonas palustris BisB18	*Bradyrhizobiaceae*	environmental sample (soil)
Rhodospirillum rubrum ATCC 11170	*Rhodospirillaceae*	environmental sample (sea mud)
Rhodovulum sp. PH10	*Rhodobacteraceae*	environmental sample (soil)
Sphingobium sp. AP49	*Sphingomonadaceae*	vegetation
Sphingomonas sp. S17	*Sphingomonadaceae*	environmental sample (stromatolite)
Tistrella mobilis KA081020-065	*Rhodospirillaceae*	environmental sample (seawater)
	Betaproteobacteria	
Acidovorax avenae subsp. *avenae* ATCC 19860	*Comamonadaceae*	environmental sample (soil)
Acidovorax ebreus TPSY	*Comamonadaceae*	environmental sample (water)
Burkholderiales bacterium 1 1 47	*Burkholderiales*	gastrointestinal tract/feces
Kingella kingae ATCC 23330	*Neisseriaceae*	oral cavity
Neisseria bacilliformis ATCC BAA-1200	*Neisseriaceae*	oral cavity
	Betaproteobacteria (continued)	
Neisseria cinerea ATCC 14685	*Neisseriaceae*	oral cavity
Neisseria flavescens SK114	*Neisseriaceae*	human clinical specimens
Neisseria lactamica 020-06	*Neisseriaceae*	oral cavity
Neisseria meningitidis A Z2491	*Neisseriaceae*	oral cavity
Neisseria mucosa C102	*Neisseriaceae*	oral cavity (expectorated sputum)
Neisseria sp. oral taxon 014 str. F0314	*Neisseriaceae*	oral cavity

Strain[a]	Class	Isolation/habitat[b]
Neisseria subflava NJ9703	*Neisseriaceae*	oral cavity
Neisseria wadsworthii 9715	*Neisseriaceae*	skin
Nitrosomonas sp. AL212	*Nitrosomonadaceae*	environmental sample (fresh water)
Parasutterella excrementihominis YIT 11859	*Alcaligenaceae*	gastrointestinal tract/feces
Ralstonia syzygii R24	*Burkholderiaceae*	environmental sample (soil)
Simonsiella muelleri ATCC 29453	*Neisseriaceae*	oral cavity
Sutterella parvirubra YIT 11816	*Alcaligenaceae*	gastrointestinal tract/feces
Sutterella wadsworthensis 3 1 45B	*Alcaligenaceae*	gastrointestinal tract/feces
Verminephrobacter aporrectodeae subsp. *tuberculatae* At4	*Comamonadaceae*	invertebrate (earthworm)
Verminephrobacter eiseniae EF01-2	*Comamonadaceae*	invertebrate (earthworm)
Gammaproteobacteria		
Actinobacillus minor NM305	*Pasteurellaceae*	pig respiratory tract
Actinobacillus pleuropneumoniae serovar 10 D13039	*Pasteurellaceae*	pig respiratory tract
Actinobacillus succinogenes 130Z	*Pasteurellaceae*	rumen
Actinobacillus suis H91-0380	*Pasteurellaceae*	pig pathogen
Actinobacillus ureae ATCC 25976	*Pasteurellaceae*	respiratory tract
Alcanivorax sp. W11-5	*Alcanivoracaceae*	extremophile (deep sea sediment)
Francisella tularensis subsp. *holarctica* LVS	*Francisellaceae*	engineered live vaccine strain
Francisella tularensis subsp. *novicida* U112	*Francisellaceae*	human/environmental sample (water)
Francisella tularensis subsp. *tularensis* WY96-3418	*Francisellaceae*	wound
gamma proteobacterium HTCC5015	*Unclassified*	environmental sample (seawater)
gammaproteobacterium HdN1	*Unclassified*	environmental sample (sewage)
Haemophilus parainfluenzae T3T1	*Pasteurellaceae*	oral cavity/genital tract
Haemophilus pittmaniae HK 85	*Pasteurellaceae*	oral cavity
Haemophilus sputorum HK 2154	*Pasteurellaceae*	oral cavity
Legionella pneumophila str. Paris	*Legionellaceae*	human clinical specimens
Pasteurella bettyae CCUG 2042	*Pasteurellaceae*	genital tract
Pasteurella multocida subsp. *gallicida* X73	*Pasteurellaceae*	bird pathogen
Pasturella multocida Pm70	*Pasteurellaceae*	bird respiratory tract/zoonotic infections
Deltaproteobacteria		
uncultured delta proteobacterium HF0070_07E19	*Unclassified*	environmental sample (seawater)
Epsilonproteobacteria		
Campylobacter coli 2962	*Campylobacteraceae*	animals/human pathogen
Campylobacter jejuni NCTC11168	*Campylobacteraceae*	bird
Campylobacter jejuni subsp. *doylei* 269-97	*Campylobacteraceae*	blood
Campylobacter lari	*Campylobacteraceae*	gastrointestinal tract/feces
Helicobacter canadensis MIT 98-5491	*Helicobacteriaceae*	gastrointestinal tract/feces
Helicobacter cinaedi CCUG 18818	*Helicobacteriaceae*	gastrointestinal tract/feces
Helicobacter hepaticus ATCC 51449	*Helicobacteriaceae*	mouse liver
Helicobacter mustelae 12198	*Helicobacteriaceae*	ferret
Helicobacter pullorum MIT 98-5489	*Helicobacteriaceae*	bird/zoonotic infections
Nitratifractor salsuginis DSM 16511	*Unclassified*	extremophile (deep sea sediment)
Wolinella succinogenes DSM 1740	*Helicobacteraceae*	rumen
Fusobacteria		
Fusobacterium nucleatum subsp. *vincentii* ATCC	*Fusobacteriaceae*	oral cavity

Strain[a]	Class	Isolation/habitat[b]
49256		
Fusobacterium sp. 1_1_41FAA	*Fusobacteriaceae*	gastrointestinal tract/feces
Fusobacterium sp. 3_1_27	*Fusobacteriaceae*	gastrointestinal tract/feces
Fusobacterium sp. 3_1_36A2	*Fusobacteriaceae*	gastrointestinal tract/feces
Ilyobacter polytropus DSM 2926	*Fusobacteriaceae*	environmental sample (sea mud)
Streptobacillus moniliformis DSM 12112	*Leptotrichiaceae*	rodent/human pathogen
	Spirochaetes	
Leptospira inadai serovar Lyme str. 10	*Leptospiraceae*	human clinical specimens
Sphaerochaeta globus str. Buddy	*Spirochaetaceae*	extremophile (marine hot spring)
Treponema denticola ATCC 35405	*Spirochaetaceae*	oral cavity
Treponema phagedenis F0421	*Spirochaetaceae*	monkey genital tracts
Treponema sp. JC4	*Spirochaetaceae*	rumen
Treponema vincentii ATCC 35580	*Spirochaetaceae*	oral cavity
	Tenericutes	
	Mollicutes	
Mycoplasma canis PG 14	*Mycoplasmataceae*	dog oral cavity
Mycoplasma cynos C142	*Mycoplasmataceae*	dog respiratory tract
Mycoplasma gallisepticum str. F	*Mycoplasmataceae*	bord pathogen
Mycoplasma iowae 695	*Mycoplasmataceae*	bird
Mycoplasma mobile 163K	*Mycoplasmataceae*	fish pathogen
Mycoplasma ovipneumoniae SC01	*Mycoplasmataceae*	goat respiratory tract
Mycoplasma synoviae 53	*Mycoplasmataceae*	bird pathogen
Solobacterium moorei F0204	*Erysipelotrichaceae*	gastrointestinal tract/feces
	Elusimicrobia	
Elusimicrobium minutum Pei191	*Elusimicrobiaceae*	invertebrate (scarab beetle)
Uncultured Termite group 1 bacterium phylotype Rs-D17	*Elusimicrobiaceae*	invertebrate
	Fibrobacteres	
Fibrobacter succinogenes S85	*Fibrobacteraceae*	rumen
	Ignavibacteria	
Ignavibacterium album JCM 16511	*Ignavibacteriaceae*	extremophile (hot water spring)
	Planktomycetes	
Blastopirellula marina DSM 3645	*Planctomycetaceae*	environmental sample (seawater)
	Verrucomicrobia	
Diplosphaera colitermitum TAV2	*Opitutaceae*	invertebrate (termite)
Akkermansia muciniphila ATCC BAA-835	*Verrucomicrobiaceae*	gastrointestinal tract/feces
	Unclassified	
candidate division TM7 single-cell isolate TM7c	*Unclassified*	oral cavity
uncultured bacterium	*Unclassified*	environmental sample (groundwater)
uncultured bacterium	*Unclassified*	environmental sample (groundwater)
uncultured bacterium T3_7_42578	*Unclassified*	invertebrate (honeybee)

[a]Single strains representing every species found to harbor the *cas9* gene are listed.

[b]The origin of the specific strain and/or typical habitat of the species are given for every strain. ND, no data available. Note that if not specified otherwise, isolates from body sites and feces are human commensals and pathogens.

Supplementary Table S5 – tracrRNA and CRISPR repeats associated with the examined type II CRISPR-Cas systems.

strain	Cas9 tree position [a]	tracrRNA sequence [b]	Repeat [c]	Number of repeats [d]	tracrRNA Length	Repeat Length [e]
Francisella novicida U112	1	AUCUAAAAUUAUAAAUGUACCAAAUAAUUAAUGCUCUG UAAUCAUUUAAAAGUAUUUUGAACGGACCUCUGUUUGA CACGUCUGAAUAAC (SEQ ID NO:2702)	GUUUCAGUUGCUGAAUAUUU GGUAAACUACUGUUAG (SEQ ID NO:2765)	14	90	37
Gamma proteobacteriuam HTCC5015	2	UCAGAAUGCAUCCCCAACACAUUCUAUACACUGAAAUCAUA GAAAAUCACGUUUGUGGCCCGACCAACUGCUUCGGCAU GUCGGGUUUUUUU (SEQ ID NO:2703)	GUUUCAGCUGUUGGUUUGUUG GGAUAAGCUCUGAAAC (SEQ ID NO:2766)	27	88	37
Parasutterella excrementihominis YIT11859	3	UUAAUUACAUUCUUUUAACAACGAAGUCGCCUUCGGGC GAGCUGAAAUCAAUUUGAUUAAAUAUUAGAUCCGGCUA CUGAGGUCUUUGACCUUAUCCGGAUUAACGAAGAGCCU CCGAGGAGGCUUUUU (SEQ ID NO:2704)	GUUUCAGUAGUUGUUAGAAGA AUGUAGUAUUGAAGCC (SEQ ID NO:2767)	>2	129	37
Sutterella wadsworthensis 3_1_45B	4	UUAGAGAUCAUAACGCUAUGAGCUAUAGGAAAUCACCU UCGGGUGAGCUGAAAUCCCCUAAAGCUAAGAUUGAAUC CGGCCACUAUCUAUUAGUAGAUAUCCGGAUAUUCU (SEQ ID NO:2705)	GUUUCAGUGCUAUAGCUCGUA GCGUUAUGAUCUUCGC (SEQ ID NO:2768)	13	111	37
Legionella pneumophila	5	UAAAUUAGAAAUCAUCUAAAUUUCGAUACCCUGAAAUC AACAAAAUUAAAGAUUGAAUCGUUUUUUCUAUGCUCGUC	GUUUCAGUGGUUGGAUUUUUA GAUGAGGGAUUAUUGG (SEQ	34	100	37

str. Paris		UUAAUAGCGAGCAUAUAACGAUUU (SEQ ID NO:2706)	ID NO:2769)			
Wolinella succinogenes DSM 1740	6	UUGUUAGAAUGUUCCCGCAACACUUUAUAGCAAAUCCG UUCGAUGCCUUGAAAUCAUCAAAAAGAUAUAAUAGACC CGCCCACUGUAUUGUACAUGGCGGGACUUUUU (SEQ ID NO:2707)	GUUUCACAGGCUAAGCGGAUU UGCUAUAAAGUGUUGC (SEQ ID NO:2770)	23	108	37
Staphylococcus pseudintermedius ED99	7	GUUUUACUUCUUUCUAAAUAAACAUAGUUAAGUUAAAA CAAGCUUAAAGCGUCAAUGUAAUAUUUUAUUAACACCC UACUGUGUCAGUGGGGUUUUUUU (SEQ ID NO:2708)	GUUUUAGCACUAUGUUUAUUU AGAAAGAGGUAAAAAC (SEQ ID NO:2771)	24	99	36
Planococcus antarctcticus DSM 14505	8	AUUUCAAAAUAUUCCCCCUUUACAUUUUUCAAAGAAA AUGUACGCUAAGAGUGUUACUACUCUGUAACAUUACAU UGGUACGUUAAAAUAAGCUUAAAGCGUAAAAGUUGGCC CUAUGAGGUCUCCCGCCAUCGACUUCGUCGGUGGCUUUU UU (SEQ ID NO:2709)	GUUUUAGACCAAUGUAAUUUU AGAGAGUAGUAAAAC (SEQ ID NO:2772)	>5	154	36
Streptococcus sanguinis SK49	9	AACUACGUUGGAACUAUUCGAAACAACACAGCCAAAAG AUUUUUCUUUUGAGUUAAAAAUAUGGUUAUCCAUAAUCA GUUAUGCGCACCGAUUCGGUGCUUUUUU (SEQ ID NO:2710)	GUUUUAGAGCUGUGUUGUUUC GAAUGGUUCCAAAAC (SEQ ID NO:2773)	11	104	36
Listeria innocua Clip11262	10	AUUGUUAGUAUUCAAAAUAACAUAGCAAGUUAAAAUAA GGCUUUGUCCGUUAUCAACUUUUAAUUAAGUAGCGCUG UUUCGGCGCUUUUU (SEQ ID NO:2711)	GUUUUAGAGCUAUGUUAUUUU GAAUGCUAACAAAAC (SEQ ID NO:2774)	11	90	36
Streptococcus pyrogenes SF370 (M1 GAS)	10	GUUGGAACCAUUCAAAACAGCAUAGCAAGUUAAAAUAA GGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGU CGGUGCUUUUUUU (SEQ ID NO:2712)	GUUUUAGAGCUAUGCUGUUUU GAAUGGUCCCCAAAAC (SEQ ID NO:2775)	7	89	36

Streptococcus thermophilus LMD-9	11	UUGUGGUUUGAAACCAUUCGAAACAACACAGCGAGUUAAAAUAAGGCUUAGUCCGUACUCAACUUGAAAAGGUGGCACCGAUUCCGUGUUUUUUUU (SEQ ID NO:2713)	9	96	36	GUUUUAGAGCUGUGUUGUUUCGAAUGGUUCCAAAAC (SEQ ID NO:2776)
Streptococcus mutans UA159	12	GUUGGAAUCAUUCGAAACACACAGCAAGUUAAAAUAAGGCAGUGAUUUUUAAUCCAGUCCGUACACAACUUGAAAAAGUGCGCACCGAUUCGGUGUCUUUUU (SEQ ID NO:2714)	6	102	36	GUUUUAGAGCUGUGUUGUUUCGAAUGGUUCCAAAAC (SEQ ID NO:2777)
Coriobacterium glomerans PW2	13	CGUCUUGAUUACCAGUCAGGACAGCACUGCGAGUCAAAAUACGGCUUUGCCAAACUUGCCUCCCCUUCGGAGGCGUCUCGUAGGAGACAAUUUGAAGCCCCUUUAGGGGCUUCAUUUUUCU (SEQ ID NO:2715)	10	120	36	GUUUUGGAGCAGUGUCGUUCUGACUGGUAAUCCAAC (SEQ ID NO:2778)
Lactobacillus farciminis KCTC 3681	14	GUUUUACUAUUUCUAGAUUCUUUAAGAUCUACAAAAAUAAGGAUUUAUUCCGAAUUUACCACCUAUUUUAAUUAAUAGGUGUUUUUUU (SEQ ID NO:2716)	9	89	36	GUUUUUGUACCUUAAAGAAUCUAGAAAUAGUAAAAC (SEQ ID NO:2779)
Catenibacterium mitsuokai DSM 15897	15	Too short contig	>4	-	36	GUUUUAGGGUUAUGUUAUUUUGAACUGAAUUAAAAC (SEQ ID NO:2780)
Lactobacillus rhamnasus GG	16	UGUUGAGACGACAUCCUCAACAACUUGAAUUGAUUGAUCUGACAUCUACGAGUUGAGAUCAAACAAAGCUUCAGCUGAGUUUCAAUUUCUGAGCCCAUGUUGGGCCAUACAUAUGCCACCCGAGUGCAAAUCGGGUGGCUUUUUUUU (SEQ ID NO:2718)	25	146	36	GUCUCAGGUAGAUGUCAGAUCAAUCAUAAGAGGC (SEQ ID NO:2781)
Bifidobacterium	17	GGAUUGUUUGGUCGCAAUCCAUGAUCAAGGUCAUUGACCUGACAGGCAUAAAUUGAAAUAAAGCAAGGUUUCGACC	45	144	36	GUUUCAGAUGCCUGUCAGAUCAAUGACUUUGACCAC (SEQ

bifidum S17		AAGCUUCAGAAGGUUUUAUACCUGGCCUUAUGGCUGUG AGGCUCCCGAUAAUGUCGGGAGCCUCUUUU (SEQ ID NO:2719)	ID NO:2782)			
Oenococcus kitaharae DSM 17330	18	UUGGGAUUGAUCAUCCCAAACAUCAUUGGGUUCUACCU CAUUGAUCUGACACACAGCAUUGAAGUAAAGCAAGAUU AAUUUCAAGCUUAAUUUUCUUCACAUUUUAUGUGCAGA AGGGCUUAUGCCCACAAUACAUAAAAAGUCCGCAUUCA CUUGCGGACUUUUAU (SEQ ID NO:2720)	GCUUCAGAUGUGUGUCAGAUC AAUGAGGUAGAACCC (SEQ ID NO:2783)	58	167	36
Fructibacillus fructosus KCTC 3544	19	CAUGGUUAGCUACCAUACAAGCAAGAAUUGUUUAGCUA ACUAUUCUUGCUAGGAAGAACCCAUUAAUCUGACAUAC AGGGUUAAAGUAACGCAAGGGCUUCAGCCCAAGCUUCA UGAACUUUUAAAAAGUUGGCCUUAUGGCCUUUUUUU (SEQ ID NO:2721)	GCUUUAGAUGUAUGUCGGAUU AAUGGGGUUUCUUCC (SEQ ID NO:2784)	>2	149	36
Finegoldia magna ATCC 29328	20	UAAUCUCAGUUUAAUACCUAUAUGAGAUUACAUCAUGA GUUCAAAUAAAAGUUUACUCAAAUCGCCCGAAAGAGCC CACAUUGGUGGACUAAACAAAUCUUCGGAUUUGUUUUU UU (SEQ ID NO:2722)	GUUUGAGAAUGAUGUAAUUUC AUAUAGGUAUUAAAC (SEQ ID NO:2785)	15	116	36
Viellonella atypical ACS-134-V-Col7a	21	AUAAGUAAUCCAAUUAGUUUUGGAGGUUUACAGAAUUA CACUACGAGUUCAAAUACAAAUUUAUUUACAAUGCCUU CGGGCCACCCGACGUAGGGUAUCAUCUCAAUUCUUCUG AAUUGGGAUUUUUUUU (SEQ ID NO:2723)	GUUUGCGAGUAGUGUAAUUCU GUAAAUCUCUAAAAC (SEQ ID NO:2786)	33	129	36
Solobacterium moorei F0204	22	GAAAUUGUCUUAUACCAGUAAGAUAAUUUACAUAGUAA GUUCAAACAAGCUUUUAGCGAAAUUACCGCUUUGCGGA UUCACAUUGUGUGAAGUUAACUCUCGAAAGAGAGUUUU UUCUUU (SEQ ID NO:2724)	CUUUGAGAACUAUGUAAAUUA UGCUGGUAGCAAAAC (SEQ ID NO:2787)	>2	120	36
Acidaminococcus sp.	23	none	GUUUGAGAGAUAUGUAAAUUC AAAGGAUAAUCAAAC (SEQ	40	-	36

D21		AUAUCAUUAUCAUUGAUUUACAAGGUGAGUUCAAACAA GGAUUUAUCCGUAAUUGAUUGCUCGCAUUGUGCGACAU UUUCUUAUGUAAAUCGUGAAGUCGGACUUUCGACUUCU UUUUUUU (SEQ ID NO:2725)	ID NO:2788)			
Eubacterium yurii subsp. Margaretiae ATCC 43715	24		GUUUGAGAACCUUGUAAAUCA AUAAGUAUGUAAAAC (SEQ ID NO:2789)	17	121	36
Coprococcus catus GD/7	25	none	GUUUGAGAAUGAUGUAAAAAU GUAUGGUACUCAAGC (SEQ ID NO:2790)	16	-	36
Fusobacterium nucleotum subsp. Vincentii ATCC 49256	26	none	GUUUGAGAGUAAUGUUAUUUU AAAUAGAUUCAAAAC (SEQ ID NO:2791)	>3	-	36
Filifactor alocis ATCC 35896	27	GUUGACUACCAUAUGAGAUUACACUACACGGUUCAAAU AAAGAAUUUUUCUAAUCGCCCAAUGGGCCCAUAUUGAU AUGGAUGAAACUCGCUUAGCGAGUUUUUUU (SEQ ID NO:2726)	GUUUGAGAGUAGUGUAAUUUC AUAUGGUAGUCAAAC (SEQ ID NO:2792)	26	106	36
Peptoniphilus duerdenii ATCC BAA-1640	28	none	GUUUGAGAGUUAUGUAAUUUC AUAUAGGACUAAAAC (SEQ ID NO:2793)	30	-	36
Treponema denticola ATCC 35405	29	AUUUAAGAUCCAUCUUAAAUUACACAACGAGUUCAAAU AAGAAUUCAUCAAAAUCGUCCCUUUUGGGACCGCUCAU UGUGGAGCAUCAAGGCUUAACAUGGUUAAGCCUUUUUU	GUUUGAGAGUUGUGUAAUUUA AGAUGGAUCUCAAAC (SEQ ID NO:2794)	58	115	36

		U (SEQ ID NO:2727)			
Staphylococcus lugenensis M23590	30	GUACUUAUACCUAAAAUUACAGAAUCUACUGAAACAAG ACAAUAUGUCGUGUUUAUCCCAUCAAUUUAUUGGUGGG AUUUUUUU (SEQ ID NO:)	GUUUUAGUACUCUGUAAUUUU AGGUAUAAGUGAUAC (SEQ ID NO:2795)	3	84 36
Eubacteriuam dolichum DSM 2991	31	UGAUCAUAAUCUAGCAAAAGUUUAUAUGAUCUAACAAA ACAAGGGUUUAUCCCGGAAUCAAGUUCCAAGUAUAUGC UUGGAGCUUUUUCUUU (SEQ ID NO:2728)	GUUUUGUUACCAUAUGGAUUU UUGCUAGAUUAAGAC (SEQ ID NO:2796)	16	92 36
Streptococcus thermophilus LMD-9	32	UGUAAGGGACGCCUUACACAGUUACUUAAAUCUUGCAG AAGCUACAAAGAUAAGGCUUCAUGCCGAAAUCAACACC CUGUCAUUUUAUGGCAGGGUGUUUUCGUUAUUU (SEQ ID NO:2729)	GUUUUUGUACUCUCAAGAUUU AAGUAACUGUACAAC (SEQ ID NO:2797)	17	109 36
Enterococcus faecalis TX0012	33	UGUAGUCGACGGACUACCGUGUUUGACGAAACACGUCU UUAAUAAUUUUACUGAUAAGAAAUUAUUGAGAAUCUAC AAAAAUAAGGCAUCUUGCCGAAUUUACCGCCCUACAUA UGUAGGGCGGUUUUUU (SEQ ID NO:2730)	GUUUUUGUACUCUCAAUAAUU UCUUAUCAGUAAAAC (SEQ ID NO:2798)	3	130 36
Eubacterium rectale ATCC 33656	34	UGUAGAGAAAAUUUUAUAGUCACGUAAAAUUUUUCAGAU CUACUAAAACAAGGCUUUAUGCCGAAAUCAGGAGCACC GACGGGUGCUCCUUUUUUU (SEQ ID NO:2731)	AUUUUAGUAAACUGAAUAAUUU ACGUGACUGUAAAAC (SEQ ID NO:2799)	45	95 36
Mycoplama mobile 163K	35	UGUAUUUCGAAAUACAGAUGUACAGUUAAGAAUACAUA AGAAUGAUACACUCACUAAAAAAAGGCUUUAUGCCGUAA CUACUACUUAUUUUCAAAAUAAGUAGUUUUUUUU (SEQ ID NO:2732)	GUUUUGGUGUAGUAUCAUUCU UAUGUAUUCUUAAAC (SEQ ID NO:2800)	64	110 36
Mycoplasma ovipneumoniae SC01	36	CUAGUAAGAAAUUGUCGCACAAAAAUAAGACGCAUUAU GCUGUCGAAUUUCCCCACCUAGUGGGGUUUUUUU (SEQ ID NO:2733)	GUUUUUGUGCUGUACAAUUUC UUACUAGAGUAAAAC (SEQ ID NO:2801)	>14	72 36

Mycoplasma gallisepticum str. F	37	AUUAUUGCUUACACAAUUAUUGUCGUGCUAAAAUAAGGCGCUGUUAAUGCAGCUGCCGCAUCCGCCAGAGCAUUUAUGCUCUGGCUUUUUUU (SEQ ID NO:2734)	GUUUUAGCACUGUACAAUACUUGUGUAAGCAAUAAC (SEQ ID NO:2802)	40	92	36
Mycoplasma synoviae 53	38	UAUAUAUUACUUACUUAACAAAAUAAUUGUACGAUUCCAAAAUAAGGCGCUUAUGUAAGAUGCAAUAAUGCACUUAUAUAAGCUGCCGUAAACGCCGAGGUAACUCGGUUUUUU (SEQ ID NO:2735)	GUUUUGGGGUUGUACAAUUAUUUUGUUAAGUAAAAC (SEQ ID NO:2803)	12	115	36
Mycoplasma canis PG 14	39	AGUACAAAUUAAUUAUUGUUUACCCAAAUAUUGUACACCUAAAUCAAGGCGCUUAAUUGCUGCCGUAAUUGCUGAAAGCGUAGCUUUCAGUUUUUUUU (SEQ ID NO:2736)	GUUUUAGUGUUGUACAAUAUUUGGGUAAACAAUAAC (SEQ ID NO:2804)	11	98	36
Wallinella succiogenes DSM 1740	40	CGUCGUCGCUGCCGCGAAAUGGCUGAGUAGGCAGCGGCUAUAAUAAGGGGUGUGGAGGCAUCCUGCGAAGUUCUACUCUACGGAGUAUCUUCU (SEQ ID NO:2737)	GUUAUAGCCGCCUACUCAGCCAUUCCUCGCUAUAAU (SEQ ID NO:2805)	10	92	36
Campylobacter jejuni subsp. jejuni NCTC 1168	41	AAGAAAUUUAAAAAGGGACUAAAAUAAAGAGUUUGCGGGACUCUGCGGGUUACAAUCCCCUAAAACCGCUUU (SEQ ID NO:2738)	GUUUUAGUCCCUUUUUAAAUUUCUUUAUGGUAAAAU (SEQ ID NO:2806)	5	73	36
Heliobacter mustelae 12198	42	none	GUUUUAGCCACUUCAUAAAUAUGUUUAUGCUAAAAU (SEQ ID NO:2807)	11	-	36
Methylosinus trichosporium OB3b	43	UAUGGGAAAUCGGAAGGGAAGCCACGGCAAGGUGGUUUCAUAGAAAUCACUGAAGGAUUACCCUCGUCACAGAAAUGUGGCGGGGGGAUUCCUAUU (SEQ ID NO:2739)	GCCGUGGCUUCCCUGCCGAUUUCCCUGUGGUAGGCU (SEQ ID NO:2808)	24	96	36

Ilyobacter polytopus DSM 2926	44	none	GUUGUACUUCCCUAAUUAUUU UAGCUAUGUUACAAU (SEQ ID NO:2809)	23	-	36
Bacillus smithii 7_3_47FAA	45	UAAGAUCAUAUCACAGCAAUGAUCUUAGGGUUACUAUG AUAAGGGCUUUCUACUUUAGGGGUAGAGAUGUCCCGCG GCGUUGGGGAUCGCCUAUUGCCCUUAAAGGGCACUCCC CAUUUUAAUUU (SEQ ID NO:2740)	GUCAUAGUUCCCCUAAGAUUA UUGCUGUGAUAUGAU (SEQ ID NO:2810)	28	125	36
Clostridium perfringens D str.JGS1721	45	UUAAUCAGGAACUAGGUAUAGCAUAUCGAGAGUUUAAC UAGUUACUAUAACAAGGCAUUAAGCCGUAAAGUAUCCC CUAUGUUCAUUUGAACCUAGGGGUAUCUUUU (SEQ ID NO:2741)	GUUAUAGUUCCUAGUAAAUUC UCGAUAUGCUAUAAU (SEQ ID NO:2811)	27	107	36
Clostridium cellulolyticum H10	46	AUGGCAUAUCGGAGCCUGAAUUGUUGCUAUAAUAAGGU GCUGGGUUUAGCCCAGACCGCCAAGUUAACCCCGGCAU UUAUUGCUGGGGUAUCUUGUUUUU (SEQ ID NO:2742)	GUUAUAGCUCCAAUUCAGGCU CCGAUAUGCUAUAAU (SEQ ID NO:2812)	9	100	36
Acidovorax ebreus TPSY	47	CGAUUGUGGUUAUCCGGGGUGAGAGCCGUUGCUGCAAU AAGGAGGGGUCGCAAGACCCCGUCCGUACCCAAAGCC UGGCAGGGAAACCUGUCAGGCUUUUUU (SEQ ID NO:2743)	GUUGUAGCUCCCUCUCUCACC CCGGAUAGCUACACU (SEQ ID NO:2813)	15	103	36
Neisseria meningitides Z2491	48	ACAUAUUGUCGCACUGCGAAAUGAGAACCGUUGCUACA AUAAGGCCGUCUGAAAAGAUGUGCCGCAACGCUCUGCC CCUUAAAGCUUCUGCUUUAAGGGGCAUCGUUUAUU (SEQ ID NO:2744)	GUUGUAGCUCCCUUUCUCAUU UCGCAGUGCUACAAU (SEQ ID NO:2814)	17	111	36
Pasteurella multocida str. Pm70	49	GCAUAUUGUUGCACUGCGAAAUGAGAGACGUUGCUACA AUAAGGCUUCUGAAAAGAAUGACCGUAACGCUCUGCCC CUUGUGAUUCUUAAUUGCAAGGGGCAUCGUUUUU (SEQ ID NO:2745)	GUUGUAGUUCCCUCUCUCAUU UCGCAGUGCUACAAU (SEQ ID NO:2815)	6	110	36

Aminomonas paucivorans DSM 12260	50	none	GUCAUAGCUCCCUGCCGCACU CCGAAAUGCUAUGCU (SEQ ID NO:2816)	7	-	36
Roseburia intestinalis L1-82	52	CUAAGAGAAUUAUAUCAUACCAAGUGAUAAUUAGGUUA UUACAAUAAGGUAAGAAACCUAAAAGCUCUAAUCCCAU UCUUCGGAAUGGGAUUAUCUUUU (SEQ ID NO:2746)	GUUGUAAUUCCCUGUUAUCAC UUGGUAUGGUAUAAU (SEQ ID NO:2817)	62	99	36
Lactobacillus corniformis subsp. Torquens KCTC 3535	51	No contig information		-	-	-
Alicyclobacillus hesperidum URH17-3-68	53	GCGAGGGAUAUCAUACCACAUCAAGGCUUGCGAGGUUG CUAUGAUAAGGCAACAGGCCGCAAAGCACUGACCCGCA UUCCAAUGAAUGCGGGUCAUCUACUUUUU (SEQ ID NO:2747)	GUCAUAGUUCCCUCACAAGCC UCGAUGUGGUAUGAU (SEQ ID NO:2818)	54	105	36
Roseburia inulinivorans DSM 16841	52	None		-	-	-
Uncult.delta proteobact. HF0070_07E19	54	none	GUCCUAGUUUCCCUUCCAAUC AAAGCCUGCUACACU (SEQ ID NO:2819)	8	-	36
Caenispiril	58	AUCACAGGGUGCCAUUACCAGAGAUGGUAGCACGGUGG	GUCCUGUAGCCCGGUCCGUUC	11	-	36

lum salinarum AK4		AACGGACCGGCACCCUACAGGACAAGUGAUCAUACACGU GACAGCCGCCUCCCCCGCCCCAGUGGCCAAGGGGAGGC GGCUUUUCU (SEQ ID NO:2748)	CACCGUGCUAGCUUC (SEQ ID NO:2820)				
Ruminococcus albus 8	55	Too short contig		-	-	-	
Trepanema sp. JC4	56	Too short cotig		-	-	-	
Alcanivorax sp. W11-5	57	none		-	-	-	
Rhodospirillum rubrum ATCC 11170	59	none	GUUCCAUGGCCCCGUCCCACA CCGCCAUGGUAGAGU (SEQ ID NO:2821)	8	-	36	
Ralstonia syzygii R24	60	GACUUUCCAGCAGAUCGGGAAUUGCGCUUUGCUACUAA CAAGCUGAAUCCGUUAGGAGUAAAUGCACCAAAUGAGA GGGCCGGCUUUUGCCGGCCCUUUGCUUUU (SEQ ID NO:2749)	GUUGUAGCCAGAGCGCAAUUC CCGAUCUGCUAACCU (SEQ ID NO:2822)	35	105	36	
Rhodovulum sp. PH10	61	CGUCUAGCAAGGAACGCGGCGUGGCCUCUCCGUUAACA AGGCACAUGCCACCAGAUCGAGGCGGGCUCCGGUCCGC CUCUUUGCUUU (SEQ ID NO:2750)	GUUGCGGUUGGGCCGCGCCGC GUUCCCUGCUAGACC (SEQ ID NO:2823)	12	87	36	
Alicycliphilus denitrificans K601	62	CACUCAGUUCACUGGGAUAUGCGCUCUGACCGCUAACA AGCUGAAAGAUGCACCAAAUGGAAAGCCCCGCAUGCGG GGCUUUCGUCUUUU (SEQ ID NO:2751)	GUUCCGGCCAGUGCGCAUAUC CCAGUGAUCUAGAAU (SEQ ID NO:2824)	75	90	36	
Cand. Puniceispirillum	63	GCCAUUAAUAAUUGAUUGCAAUAACACUCUGGUGAUUG AGGAGCCUAUGGUUAACAAGUGGGUUUCCUGCACAAAU CUAAGAGCUGCCUCCGGGCGGCUCUUUUGCUUU (SEQ	GUUGCUCUAGGCUCUCCAAUCA CCAGAGUGCUAUACU (SEQ	27	109	36	

marinum IMCC1322		ID NO:2752)	ID NO:2825)			
Azospirillum sp. B510	64	UGGAAAUUCGAUGGGGAUCGGGGGUCCAGCCGUUAACA UGUUCCCUUCGGGGAGCACGAAAUGCGGGGCGGGCCAC GGUCCGCCCCUUUUUUU (SEQ ID NO:2753)	GUUGCGGCUGGACCCCCGAUC CCCAUCGGCUACACU (SEQ ID NO:2826)	25	93	36
Dinoroseobacter shibae DFL 12	67	GUUCAGAAUUCGCGGUCGAGCCGUUAACAAGCUCGAAG AAGCACCACAUUAAAACGCGUCCUGCGGGGCGCGUUUU CUUUUUU (SEQ ID NO:2754)	GUUGCGGCUGGACCCCGAAUU CUGAACAGCUAAACU (SEQ ID NO:2827)	19		36
Nitrobacter hamburgensis X14	65	None		-	-	-
Bradyrhizobium sp. BTAi1	66	none		-	-	-
Parvibaculm lavamentivorans DS-1	68	UAGCAAAUCGAGAGGCGGUCGCUUUUCGCAAGCAAAUU GACCCCUUGUGCGGGCUCGGCAUCCCAAGGUCAGCUGC CGGUUAUUAUCGAAAAGGCCCACCGCAAGCAGCGCGUG GGCCUUUUUUU (SEQ ID NO:2755)	GCUGCGGAUUGCGGCCGUCUC UCGAUUUGCUACUCU (SEQ ID NO:2828)	49	125	36
Bacteroides sp. 20_3	70	None	GUUGUGAUUUGUUUUUAAAUU AGUAUCUUUGAUCCAUUGGAA ACAGC (SEQ ID NO:2829)	10	-	47
Bergeyella zoohelcum ATCC 43767	69	Too short contig		-	-	-

Ignavibacterium album JCM 16511	71	UUCUGUCCCAUUUGUUGUGAUUUGCUUUUGCACAGCAUCCUUUGGACAACUUGUUCUUUGAGGAUAUUAAAACCAACCUAUCUGUUUAAGAUAGUCAAUAUCUUUUUU (SEQ ID NO:2756)	GUUGGUUUAAUAUCCUAAAGAACAAGUUGAAAGCAAAUCACAAC (SEQ ID NO:2830)	9	107	45
Bacteroides fragilis NCTC 9343	72	none	GUUGUGAUUUGCUUUCAAAUUAGUAUCUUUGAACCAUUGGAAACAGCG (SEQ ID NO:2831)	28	4	48
Barnesiella intestinihominis YIT 11860	74	UUAAUGUGUAAAUAUAAAAAAAUUAUCGAAAAAUACAAUAGUAUUAAAAAAUUAUAUGUAUAUUUGUCAACACAAAUUUGAAAGCAAAUCACAAUAAGGAUUAUUCCGUUGUGAAAACAUUUGGAAGGGGGAGUAUUUAUACUCCUCGUUCUUUUUU (SEQ ID NO:2757)	GUUGUGAUUCGCUUUCAAAUUUGUAUCUUUGACAUAUUAAAUACAGC (SEQ ID NO:2832)	>2	158	47
Porphyromonas sp. Oral taxon 279 str. F0450	73	none		-	-	-
Odoribacter laneus YIT 12061	75	CGUUGAUUAAACAAAUCAAUUUUUUACACUCUUAUCACAGCAAGGCUAUAUGCCCGAAGGAUGUAAUCCUAUACUCUCCCGCUUCGGUGGGAGUUUUUU (SEQ ID NO:2758)	GCUGUGAUUUGAUGUAAAUACUUGAUAAGAUAUACC (SEQ ID NO:2833)	>3	94	36
Flavobacterium branchiophilum FL-15	76	AUGUUUUAUAUAUUUGCAGCAUGAUUAAUAUUUCUAAUCUUUAAUCUUAUCACAAUAAGGCUAUAUGCCGUAGAUGAAAAUCUUUAGUCCUGCUUCGGUGGGACUUUUUUUU (SEQ ID NO:2759)	GUUGUGGUUUGAUUAAAGAUUAGAAACACGAUAUG (SEQ ID NO:2834)	29	112	36
Prevotella sp. C561	77	GUUUGUUUUAUCAGAAAUAAGUUGUAUAUUUGCACUCAGAUACACAGUGAAGACUUUUCACAACAAGGCUAUAAGCCGAAGAUUUUCUUGUACCCUGCGGUCAACCACAGGGUC	GUUGUACGUGCUAAUGCAAAGAUACACAUUUUGAAGCAAAUCACAAC (SEQ ID	>3	121	47

		UUUUUUU (SEQ ID NO:2760)	NO:2835)			
Prevotella timonensis CRIS 5C-B1	78	None	GUUGUGGUUUGAUGUAGAAUC AAAAUAUGAAGCAAC (SEQ ID NO:2836)	>3	-	36
Elusimicrobium minutum Pei191	79	none	GUUAGGGUUGCCCUCCGAGAA UUGAUUUUAUAGAAU (SEQ ID NO: 2837)	15	-	36
Sphaerochaeta globus str. Buddy	80	GUUAUAUCUUAACAAAAACCAGCGAUUAUCUCUAAUAA GACUUAAGUCGCAAAAUGCUCCCUAUUUUGGGAGCUUU UUUU (SEQ ID NO:2761)	GUUGGGGAUGACCGCUGAUUU UUGUUAAGAUUGACC (SEQ ID NO:2838)	43	80	36
Acidothermus cellulyticus 11B	81	GAGACAGGCUACCUAGCAAGACCCCUUCGUGGGGUCGC AUUCUUCACCCCCUCGCAGCGAGGGGGUUCGUUU (SEQ ID NO:2762)	GCUGGGGAGCCUGUCUCAAUC CCCCGGCUAAAAUGG (SEQ ID NO:2839)	24	75	36
Actinomyces sp. Oral taxon 180 str. F0310	82	None	GCUGGGAAUCAAUCACCACUC CCCUUUGAUAUACUG (SEQ ID NO:2840)	21	-	36
Bifidobacterium longum DJO10A	84	none	GCUGGGAAUUAGCAUUCACCC UUCUUGAUAAGCUUG (SEQ ID NO:2841)	43	-	36
Akkermansia mciniphila ATCC BAA-835	85	ACAAAACACUCUGAACACUCACUUUAACUCCCAACGGAUU CAAGACAAAAUUUGAAAUGCAAACCGAUUUUCCUGACU GCCAGCCAGUCACACCGGUAACAAAAGCAUUUU (SEQ ID NO:2763)	GUUUUGCCUUGAAUCCAAAAU AAGGCACAGUACAAC (SEQ ID NO:2842)	12	109	36

Nitratifractor salsuginis DSM 16511	86	GUUGUAACAGGGUAGGGUUUUUUGAGGGGUCUUAAAAU CAAGAACUGUUACAACAGUUCCAUUCUAGGGCCCAUCU UCGGACGGGCCUCAGCCUUUUUUU (SEQ ID NO:2764)	GUUUUAAGACCCCUCAAAACC CCACCCUGUUACAAU (SEQ ID NO:2843)	12	100	36

[a] The position of the strain on the Cas9 tree is given. Color shading corresponds to the color branch of the tree.

[b] Predicted or previously validated tracrRNA sequence is given. none, no tracrRNA was found; too short contig, the type II CRISPR-Cas locus is at the end of the genomic sequence contig and it was not possible to identify a tracrRNA ortholog; no contig information, genomic sequence contig encoding a type II CRISPR-Cas locus was not available.

[c] Predicted or previously validated CRISPR repeat sequence is given. none, no repeat-spacer array was found; too short contig, the type III CRISPR-Cas locus is at the end of the genomic sequence contig and it was not possible to identify a repeat-spacer array; no contig information, genomic sequence contig encoding a type II CRISPR-Cas locus was not available.

[d] Amount of the CRISPR repeats of the repeat-spacer array is given. Values preceded by ">" indicate a minimal amount of repeats in the array given that the array is at the end of the genomic sequence contig.

[e] The length of the CRISPR repeats is given. Values higher than the typical 36 nt are highlighted.

Supplemental Table S6

Strain	Cas9 GI	Cluster
Acidaminococcus intestini RyC-MR95	352684361	1
Acidaminococcus sp. D21	227824983	1
Anaerococcus tetradius ATCC 35098	227501312	1
Bifidobacterium bifidum S17	310286728	1
Catenibacterium mitsuokai DSM 15897	224543312	1
Coprococcus catus GD/7	291520705	1
Coriobacterium glomerans PW2	328956315	1
Dolosigranulum pigrum ATCC 51524	375088882	1
Dorea longicatena DSM 13814	153855454	1
Eggerthella sp. YY7918	339445983	1
Enterococcus faecalis ATCC 29200	229548613	1
Enterococcus faecalis ATCC 4200	256617555	1
Enterococcus faecalis D6	257086028	1
Enterococcus faecalis E1Sol	257080914	1
Enterococcus faecalis OG1RF	384512368	1
Enterococcus faecalis TX0470	312900261	1
Enterococcus faecalis TX4244	422695652	1
Enterococcus faecium 1,141,733	257888853	1
Enterococcus faecium 1,231,408	257893735	1
Enterococcus faecium E1133	430847551	1
Enterococcus faecium E3083	431757680	1
Enterococcus faecium PC4.1	293379700	1
Enterococcus faecium TX1330	227550972	1
Enterococcus faecium TX1337RF	424765774	1
Enterococcus hirae ATCC 9790	392988474	1
Enterococcus italicus DSM 15952	315641599	1
Eubacterium sp. AS15	402309258	1
Eubacterium yurii subsp. margaretiae ATCC 43715	306821691	1
Filifactor alocis ATCC 35896	374307738	1
Finegoldia magna ACS-171-V-Col3	302380288	1
Finegoldia magna ATCC 29328	169823755	1
Finegoldia magna SY403409CC001050417	417926052	1
Fructobacillus fructosus KCTC 3544	339625081	1
Fusobacterium nucleatum subsp. vincentii ATCC 49256	34762592	1
Fusobacterium sp. 1_1_41FAA	294782278	1

Fusobacterium sp. 3_1_27	294785695	1
Fusobacterium sp. 3_1_36A2	256845019	1
Gemella haemolysans ATCC 10379	241889924	1
Gemella morbillorum M424	317495358	1
Gordonibacter pamelaeae 7-10-1-b	295106015	1
Helcococcus kunzii ATCC 51366	375092427	1
Lactobacillus animalis KCTC 3501	335357451	1
Lactobacillus brevis subsp. gravesensis ATCC 27305	227509761	1
Lactobacillus buchneri CD034	406027703	1
Lactobacillus buchneri NRRL B-30929	331702228	1
Lactobacillus casei BL23	191639137	1
Lactobacillus casei Lc-10	418010298	1
Lactobacillus casei M36	417996992	1
Lactobacillus casei str. Zhang	301067199	1
Lactobacillus casei T71499	417999832	1
Lactobacillus casei UCD174	418002962	1
Lactobacillus casei W56	409997999	1
Lactobacillus coryniformis subsp. coryniformis KCTC 3167	333394446	1
Lactobacillus curvatus CRL 705	354808135	1
Lactobacillus farciminis KCTC 3681	336394882	1
Lactobacillus fermentum 28-3-CHN	260662220	1
Lactobacillus fermentum ATCC 14931	227514633	1
Lactobacillus florum 2F	408790128	1
Lactobacillus gasseri JV-V03	300361537	1
Lactobacillus hominis CRBIP 24.179	395244248	1
Lactobacillus iners LactinV 11V1-d	309803917	1
Lactobacillus jensenii 269-3	238854567	1
Lactobacillus jensenii 27-2-CHN	256852176	1
Lactobacillus johnsonii DPC 6026	385826041	1
Lactobacillus mucosae LM1	377831443	1
Lactobacillus paracasei subsp. paracasei 8700:2	239630053	1
Lactobacillus pentosus IG1	339637353	1
Lactobacillus pentosus KCA1	392947436	1
Lactobacillus pentosus MP-10	334881121	1
Lactobacillus plantarum ZJ316	448819853	1
Lactobacillus rhamnosus GG	258509199	1
Lactobacillus rhamnosus HN001	199597394	1
Lactobacillus rhamnosus R0011	418072660	1
Lactobacillus ruminis ATCC 25644	323340068	1

Lactobacillus salivarius SMXD51	418960525	1
Lactobacillus sanfranciscensis TMW 1.1304	347534532	1
Lactobacillus sp. 66c	408410332	1
Lactobacillus versmoldensis KCTC 3814	365906066	1
Leuconostoc gelidum KCTC 3527	333398273	1
Listeria innocua ATCC 33091	423101383	1
Listeria innocua Clip11262	16801805	1
Listeria innocua FSL S4-378	422414122	1
Listeria ivanovii FSL F6-596	315305353	1
Listeria monocytagenes 10403S	386044902	1
Listeria monocytogenes FSL J1-175	255520581	1
Listeria monocytogenes FSL J1-194	254825045	1
Listeria monocytagenes FSL J1-208	422810631	1
Listeria monocytogenes FSL N3-165	254829042	1
Listeria monocytogenes FSL R2-503	254854201	1
Listeria monocytogenes str. 1/2a F6854	47097148	1
Megasphaera sp. UPII 135-E	342218215	1
Oenococcus kitaharae DSM 17330	366983953	1
Oenococcus kitaharae DSM 17330	372325145	1
Olsenella uli DSM 7084	302336020	1
Pediococcus acidilactici DSM 20284	304386254	1
Pediococcus acidilactici MA18/5M	418068659	1
Peptoniphilus duerdenii ATCC BAA-1640	304438954	1
Planococcus antarcticus DSM 14505	389815359	1
Psychroflexus torquis ATCC 700755	408489713	1
Ruminococcus lactaris ATCC 29176	197301447	1
Scardovia wiggsiae F0424	423349694	1
Solobacterium moorei F0204	320528778	1
Staphylococcus pseudintermedius ED99	323463801	1
Staphylococcus pseudintermedius ED99	386318630	1
Staphylococcus simulans ACS-120-V-Sch1	414160476	1
Streptococcus agalactiae 2603V/R	22537057	1
Streptococcus agalactiae 515	77413160	1
Streptococcus agalactiae A909	76788458	1
Streptococcus agalactiae ATCC 13813	339301617	1
Streptococcus agalactiae CJB111	77411010	1
Streptococcus agalactiae COH1	77407964	1
Streptococcus agalactiae FSL S3-026	417005168	1
Streptococcus agalactiae GB00112	421147428	1

Streptococcus agalactiae H36B	77405721	1
Streptococcus agalactiae NEM316	25010965	1
Streptococcus agalactiae SA20-06	410594450	1
Streptococcus agalactiae STIR-CD-17	421532069	1
Streptococcus anginosus F0211	315223162	1
Streptococcus anginosus SK1138	421490579	1
Streptococcus anginosus SK52 = DSM 20563	335031483	1
Streptococcus bovis ATCC 700338	306833855	1
Streptococcus canis FSL Z3-227	392329410	1
Streptococcus constellatus subsp. constellatus SK53	418965022	1
Streptococcus dysgalactiae subsp. equisimilis AC-2713	410494913	1
Streptococcus dysgalactiae subsp. equisimilis ATCC 12394	386317166	1
Streptococcus dysgalactiae subsp. equisimilis GGS_124	251782637	1
Streptococcus dysgalactiae subsp. equisimilis RE378	408401787	1
Streptococcus equi subsp. zooepidemicus MGCS10565	195978435	1
Streptococcus equinus ATCC 9812	320547102	1
Streptococcus gallolyticus subsp. gallolyticus ATCC BAA-2069	325978669	1
Streptococcus gallolyticus subsp. gallolyticus TX20005	306831733	1
Streptococcus gallolyticus UCN34	288905639	1
Streptococcus infantarius subsp. infantarius CJ18	379705580	1
Streptococcus iniae 9117	406658208	1
Streptococcus macacae NCTC 11558	357636406	1
Streptococcus mitis SK321	307710946	1
Streptococcus mutans 11SSST2	449165720	1
Streptococcus mutans 11SSST2	449951835	1
Streptococcus mutans 11VS1	449976542	1
Streptococcus mutans 14D	450149988	1
Streptococcus mutans 15VF2	449170557	1
Streptococcus mutans 15VF2	449965974	1
Streptococcus mutans 1SM1	449158457	1
Streptococcus mutans 1SM1	449920643	1
Streptococcus mutans 24	449247589	1
Streptococcus mutans 24	450180942	1
Streptococcus mutans 2VS1	449174812	1
Streptococcus mutans 2VS1	449968746	1
Streptococcus mutans 3SN1	449162653	1
Streptococcus mutans 3SN1	449931425	1
Streptococcus mutans 4SM1	449159838	1
Streptococcus mutans 4SM1	449927152	1

Streptococcus mutans 4VF1	449167132	1
Streptococcus mutans 4VF1	449961027	1
Streptococcus mutans 5SM3	449176693	1
Streptococcus mutans 5SM3	449980571	1
Streptococcus mutans 66-2A	449240165	1
Streptococcus mutans 66-2A	450160342	1
Streptococcus mutans 8ID3	449154769	1
Streptococcus mutans 8ID3	449872064	1
Streptococcus mutans A19	449187668	1
Streptococcus mutans A19	450013175	1
Streptococcus mutans B	450166294	1
Streptococcus mutans G123	450029806	1
Streptococcus mutans GS-5	397650022	1
Streptococcus mutans LJ23	387785882	1
Streptococcus mutans M21	449194333	1
Streptococcus mutans M21	450036249	1
Streptococcus mutans M230	449260994	1
Streptococcus mutans M230	449903532	1
Streptococcus mutans M2A	449209586	1
Streptococcus mutans M2A	450074072	1
Streptococcus mutans N29	449182997	1
Streptococcus mutans N29	450003067	1
Streptococcus mutans N3209	449210660	1
Streptococcus mutans N3209	450077860	1
Streptococcus mutans N66	449212466	1
Streptococcus mutans N66	450083993	1
Streptococcus mutans NFSM1	449202104	1
Streptococcus mutans NFSM1	450051112	1
Streptococcus mutans NLML1	450140393	1
Streptococcus mutans NLML4	449202681	1
Streptococcus mutans NLML4	450059882	1
Streptococcus mutans NLML9	449209148	1
Streptococcus mutans NLML9	450066176	1
Streptococcus mutans NMT4863	449186850	1
Streptococcus mutans NMT4863	450007078	1
Streptococcus mutans NN2025	290580220	1
Streptococcus mutans NV1996	450086338	1
Streptococcus mutans NVAB	449181424	1

Streptococcus mutans NVAB	449990810	1
Streptococcus mutans R221	449258042	1
Streptococcus mutans R221	449899675	1
Streptococcus mutans S1B	449251227	1
Streptococcus mutans S1B	449877120	1
Streptococcus mutans SF1	450098705	1
Streptococcus mutans SF14	449221374	1
Streptococcus mutans SF14	450107816	1
Streptococcus mutans SM1	449245264	1
Streptococcus mutans SM1	450176410	1
Streptococcus mutans SM4	449246010	1
Streptococcus mutans SM4	450170248	1
Streptococcus mutans SM6	449223000	1
Streptococcus mutans SM6	450112022	1
Streptococcus mutans ST6	449227252	1
Streptococcus mutans ST6	450123011	1
Streptococcus mutans UA159	24379809	1
Streptococcus mutans W6	450094364	1
Streptococcus oralis SK304	421488030	1
Streptococcus oralis SK610	419782534	1
Streptococcus pseudoporcinus LQ 940-04	416852857	1
Streptococcus pyogenes M1	13622193	1
Streptococcus pyogenes MGAS10750	94543903	1
Streptococcus pyogenes MGAS15252	94994317	1
Streptococcus pyogenes MGAS2096	383479946	1
Streptococcus pyogenes MGAS315	94992340	1
Streptococcus pyogenes MGAS5005	21910213	1
Streptococcus pyogenes MGAS6180	71910582	1
Streptococcus pyogenes MGAS9429	71903413	1
Streptococcus pyogenes NZ131	94988516	1
Streptococcus pyogenes SSI-1	28896088	1
Streptococcus ratti FA-1 = DSM 20564	400290495	1
Streptococcus salivarius K12	421452908	1
Streptococcus sanguinis SK115	422848603	1
Streptococcus sanguinis SK330	422860049	1
Streptococcus sanguinis SK353	422821159	1
Streptococcus sanguinis SK49	422884106	1
Streptococcus sp. C300	322375978	1

Streptococcus sp. F0441	414157437	1
Streptococcus sp. M334	322378004	1
Streptococcus sp. oral taxon 56 str. F0418	339640839	1
Streptococcus sp. oral taxon 71 str. 73H25AP	306826314	1
Streptococcus suis ST1	389856936	1
Streptococcus thermophilus	343794781	1
Streptococcus thermophilus LMD-9	116628213	1
Streptococcus thermophilus MN-ZLW-002	387910220	1
Streptococcus thermophilus ND03	386087120	1
Treponema denticola AL-2	449103686	1
Treponema denticola ASLM	449106292	1
Treponema denticola ATCC 35405	42525843	1
Treponema denticola H1-T	449118593	1
Treponema denticola H-22	449117322	1
Treponema denticola OTK	449125136	1
Treponema denticola SP37	449130155	1
Veillonella atypica ACS-134-V-Col7a	303229466	1
Veillonella parvula ATCC 17745	282849530	1
Veillonella sp. 6_1_27	294792465	1
Veillonella sp. oral taxon 780 str. F0422	342213964	1
Streptococcus pyogenes SF370 (M1 GAS)	209559356	1
Streptococcus pyogenes MGAS10270	56808315	1
Acidovorax ebreus TPSY	222109285	2
Actinobacillus minor NM305	240949037	2
Actinobacillus pleuropneumoniae serovar 10 str. D13039	307256472	2
Actinobacillus succinogenes 130Z	152978060	2
Actinobacillus suis H91-0380	407692091	2
Alicyclobacillus hesperidum URH17-3-68	403744858	2
Aminomonas paucivorans DSM 12260	312879015	2
Bacillus cereus BAG4X12-1	423439645	2
Bacillus cereus BAG4X2-1	423445130	2
Bacillus cereus Rock1-15	229113166	2
Bacillus smithii 7_3_47FAA	365156657	2
Bacillus thuringiensis serovar finitimus YBT-020	384183447	2
Bacteroides sp. 3_1_33FAA	265750948	2
Brevibacillus laterosporus GI-9	421874297	2
Campylobacter coli 1098	419564797	2
Campylobacter coli 111-3	419536531	2

Campylobacter coli 132-6	419572019	2
Campylobacter coli 151-9	419603415	2
Campylobacter coli 1909	419576091	2
Campylobacter coli 1957	419581876	2
Campylobacter coli 2692	419553162	2
Campylobacter coli 59-2	419578074	2
Campylobacter coli 67-8	419587721	2
Campylobacter coli 80352	419559505	2
Campylobacter coli 80352	419558307	2
Campylobacter jejuni subsp. doylei 269.97	153952471	2
Campylobacter jejuni subsp. jejuni 110-21	419676124	2
Campylobacter jejuni subsp. jejuni 129-258	419619138	2
Campylobacter jejuni subsp. jejuni 1336	283956897	2
Campylobacter jejuni subsp. jejuni 140-16	419681578	2
Campylobacter jejuni subsp. jejuni 1577	419685099	2
Campylobacter jejuni subsp. jejuni 1854	419689467	2
Campylobacter jejuni subsp. jejuni 1997-10	419666522	2
Campylobacter jejuni subsp. jejuni 2008-1025	419650041	2
Campylobacter jejuni subsp. jejuni 2008-872	419654778	2
Campylobacter jejuni subsp. jejuni 2008-979	419660762	2
Campylobacter jejuni subsp. jejuni 2008-988	419655317	2
Campylobacter jejuni subsp. jejuni 2008-988	419656328	2
Campylobacter jejuni subsp. jejuni 260.94	86152042	2
Campylobacter jejuni subsp. jejuni 414	283953849	2
Campylobacter jejuni subsp. jejuni 51037	419674189	2
Campylobacter jejuni subsp. jejuni 51494	419619463	2
Campylobacter jejuni subsp. jejuni 53161	419647275	2
Campylobacter jejuni subsp. jejuni 60004	419629136	2
Campylobacter jejuni subsp. jejuni 81116	157415744	2
Campylobacter jejuni subsp. jejuni 84-25	88596565	2
Campylobacter jejuni subsp. jejuni 87459	419680124	2
Campylobacter jejuni subsp. jejuni ATCC 33560	419643715	2
Campylobacter jejuni subsp. jejuni CF93-6	86149266	2
Campylobacter jejuni subsp. jejuni CG8486	148925683	2
Campylobacter jejuni subsp. jejuni HB93-13	86152450	2
Campylobacter jejuni subsp. jejuni LMG 23210	419696801	2
Campylobacter jejuni subsp. jejuni LMG 23211	419697443	2
Campylobacter jejuni subsp. jejuni LMG 23263	419628620	2

Campylobacter jejuni subsp. jejuni LMG 23264	419632476	2
Campylobacter jejuni subsp. jejuni LMG 23269	419634246	2
Campylobacter jejuni subsp. jejuni LMG 23357	419641132	2
Campylobacter jejuni subsp. jejuni NCTC 11168	218563121	2
Campylobacter jejuni subsp. jejuni NW	424845990	2
Campylobacter jejuni subsp. jejuni PT14	407942868	2
Campylobacter lari	345468028	2
Clostridium cellulolyticum H10	220930482	2
Clostridium perfringens C str. JGS1495	169343975	2
Clostridium perfringens D str. JGS1721	182624245	2
Haemophilus parainfluenzae ATCC 33392	325578067	2
Haemophilus parainfluenzae CCUG 13788	359298684	2
Haemophilus parainfluenzae T3T1	345430422	2
Haemophilus sputorum HK 2154	402304649	2
Helicobacter canadensis MIT 98-5491	253828136	2
Helicobacter cinaedi ATCC BAA-847	396079277	2
Helicobacter cinaedi CCUG 18818	313144862	2
Helicobacter cinaedi PAGU611	386762035	2
Helicobacter mustelae 12198	291276265	2
Ilyobacter polytropus DSM 2926	310780384	2
Kingella kingae PYKK081	381401699	2
Lactobacillus coryniformis subsp. torquens KCTC 3535	336393381	2
Neisseria bacilliformis ATCC BAA-1200	329117879	2
Neisseria cinerea ATCC 14685	261378287	2
Neisseria flavescens SK114	241759613	2
Neisseria lactamica 020-06	313669044	2
Neisseria meningitidis 053442	161869390	2
Neisseria meningitidis 2007056	433531983	2
Neisseria meningitidis 63049	433514137	2
Neisseria meningitidis 8013	385324780	2
Neisseria meningitidis 92045	421559784	2
Neisseria meningitidis 93003	421538794	2
Neisseria meningitidis 93004	421541126	2
Neisseria meningitidis 96023	433518260	2
Neisseria meningitidis 98008	421555531	2
Neisseria meningitidis alpha14	254804356	2
Neisseria meningitidis alpha275	254672046	2
Neisseria meningitidis ATCC 13091	304388355	2

Neisseria meningitidis N1568	416164244	2
Neisseria meningitidis NM140	421545139	2
Neisseria meningitidis NM220	418291220	2
Neisseria meningitidis NM233	418288950	2
Neisseria meningitidis WUE 2594	385337435	2
Neisseria meningitidis Z2491	218767588	2
Neisseria sp. oral taxon 14 str. F0314	298369677	2
Neisseria wadsworthii 9715	350570326	2
Pasteurella multocida subsp. gallicida X73	425063822	2
Pasteurella multocida subsp. multocida str. P52VAC	421263876	2
Pasteurella multocida subsp. multocida str. Pm70	15602992	2
Phascolarctobacterium succinatutens YIT 12067	323142435	2
Roseburia intestinalis L1-82	257413184	2
Roseburia intestinalis M50/1	291537230	2
Roseburia inulinivorans DSM 16841	225377804	2
Simonsiella muelleri ATCC 29453	404379108	2
Sporalactobacillus vineae DSM 21990 = SL153	404330915	2
Subdoligranulum sp. 4_3_54A2FAA	365132400	2
Wolinella succinogenes DSM 1740	34557790	2
Catellicoccus marimammalium M35/04/3	424780480	3
Clostridium spiroforme DSM 1552	169349750	3
Enterococcus faecalis Fly1	257084992	3
Enterococcus faecalis R508	424761124	3
Enterococcus faecalis T11	257419486	3
Enterococcus faecalis TX0012	315149830	3
Enterococcus faecalis TX0012	422729710	3
Enterococcus faecalis TX1342	422701955	3
Eubacterium dolichum DSM 3991	160915782	3
Eubacterium rectale ATCC 33656	238924075	3
Eubacterium ventriosum ATCC 27560	154482474	3
Facklamia hominis CCUG 36813	406671118	3
Lactobacillus farciminis KCTC 3681	336394701	3
Listeriaceae bacterium TTU M1-001	381184145	3
Staphylococcus aureus subsp. aureus	403411236	3
Staphylococcus lugdunensis M23590	315659848	3
Streptococcus anginosus 1_2_62CV	319939170	3
Streptococcus gallolyticus UCN34	288905632	3
Streptococcus gordonii str. Challis substr. CH1	157150687	3

Streptococcus infantarius ATCC BAA-102	171779984	3
Streptococcus macedonicus ACA-DC 198	374338350	3
Streptococcus mitis ATCC 6249	306829274	3
Streptococcus mutans NLML5	449203378	3
Streptococcus mutans NLML5	450064617	3
Streptococcus mutans NLML8	449151037	3
Streptococcus mutans NLML8	450133520	3
Streptococcus mutans ST1	449228751	3
Streptococcus mutans ST1	450114718	3
Streptococcus mutans U2A	449232458	3
Streptococcus mutans U2A	450125471	3
Streptococcus oralis SK1074	418974877	3
Streptococcus oralis SK313	417940002	3
Streptococcus parasanguinis F0449	419799964	3
Streptococcus pasteurianus ATCC 43144	336064611	3
Streptococcus salivarius JIM8777	387783792	3
Streptococcus salivarius PS4	419707401	3
Streptococcus sp. BS35b	401684660	3
Streptococcus sp. C150	322372617	3
Streptococcus sp. GMD6S	406576934	3
Streptococcus suis 89/1591	223932525	3
Streptococcus suis D9	386584496	3
Streptococcus suis ST3	330833104	3
Streptococcus thermophilus CNRZ1066	55822627	3
Streptococcus thermophilus JIM 8232	386344353	3
Streptococcus thermophilus LMD-9	116627542	3
Streptococcus thermophilus LMG 18311	55820735	3
Streptococcus thermophilus MN-ZLW-002	387909441	3
Streptococcus thermophilus MTCC 5460	445374534	3
Streptococcus thermophilus ND03	386086348	3
Streptococcus vestibularis ATCC 49124	322517104	3
Anaerophaga sp. HS1	371776944	4
Anaerophaga thermohalophila DSM 12881	346224232	4
Bacteroides coprophilus DSM 18228	224026357	4
Bacteroides coprosuis DSM 18011	333031006	4
Bacteroides dorei DSM 17855	212694363	4
Bacteroides eggerthii 1_2_48FAA	317474201	4
Bacteroides faecis 27-5	380696107	4

Bacteroides fluxus YIT 12057	329965125	4
Bacteroides nordii CL02T12C05	393788929	4
Bacteroides sp. 20_3	301311869	4
Bacteroides sp. D2	383115507	4
Bacteroides uniformis CL03T00C23	423303159	4
Bacteroides vulgatus CL09T03C04	423312075	4
Bergeyella zoohelcum ATCC 43767	423317190	4
Capnocytophaga gingivalis ATCC 33624	228473057	4
Capnocytophaga sp. CM59	402830627	4
Capnocytophaga sp. oral taxon 324 str. F0483	429756885	4
Capnocytophaga sp. oral taxon 326 str. F0382	429752492	4
Capnocytophaga sp. oral taxon 412 str. F0487	393778597	4
Chryseobacterium sp. CF314	399023756	4
Fibrobacter succinogenes subsp. succinogenes S85	261414553	4
Flavobacteriaceae bacterium S85	372210605	4
Flavobacterium columnare ATCC 49512	365960762	4
Fluviicola taffensis DSM 16823	327405121	4
Ignavibacterium album JCM 16511	385811609	4
Mucilaginibacter paludis DSM 18603	373954054	4
Myroides odoratus DSM 2801	374597806	4
Ornithobacterium rhinotracheale DSM 15997	392391493	4
Prevotella bivia JCVIHMP010	282858617	4
Prevotella buccae ATCC 33574	315607525	4
Prevotella nigrescens ATCC 33563	340351024	4
Prevotella sp. MSX73	402307189	4
Prevotella timonensis CRIS 5C-B1	282881485	4
Prevotella veroralis F0319	260592128	4
Sphingobacterium spiritivorum ATCC 33861	300771242	4
Weeksella virosa DSM 16922	325955459	4
Acidovorax avenae subsp. avenae ATCC 19860	326315085	5
Alicycliphilus denitrificans BC	319760940	5
Alicycliphilus denitrificans K601	330822845	5
Azospirillum sp. B510	288957741	5
Bradyrhizabium sp. BTAi1	148255343	5
Candidatus Puniceispirillum marinum IMCC1322	294086111	5
Dinoroseabacter shibae DFL 12	159042956	5
gamma proteobacterium HdN1	304313029	5

Nitrobacter hamburgensis X14	92109262	5
Nitrosomonas sp. AL212	325983496	5
Ralstonia syzygii R24	344171927	5
Rhodovulum sp. PH10	402849997	5
Sphingobium sp. AP49	398385143	5
Sphingomonas sp. S17	332188827	5
Verminephrobacter eiseniae EF01-2	121608211	5
Bacteroides fragilis 638R	375360193	6
Bacteroides fragilis NCTC 9343	60683389	6
Bacteroides sp. 2_1_16	265767599	6
Bacteroides sp. 3_1_19	298377533	6
Bacteroides sp. D2	383110723	6
Bacteroidetes oral taxon 274 str. F0058	298373376	6
Barnesiella intestinihominis YIT 11860	404487228	6
Belliella baltica DSM 15883	390944707	6
Bergeyella zoohelcum CCUG 30536	406673990	6
Capnocytophaga canimorsus Cc5	340622236	6
Capnocytophaga ochracea DSM 7271	256819408	6
Capnocytophaga sp. oral taxon 329 str. F0087	332882466	6
Capnocytophaga sp. oral taxon 335 str. F0486	420149252	6
Capnocytophaga sp. oral taxon 380 str. F0488	429748017	6
Capnocytophaga sputigena Capno	213962376	6
Flavobacterium psychrophilum JIP02/86	150025575	6
Galbibacter sp. ck-I2-15	408370397	6
Indibacter alkaliphilus LW1	404451234	6
Joostella marina DSM 19592	386818981	6
Kordia algicida OT-1	163754820	6
Marinilabilia sp. AK2	410030899	6
Myroides injenensis M09-0166	399927444	6
Niabella soli DSM 19437	374372722	6
Parabacteroides johnsonii DSM 18315	218258638	6
Parabacteroides sp. D13	256840409	6
Porphyromonas sp. oral taxon 279 str. F0450	402847315	6
Prevotella histicola F0411	357042839	6
Prevotella intermedia 17	387132277	6
Prevotella nigrescens F0103	445119230	6
Prevotella oralis ATCC 33269	323344874	6
Prevotella sp. oral taxon 306 str. F0472	383811446	6

Riemerella anatipestifer RA-CH-1	407451859	6
Riemerella anatipestifer RA-GD	386321727	6
Zunongwangia profunda SM-A87	295136244	6
Actinomyces coleocanis DSM 15436	227494853	7
Actinomyces georgiae F0490	420151340	7
Actinomyces naeslundii str. Howell 279	400293272	7
Actinomyces sp. ICM47	396585058	7
Actinomyces sp. oral taxon 175 str. F0384	343523232	7
Actinomyces sp. oral taxon 181 str. F0379	429758968	7
Actinomyces sp. oral taxon 848 str. F0332	269219760	7
Actinomyces turicensis ACS-279-V-Col4	405979650	7
Bifidobacterium dentium Bd1	283456135	7
Bifidobacterium longum DJO10A	189440764	7
Bifidobacterium longum subsp. longum 2-2B	419852381	7
Bifidobacterium longum subsp. longum KACC 91563	384200944	7
Bifidobacterium sp. 12_1_47BFAA	317482066	7
Corynebacterium accolens ATCC 49725	227502575	7
Corynebacterium accolens ATCC 49726	306835141	7
Corynebacterium diphtheriae 241	375289763	7
Corynebacterium diphtheriae 31A	376283539	7
Corynebacterium diphtheriae BH8	376286566	7
Corynebacterium diphtheriae bv. intermedius str. NCTC 5011	419861895	7
Corynebacterium diphtheriae C7 (beta)	376289243	7
Corynebacterium diphtheriae HC02	376292154	7
Corynebacterium diphtheriae NCTC 13129	38232678	7
Corynebacterium diphtheriae VA01	376256051	7
Corynebacterium matruchotii ATCC 14266	305681510	7
Corynebacterium matruchotii ATCC 33806	225021644	7
Gardnerella vaginalis 1500E	415717744	7
Gardnerella vaginalis 284V	415703177	7
Gardnerella vaginalis 5-1	298252606	7
Mobiluncus curtisii subsp. holmesii ATCC 35242	315656340	7
Mobiluncus mulieris 28-1	269977848	7
Mobiluncus mulieris FB024-16	307700167	7
Scardovia inopinata F0304	294790575	7
Actinomyces sp. oral taxon 180 str. F0310	315605738	8

Gluconacetobacter diazotrophicus PAl 5	209542524	8
Gluconacetobacter diazotrophicus PAl 5	162147907	8
Methylocystis sp. ATCC 49242	323139312	8
Methylosinus trichosporium OB3b	296446027	8
Rhodopseudomonas palustris BisB18	90425961	8
Rhodopseudomonas palustris BisB5	91975509	8
Tistrella mobilis KA081020-065	389874754	8
Mycoplasma canis PG 14	384393286	9
Mycoplasma canis PG 14	419703974	9
Mycoplasma canis UF31	384937953	9
Mycoplasma canis UF33	419704625	9
Mycoplasma canis UFG1	419705269	9
Mycoplasma canis UFG4	419705920	9
Mycoplasma cynos C142	433625054	9
Mycoplasma gallisepticum NC95_13295-2-2P	401767318	9
Mycoplasma gallisepticum NY01_2001.047-5-1P	401768851	9
Mycoplasma gallisepticum str. F	284931710	9
Mycoplasma gallisepticum str. F	385326554	9
Mycoplasma gallisepticum str. R(low)	294660600	9
Mycoplasma synoviae 53	71894592	9
Mycoplasma synoviae 53	144575181	9
Prevotella buccalis ATCC 35310	282878504	10
Prevotella ruminicola 23	294674019	10
Prevotella stercorea DSM 18206	359406728	10
Prevotella tannerae ATCC 51259	258648111	10
Prevotella timonensis CRIS 5C-B1	282880052	10
Burkholderiales bacterium 1_1_47	303257695	11
Parasutterella excrementihominis YIT 11859	331001027	11
Sutterella wadsworthensis 3_1_45B	319941583	11
Elusimicrobium minutum Pei191	187250660	12
Sphaerochaeta globus str. Buddy	325972003	12
uncultured Termite group 1 bacterium phylotype Rs-D17	189485059	12
Flavobacterium branchiophilum FL-15	347536497	13
Flavobacterium columnare ATCC 49512	365959402	13
Odoribacter laneus YIT 12061	374384763	13
Prevotella denticola CRIS 18C-A	325859619	14
Prevotella micans F0438	373501184	14
Prevotella sp. C561	345885718	14

Francisella tularensis subsp. tularensis WY96-3418	134302318	15	
Francisella cf. novicida 3523	387824704	16	
Francisella cf. novicida Fx1	385792694	16	
Francisella novicida FTG	208779141	16	
Francisella novicida GA99-3548	254374175	16	
Francisella novicida U112	118497352	16	
Francisella tularensis subsp. novicida GA99-3549	254372717	16	
Wolinella succinogenes DSM 1740	34557932	17	
gamma proteobacterium HTCC5015	254447899	18	
Legionella pneumophila 130b	307608922	19	
Legionella pneumophila str. Paris	54296138	19	
Mycoplasma ovipneumoniae SC01	363542550	20	
Streptobacillus moniliformis DSM 12112	269123826	21	
Mycoplasma mobile 163K	47458868	22	
Alcanivorax sp. W11-5	407803669	23	
Caenispirillum salinarum AK4	427429481	23	
Rhodospirillum rubrum ATCC 11170	83591793	23	
Treponema sp. JC4	384109266	23	
Ruminococcus albus 8	325677756	24	
uncultured delta proteobacterium HF0070_07E19	297182908	24	
Acidothermus cellulolyticus 11B	117929158	25	
Nitratifractor salsuginis DSM 16511	319957206	26	
Akkermansia muciniphila ATCC BAA-835	187736489	27	
Parvibaculum lavamentivorans DS-1	154250555	28	

Example 6

[0001] Phylogenetic clustering of Cas9 defines dual-RNA:Cas9 exchangeability.

[0002] As described above, clustering of Cas9 orthologs correlates with the ability to substitute for the RNA-stabilizing role of *S. pyogenes* Cas9 in tracrRNA:pre-crRNA processing by RNase III *in vivo* (Figure 4B). The exchangeability between Cas9 and dual-RNA in closely related CRISPR-Cas systems was investigated at the level of DNA interference.

[0003] Plasmid cleavage assays were performed using *S. pyogenes* Cas9 complexed with dual-RNAs from selected CRISPR-Cas systems representative of the clustering of the type II CRISPR-Cas systems. As shown in Figure 6A (upper panel), *S. pyogenes* Cas9 can cleave target DNA in the presence of dual-RNAs from *S. mutans* and *S. thermophilus** (type II-A, yellow subcluster), but not from any other tested species. The same result was observed when the dual-RNA from *S. pyogenes* was incubated with Cas9 orthologs from different bacteria (Figure 6A, lower panel). Cleavage assays were also performed with all Cas9 orthologs incubated with cognate and non-cognate dual-RNAs on their PAM-specific plasmid DNA. Only the combinations of Cas9 and dual-RNA within the same type II subcluster conferred dsDNA cleavage activity (Figure 6B, Supplementary Figure S10). More striking was the gradient of activity dependent on how closely related the species are in the corresponding type II group. This effect can be observed for *C. jejuni* Cas9 that is able to cleave DNA in the presence of dual-RNA from *P. multocida* and *N. meningitidis*, but not as efficient as with its own RNA (type II-C, blue subcluster). This finding is in good agreement with the phylogenetic tree of Cas9 (Figure 1A) showing that all three Cas9 orthologs belong to type II-C but *C. jejuni* Cas9 clusters more distantly from *P. multocida* and *N. meningitidis* Cas9. This effect was even greater for *S. thermophilus*** Cas9 which belongs to type II-A together with *S. pyogenes*, *S. mutans* and *S. thermophilus**. However, none of the dual-RNAs from the three latter loci could direct DNA cleavage by *S. thermophilus*** Cas9. This result supports the recent findings demonstrating the lack of exchangeability between Cas9 from CRISPR1 and CRISPR3 of *S. thermophilus* DGCC7710 with regard to dual-RNA binding (17). Cas9 and tracrRNA:crRNA interchangeabilty is contemplated to directly result from Cas9 co-evolution with dual-RNA and follows the Cas9 phylogeny that may differ from the phylogeny of the respective bacterial species due to horizontal transfer.

[0004] Thus, to investigate the interchangeability between type II subgroups at the level of DNA interference, the PAMs specific for each of the 8 selected Cas9 orthologs (28) were determined. By aligning potential crRNA-targeted sequences, conserved motifs adjacent to the protospacers in all selected species were identified. These motifs were then shown to be essential for DNA interference activity of the cognate dual-RNA:Cas9 complex *in vitro*. The interchangeability between dual-RNA and Cas9 from different subclusters was tested using plasmid cleavage assays. Only closely related Cas9

proteins can exchange their cognate dual-RNAs and still exert cleavage activity when using the Cas9 specific PAM. The specificity of Cas9 towards dual-RNAs is highly sensitive to the Cas9 sequence relatedness. This sensitivity is observed with Cas9 from *C. jejuni* that displays full cleavage activity with its cognate dual-RNA but reduced activity with dual-RNAs from *N. meningitidis* or *P. multocida* which belong to different subclusters of type II-C. It is contemplated that Cas9 possesses specificity for the secondary structure of dual-RNAs, given that bioinformatics predictions suggest similar structures of repeat:antirepeat duplexes in closely related CRISPR-Cas systems (Supplementary Figure S12).

[0005] While the present invention has been described in terms of specific embodiments, it is understood that variations and modifications will occur to those skilled in the art. Accordingly, only such limitations as appear in the claims should be placed on the invention.

Documents Cited

[0006] All publications and patents cited in this specification are herein incorporated by reference as if each individual publication or patent were specifically and individually indicated to be incorporated by reference and are incorporated herein by reference to disclose and describe the methods and/or materials in connection with which the publications are cited. The citation of any publication is for its disclosure prior to the filing date and should not be construed as an admission that the present invention is not entitled to antedate such publication by virtue of prior invention. Further, the dates of publication provided may be different from the actual publication dates which may need to be independently confirmed.

1. Cho, S.W., Kim, S., Kim, J.M. and Kim, J.S. (2013) Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. *Nat. Biotechnol.*, **31**, 230-232.
2. Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A. *et al.* (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science*, **339**, 819-823.
3. DiCarlo, J.E., Norville, J.E., Mali, P., Rios, X., Aach, J. and Church, G.M. (2013) Genome engineering in *Saccharomyces cerevisiae* using CRISPR-Cas systems. *Nucleic Acids Res.*, **41**, 4336-4343.
4. Friedland, A.E., Tzur, Y.B., Esvelt, K.M., Colaiacovo, M.P., Church, G.M. and Calarco, J.A. (2013) Heritable genome editing in *C. elegans* via a CRISPR-Cas9 system. *Nat. Methods*, **10**, 741-743.
5. Gratz, S.J., Cummings, A.M., Nguyen, J.N., Hamm, D.C., Donohue, L.K., Harrison, M.M., Wildonger, J. and O'Connor-Giles, K.M. (2013) Genome engineering of Drosophila with the CRISPR RNA-guided Cas9 nuclease. *Genetics*, **194**, 1029-1035.
6. Hwang, W.Y., Fu, Y., Reyon, D., Maeder, M.L., Tsai, S.Q., Sander, J.D., Peterson, R.T., Yeh, J.R. and Joung, J.K. (2013) Efficient genome editing in zebrafish using a CRISPR-Cas system. *Nat. Biotechnol.*, **31**, 227-229.
7. Jiang, W., Bikard, D., Cox, D., Zhang, F. and Marraffini, L.A. (2013) RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nat. Biotechnol.*, **31**, 233-239.
8. Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., Dicarlo, J.E., Norville, J.E. and Church, G.M. (2013) RNA-guided human genome engineering via Cas9. *Science*, **339**, 823-826.
9. Shen, B., Zhang, J., Wu, H., Wang, J., Ma, K., Li, Z., Zhang, X., Zhang, P. and Huang, X. (2013) Generation of gene-modified mice via Cas9/RNA-mediated gene targeting. *Cell Res.*, **23**, 720-723.
10. Wang, H., Yang, H., Shivalila, C.S., Dawlaty, M.M., Cheng, A.W., Zhang, F. and Jaenisch, R. (2013) One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. *Cell*, **153**, 910-918.
11. Jinek, M., East, A., Cheng, A., Lin, S., Ma, E. and Doudna, J. (2013) RNA-programmed genome editing in human cells. *eLIFE*, **2**, e00471.
12. Li, J.F., Norville, J.E., Aach, J., McCormack, M., Zhang, D., Bush, J., Church, G.M. and Sheen, J. (2013) Multiplex and homologous recombination-mediated genome editing in *Arabidopsis* and *Nicotiana benthamiana* using guide RNA and Cas9. *Nat. Biotechnol.*, **31**, 688-691.

13. Nekrasov, V., Staskawicz, B., Weigel, D., Jones, J.D. and Kamoun, S. (2013) Targeted mutagenesis in the model plant *Nicotiana benthamiana* using Cas9 RNA-guided endonuclease. *Nat. Biotechnol.*, **31**, 691-693.

14. Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A. and Charpentier, E. (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, **337**, 816-821.

15. Chylinski, K., Le Rhun, A. and Charpentier, E. (2013) The tracrRNA and Cas9 families of type II CRISPR-Cas immunity systems. *RNA Biol.*, **10**, 726-737.

16. Deltcheva, E., Chylinski, K., Sharma, C.M., Gonzales, K., Chao, Y., Pirzada, Z.A., Eckert, M.R., Vogel, J. and Charpentier, E. (2011) CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature*, **471**, 602-607.

17. Karvelis, T., Gasiunas, G., Miksys, A., Barrangou, R., Horvath, P. and Siksnys, V. (2013) crRNA and tracrRNA guide Cas9-mediated DNA interference in *Streptococcus thermophilus*. *RNA Biol.*, **10**, 841-851.

18. Garneau, J.E., Dupuis, M.E., Villion, M., Romero, D.A., Barrangou, R., Boyaval, P., Fremaux, C., Horvath, P., Magadan, A.H. and Moineau, S. (2010) The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature*, **468**, 67-71.

19. Magadan, A.H., Dupuis, M.E., Villion, M. and Moineau, S. (2012) Cleavage of phage DNA by the *Streptococcus thermophilus* CRISPR3-Cas system. *PLoS One*, **7**, e40913.

20. Haft, D.H., Selengut, J., Mongodin, E.F. and Nelson, K.E. (2005) A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLoS Comput. Biol.*, **1**, e60.

21. Makarova, K.S., Grishin, N.V., Shabalina, S.A., Wolf, Y.I. and Koonin, E.V. (2006) A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol. Direct*, **1**, 7.

22. Gasiunas, G., Barrangou, R., Horvath, P. and Siksnys, V. (2012) Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proc. Natl. Acad. Sci. U. S. A.*, **109**, E2579-2586.

23. Sapranauskas, R., Gasiunas, G., Fremaux, C., Barrangou, R., Horvath, P. and Siksnys, V. (2011) The *Streptococcus thermophilus* CRISPR/Cas system provides immunity in *Escherichia coli*. *Nucleic Acids Res.*, **39**, 9275-9282.

24. Mali, P., Aach, J., Stranges, P.B., Esvelt, K.M., Moosburner, M., Kosuri, S., Yang, L. and Church, G.M. (2013) Cas9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat Biotechnol.*, **31**, 833-838.

25. Ran, F.A., Hsu, P.D., Lin, C.Y., Gootenberg, J.S., Konermann, S., Trevino, A.E., Scott, D.A., Inoue, A., Matoba, S., Zhang, Y. *et al.* (2013) Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell*, **154**, 1380-1389.

26. Deveau, H., Barrangou, R., Garneau, J.E., Labonte, J., Fremaux, C., Boyaval, P., Romero, D.A., Horvath, P. and Moineau, S. (2008) Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. *J. Bacteriol.*, **190**, 1390-1400.

27. Horvath, P., Romero, D.A., Coute-Monvoisin, A.C., Richards, M., Deveau, H., Moineau, S., Boyaval, P., Fremaux, C. and Barrangou, R. (2008) Diversity, activity, and evolution of CRISPR loci in *Streptococcus thermophilus*. *J. Bacteriol.*, **190**, 1401-1412.

28. Mojica, F.J., Diez-Villasenor, C., Garcia-Martinez, J. and Almendros, C. (2009) Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology*, **155**, 733-740.

29. Bikard, D., Jiang, W., Samai, P., Hochschild, A., Zhang, F. and Marraffini, L.A. (2013) Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system. *Nucleic Acids Res.*, **41**, 7429-7437.

30. Qi, L.S., Larson, M.H., Gilbert, L.A., Doudna, J.A., Weissman, J.S., Arkin, A.P. and Lim, W.A. (2013) Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell*, **152**, 1173-1183.

31. Charpentier, E. and Doudna, J.A. (2013) Biotechnology: Rewriting a genome. *Nature*, **495**, 50-51.

32. Horvath, P. and Barrangou, R. (2013) RNA-guided genome editing a la carte. *Cell Res.*, **23**, 733-734.

33. van der Oost, J. (2013) Molecular biology. New tool for genome surgery. *Science*, **339**, 768-770.

34. Hou, Z., Zhang, Y., Propson, N.E., Howden, S.E., Chu, L.F., Sontheimer, E.J. and Thomson, J.A. (2013) Efficient genome engineering in human pluripotent stem cells using Cas9 from *Neisseria meningitidis. Proc. Natl. Acad. Sci. U. S. A.*, **110**, 15644-15649.

35. Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989) *Molecular Cloning: a Laboratory Manual.* 2nd edn. Cold Spring Harbor, N. Y. ed. Cold Spring Harbor Laboratory Press.

36. Caparon, M.G. and Scott, J.R. (1991) Genetic manipulation of pathogenic streptococci. *Methods Enzymol.*, **204**, 556-586.

37. Kirsch, R.D. and Joly, E. (1998) An improved PCR-mutagenesis strategy for two-site mutagenesis or sequence swapping between related genes. *Nucleic Acids Res.*, **26**, 1848-1850.

38. Siller, M., Janapatla, R.P., Pirzada, Z.A., Hassler, C., Zinkl, D. and Charpentier, E. (2008) Functional analysis of the group A streptococcal *luxS*/AI-2 system in metabolism, adaptation to stress and interaction with host cells. *BMC Microbiol.*, **8**, 188.

39. Mangold, M., Siller, M., Roppenser, B., Vlaminckx, B.J., Penfound, T.A., Klein, R., Novak, R., Novick, R.P. and Charpentier, E. (2004) Synthesis of group A streptococcal virulence factors is controlled by a regulatory RNA molecule. *Mol. Microbiol.*, **53**, 1515-1527.

40. Herbert, S., Barry, P. and Novick, R.P. (2001) Subinhibitory clindamycin differentially inhibits transcription of exoprotein genes in *Staphylococcus aureus. Infect. Immun.*, **69**, 2996-3003.

41. Pall, G.S. and Hamilton, A.J. (2008) Improved northern blot method for enhanced detection of small RNA. *Nat. Protoc.*, **3**, 1077-1084.

42. Urban, J.H. and Vogel, J. (2007) Translational control and target recognition by *Escherichia coli* small RNAs *in vivo. Nucleic Acids Res.*, **35**, 1018-1037.

43. McClelland, M., Hanish, J., Nelson, M. and Patel, Y. (1988) KGB: a single buffer for all restriction endonucleases. *Nucleic Acids Res.*, **16**, 364.

44. Makarova, K.S., Haft, D.H., Barrangou, R., Brouns, S.J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F.J., Wolf, Y.I., Yakunin, A.F. *et al.* (2011) Evolution and classification of the CRISPR-Cas systems. *Nat. Rev. Microbiol.*, **9**, 467-477.

45. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389-3402.

46. Wheeler, D. and Bhagwat, M. (2007) BLAST QuickStart: example-driven web-based BLAST tutorial. *Methods Mol. Biol.*, **395**, 149-176.

47. Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792-1797.

48. Soding, J., Biegert, A. and Lupas, A.N. (2005) The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.*, **33**, W244-248.

49. Price, M.N., Dehal, P.S. and Arkin, A.P. (2010) FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.

50. Bernhart, S.H., Tafer, H., Muckstein, U., Flamm, C., Stadler, P.F. and Hofacker, I.L. (2006) Partition function and base pairing probabilities of RNA heterodimers. *Algorithms Mol. Biol.*, **1**, 3.

51. Hofacker, I.L., Fekete, M. and Stadler, P.F. (2002) Secondary structure prediction for aligned RNA sequences. *Journal of molecular biology*, **319**, 1059-1066.

52. Darty, K., Denise, A. and Ponty, Y. (2009) VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, **25**, 1974-1975.

53. Bhaya, D., Davison, M. and Barrangou, R. (2011) CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation. *Annu. Rev. Genet.*, **45**, 273-297.

54. Zhang, Y., Heidrich, N., Ampattu, B.J., Gunderson, C.W., Seifert, H.S., Schoen, C., Vogel, J. and Sontheimer, E.J. (2013) Processing-independent CRISPR RNAs limit natural transformation in *Neisseria meningitidis*. *Mol. Cell*, **50**, 488-503.

55. Takeuchi, N., Wolf, Y.I., Makarova, K.S. and Koonin, E.V. (2012) Nature and intensity of selection pressure on CRISPR-associated genes. *J. Bacteriol.*, **194**, 1216-1225.

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

57. Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A. and Horvath, P. (2007) CRISPR provides acquired resistance against viruses in prokaryotes. *Science*, **315**, 1709-1712.

58. Sun, W., Li, G. and Nicholson, A.W. (2004) Mutational analysis of the nuclease domain of *Escherichia coli* ribonuclease III. Identification of conserved acidic residues that are important for catalytic function *in vitro*. *Biochemistry*, **43**, 13054-13062.

59. Sun, W., Jun, E. and Nicholson, A.W. (2001) Intrinsic double-stranded-RNA processing activity of *Escherichia coli* ribonuclease III lacking the dsRNA-binding domain. *Biochemistry*, **40**, 14976-14984.

Claims

We claim:

1. A single-molecule guide RNA comprising:

a DNA-targeting segment and a protein-binding segment wherein the protein-binding segment comprises a tracrRNA set out in Supplementary Table S5.

2. The single-molecule guide RNA of claim 1 wherein the protein-binding segment comprises a CRISPR repeat set out in Supplementary Table S5 that is the cognate CRISPR repeat of the tracrRNA of the protein-binding segment.

3. The single-molecule guide RNA of claim 1 or 2 wherein the DNA-targeting segment further comprises RNA complementary to a protospacer-like sequence in a target DNA 5' to a PAM sequence.

4. The single-molecule guide RNA of claim 3 wherein the tracrRNA and CRISPR repeat are respectively at least 80% identical to the C. jejuni tracrRNA and CRISPR repeat set out in Supplementary Table S5 and wherein the PAM sequence is NNNNACA.

5. The single-molecule guide RNA of claim 4 or 8 wherein the RNA complementary to a protospacer-like sequence is RNA complementary to the target sequences set out in one of SEQ ID NOs: 801-973, 1079-1222, 1313-1348, 1372-1415, 1444-1900, 2163-2482 or 2667-2686.

6. A single-molecule guide RNA comprising:

a DNA-targeting segment and a protein-binding segment, wherein the protein-binding segment comprises a tracrRNA at least 80% identical over at least 20 nucleotides to a tracrRNA set out in Supplementary Table S5.

7. A single-molecule guide RNA comprising:

a DNA-targeting segment and a protein-binding segment, wherein the protein-binding segment comprises a tracrRNA at least 80% identical over at least 20 nucleotides to a tracrRNA set out in Supplementary Table S5, a CRISPR repeat at least 80% identical to a CRISPR repeat set out in Supplementary Table S5, or both.

8. The single-molecule guide RNA of claim 7 wherein the tracrRNA and CRISPR repeat are respectively at least 80% identical to the C. jejuni tracrRNA and CRISPR repeat set out in Supplementary Table S5 and wherein the PAM sequence is NNNNACA.

9. The single-molecule guide RNA of claim 1 or 6 comprising a linker between the DNA-targeting segment and the protein-binding segment.

10. A DNA encoding a single-molecule guide RNA comprising:

a DNA-targeting segment and a protein-binding segment, wherein the protein-binding segment comprises a tracrRNA set out in Supplementary Table S5.

11. A DNA encoding a single-molecule guide RNA comprising:

a DNA-targeting segment and a protein-binding segment, wherein the protein-binding segment comprises a tracrRNA at least 80% identical over at least 20 nucleotides to a tracrRNA set out in Supplementary Table S5, a CRISPR repeat at least 80% identical to a CRISPR repeat set out in Supplementary Table S5, or both.

12. A vector comprising a DNA encoding a single-molecule guide RNA comprising:

a DNA-targeting segment and a protein-binding segment wherein the protein-binding segment comprises a tracrRNA set out in Supplementary Table S5.

13. A vector comprising a DNA encoding a single-molecule guide RNA comprising:

a DNA-targeting segment and a protein-binding segment, wherein the protein-binding segment comprises a tracrRNA at least 80% identical over at least 20 nucleotides to a tracrRNA set out in Supplementary Table S5, a CRISPR repeat at least 80% identical to a CRISPR repeat set out in Supplementary Table S5, or both.

14. A cell comprising a DNA encoding a single-molecule guide RNA comprising:

a DNA-targeting segment and a protein-binding segment, wherein the protein-binding segment comprises a tracrRNA set out in Supplementary Table S5.

15. A cell comprising a DNA encoding a single-molecule guide RNA comprising:

a DNA-targeting segment and a protein-binding segment, wherein the protein-binding segment comprises a tracrRNA at least 80% identical over at least 20 nucleotides to a tracrRNA set out in Supplementary Table S5, a CRISPR repeat at least 80% identical to a CRISPR repeat set out in Supplementary Table S5, or both.

16. A double-molecule guide RNA comprising:

a targeter-RNA and an activator-RNA complementary thereto, wherein the activator-RNA comprises a tracrRNA set out in Supplementary Table S5, and

wherein the guide RNA comprises a modified backbone, a non-natural internucleoside linkage, a nucleic acid mimetic, a modified sugar moiety, a base modification, a modification or sequence that provides for modified or regulated stability, a modification or sequence that provides for subcellular tracking, a modification or sequence that provides for tracking, or a modification or sequence that provides for a binding site for a protein or protein complex.

17. The double-molecule guide RNA of claim 16, wherein the targeter-RNA comprises a CRISPR repeat set out in Supplementary Table S5 that is the cognate CRISPR repeat of the tracrRNA of the protein-binding segment.

18. The double-molecule guide RNA of claim 16 or 17 wherein the targeter-RNA further comprises RNA complementary to a protospacer-like sequence in a target DNA 5' to a PAM sequence.

19. The double-molecule guide RNA of claim 18 wherein the tracrRNA and CRISPR repeat are respectively at least 80% identical to the C. jejuni tracrRNA and CRISPR repeat set out in Supplementary Table S5 and wherein the PAM sequence is NNNNACA.

20. The double-molecule guide RNA of claim 19 or claim 23 wherein the RNA complementary to a protospacer-like sequence is RNA complementary to the target sequences set out in one of SEQ ID NOs: 801-973, 1079-1222, 1313-1348, 1372-1415, 1444-1900, 2163-2482 or 2667-2686.

21. A double-molecule guide RNA comprising:

a targeter-RNA and a activator-RNA, wherein the activator-RNA comprises a tracrRNA at least 80% identical over at least 20 nucleotides to a tracrRNA set out in Supplementary Table S5.

22. The double-molecule guide RNA of claim 21, wherein the targeter-RNA comprises a CRISPR repeat set out in Supplementary Table S5, the cognate CRISPR repeat of the tracrRNA of the activator-RNA set out in Supplementary Table S5, or a CRISPR repeat at least 80% identical to a CRISPR repeat set out in Supplementary Table S5.

23. The double-molecule guide RNA of claim 21 wherein the tracrRNA and CRISPR repeat are respectively at least 80% identical to the C. jejuni tracrRNA and CRISPR repeat set out in Supplementary Table S5 and wherein the PAM sequence is NNNNACA.

24. The double-molecule guide RNA of claim 16 or 21 comprising a linker between the targeter-RNA and the activator-RNA.

25. A DNA encoding a double-molecule guide RNA comprising:

a targeter-RNA and a activator-RNA complementary thereto, wherein the activator-RNA comprises a tracrRNA set out in Supplementary Table S5.

26. A DNA encoding a double-molecule guide RNA comprising:

a targeter-RNA and a activator-RNA complementary thereto, wherein the activator-RNA comprises a tracrRNA at least 80% identical over at least 20 nucleotides to a tracrRNA set out in Supplementary Table S5, a CRISPR repeat at least 80% identical to a CRISPR repeat set out in Supplementary Table S5, or both.

27. A vector comprising a DNA encoding a double-molecule guide RNA comprising:

a targeter-RNA and a activator-RNA complementary thereto, wherein the activator-RNA comprises a tracrRNA set out in Supplementary Table S5.

28. A vector comprising a DNA encoding a double-molecule guide RNA comprising:

a targeter-RNA and a activator-RNA complementary thereto, wherein the activator-RNA comprises a tracrRNA at least 80% identical over at least 20 nucleotides to a tracrRNA set out in Supplementary Table S5, a CRISPR repeat at least 80% identical to a CRISPR repeat set out in Supplementary Table S5, or both.

29. A cell comprising a DNA encoding a double-molecule guide RNA comprising:

a targeter-RNA and a activator-RNA complementary thereto, wherein the activator-RNA comprises a tracrRNA set out in Supplementary Table S5.

30. A cell comprising a DNA encoding a double-molecule guide RNA comprising:

a targeter-RNA and a activator-RNA complementary thereto, wherein the activator-RNA comprises a tracrRNA at least 80% identical over at least 20 nucleotides to a tracrRNA set out in Supplementary Table S5, a CRISPR repeat at least 80% identical to a CRISPR repeat set out in Supplementary Table S5, or both.

31. A method for manipulating DNA in a cell, comprising contacting the DNA with a Cas9 ortholog-guideRNA complex, wherein the complex comprises:

(a) a C. jejuni Cas9 endonuclease heterologous to the cell or an endonuclease with an activity portion at least 90% identical to the activity portion of the C. jejuni Cas9 endonuclease, and a guide RNA targeting the complex to a protospacer-like sequence in the DNA 5' to the PAM sequence NNNNACA;

(b) a P. multocida Cas9 endonuclease heterologous to the cell or an endonuclease with an activity portion at least 90% identical to the activity portion of the P. multocida Cas9 endonuclease, and a

guide RNA targeting the complex to a protospacer-like sequence in the DNA 5' to the PAM sequence GNNNCNNA or NNNNC;

(c) an F. novicida Cas9 endonuclease heterologous to the cell or an endonuclease with an activity portion at least 90% identical to the activity portion of the F. novicida Cas9 endonuclease, and a guide RNA targeting the complex to a protospacer-like sequence in the DNA 5' to the PAM sequence NG;

(d) an S. thermophilus** Cas9 endonuclease heterologous to the cell or an endonuclease with an activity portion at least 90% identical to the activity portion of the S. thermophilus** Cas9 endonuclease, and a guide RNA targeting the complex to a protospacer-like sequence in the DNA 5' to the PAM sequence NNAAAAW;

(e) an L. innocua Cas9 endonuclease heterologous to the cell or an endonuclease with an activity portion at least 90% identical to the activity portion of the L. innocua Cas9 endonuclease, and a guide RNA targeting the complex to a protospacer-like sequence in the DNA 5' to the PAM sequence NGG; or

(f) an S. dysgalactiae Cas9 endonuclease heterologous to the cell or an endonuclease with an activity portion at least 90% identical to the activity portion of the S. dysgalactiae Cas9 endonuclease, and a guide RNA targeting the complex to a protospacer-like sequence in the DNA 5' to the PAM sequence NGG.

32. The method of claim 31 wherein the cell is a bacterial cell, a fungal cell, an archaea cell, a plant cell or an animal cell.

33. The method of claim 31 wherein the guide RNA is a single-molecule guide RNA.

34. The method of claim 31 wherein the guide RNA is a double-molecule guide RNA.

35. The method of claim 31 wherein the endonuclease is a nickase.

36. The method of claim 31 wherein the endonuclease comprises a mutation corresponding to S pyogenes E762A, HH983AA or D986A.

37. The method of claim 31 wherein the endonuclease is a dead mutant/DNA binding protein.

38. The method of claim 31 wherein the protospacer-like sequence targeted is in a CCR5, CXCR4, KRT5, KRT14, PLEC or COL7A1 gene.

39. The method of claim 31 wherein the protospacer-like sequence is in a chronic granulomatous disease (CGD)-related gene CYBA, CYBB, NCF1, NCF2 or NCF4.

40. The method of claim 31 wherein the protospacer-like sequence targeted is in, or is up to 1000 nucleotides upstream of, a gene encoding B-cell lymphoma/leukemia IIA (BCL11A) protein, an erythroid enhancer of BCL11A or a BCL11A binding site.

41. The method of claim 31 wherein the endonuclease and the guide RNA are introduced to the cell by the same or different recombinant vectors encoding the endonuclease and the guide RNA.

42. The method of claim 31 wherein at least one recombinant vector is a recombinant viral vector.

43. A recombinant vector encoding:

(a) a guide RNA, wherein the guide RNA comprises a DNA-targeting segment complementary to a protospacer-like sequence in the DNA 5' to the PAM sequence NNNNACA; and

(b) s C. jejuni Cas9 endonuclease or an endonuclease with an activity portion at least 90% identical to the activity portion of the C. jejuni Cas9 endonuclease.

44. A recombinant vector encoding:

(a) a guide RNA, wherein the guide RNA comprises a DNA-targeting segment complementary to a protospacer-like sequence in the DNA 5' to the PAM sequence GNNNCNNA or NNNNC; and

(b) a P. multocida Cas9 endonuclease or an endonuclease with an activity portion at least 90% identical to the activity portion the P. multocida Cas9 endonuclease.

45. A recombinant vector encoding:

(a) a guide RNA, wherein the guide RNA comprises a DNA-targeting segment complementary to a protospacer-like sequence in the DNA 5' to the PAM sequence NG; and

(b) a F. novicida Cas9 endonuclease or an endonuclease with an activity portion at least 90% identical to the activity portion of the F. novicida Cas9 endonuclease.

46. A recombinant vector encoding:

(a) a guide RNA, wherein the guide RNA comprises a DNA-targeting segment complementary to a protospacer-like sequence in the DNA 5' to the PAM sequence NNAAAAW; and

(b) a S. thermophilus** Cas9 endonuclease or an endonuclease with an activity portion at least 90% identical to the activity portion of the S. thermophilus** Cas9 endonuclease.

47. A recombinant vector encoding:

(a) a guide RNA, wherein the guide RNA comprises a DNA-targeting segment complementary to a protospacer-like sequence in the DNA 5' to the PAM sequence NGG; and

(b) a L. innocua Cas9 endonuclease or an endonuclease with an activity portion at least 90% identical to the activity portion of the L. innocua Cas9 endonuclease.

48. A recombinant vector encoding:

(a) a guide RNA, wherein the guide RNA comprises a DNA-targeting segment complementary to a protospacer-like sequence in the DNA 5' to the PAM sequence NGG; and

(b) a S. dysgalactiae Cas9 endonuclease or an endonuclease with an activity portion at least 90% identical to the activity portion of the S. dysgalactiae Cas9 endonuclease.

49. The recombinant vector of claim 43, 44, 45, 46, 47 or 48 wherein the recombinant vector is a recombinant viral vector.

50. A modified Cas9 endonuclease comprising one or more mutations corresponding to S. pyogenes mutation E762A, HH983AA or D986A.

51. The modified Cas 9 endonuclease of claim 50 further comprising one or more mutations corresponding to S. pyogenes mutation D10A, H840A, G12A, G17A, N854A, N863A, N982A or A984A.

52. A method for manipulating DNA in a cell, comprising contacting the DNA with a Cas9 ortholog-guide RNA complex, wherein the complex comprises:

(a) a Cas9 endonuclease heterologous to the cell and

(b) a cognate guide RNA of the Cas9 endonuclease comprising a tracrRNA set out in Supplementary Table S5 or a guide RNA comprising a tracrRNA at least 80% identical to a cognate tracrRNA set out in Supplementary Table S5 over at least 20 nucleotides.

53. The method of claim 52 wherein the cell is a bacterial cell, a fungal cell, an archaea cell, a plant cell or an animal cell.

54. The method of claim 52 wherein the guide RNA is a single-molecule guide RNA.

55. The method of claim 52 wherein the guide RNA is a double-molecule guide RNA.

56. The method of claim 52 wherein the endonuclease is a nickase.

57. The method of claim 52 wherein the endonuclease comprises a mutation corresponding to S pyogenes mutations E762, HH983AA or D986A.

58. The method of claim 52 wherein the endonuclease is a dead mutant/DNA binding protein.

59. The method of claim 52 wherein the protospacer-like sequence targeted is in a CCR5, CXCR4, KRT5, KRT14, PLEC or COL7A1 gene or a sequence up to 1000 nucleotides upstream of the gene.

60. The method of claim 52 wherein the protospacer-like sequence is in a chronic granulomatous disease (CGD)-related gene CYBA, CYBB, NCF1, NCF2 or NCF4 or a sequence up to 1000 nucleotides upstream of the gene.

61. The method of claim 52 wherein the protospacer-like sequence targeted is in, or is up to 1000 nucleotides upstream of, a gene encoding B-cell lymphoma/leukemia IIA (BCL11A) protein, an erythroid enhancer of BCL11A or a BCL11A binding site.

62. The method of claim 52 wherein the enodnuclease and the guide RNA are introduced to the cell by the same or different recombinant vectors encoding the endonuclease and the guide RNA.

63. The method of claim 52 wherein at least one recombinant vector is a recombinant viral vector.

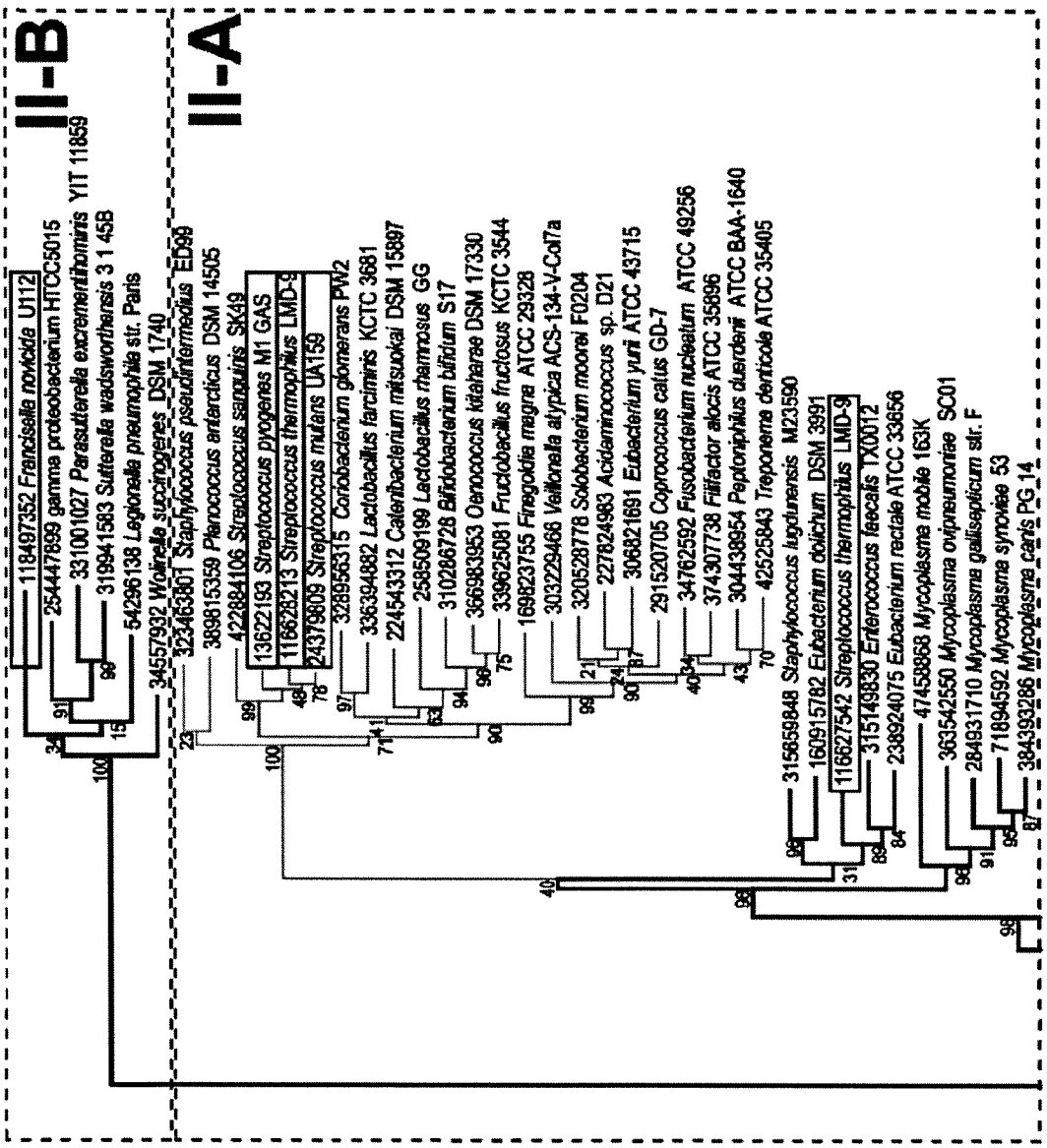

II-B
II-A
118497352 Francisella novicida U112
254447899 gamma proteobacterium HTCC5015
331001027 Parasutterella excrementihominis YIT 11859
319941583 Sutterella wadsworthensis 3 1 45B
54296138 Legionella pneumophila str. Paris
34577932 Wolinella succinogenes DSM 1740
323463801 Staphylococcus pseudintermedius ED99
389815359 Planococcus antarcticus DSM 14505
422884106 Streptococcus sanguinis SK49
13622193 Streptococcus pyogenes M1 GAS
116628213 Streptococcus thermophilus LMD-9
24379809 Streptococcus mutans UA159
328956315 Coriobacterium glomerans PW2
336394882 Lactobacillus farciminis KCTC 3681
224543312 Caterribacterium mitsuokai DSM 15897
258509199 Lactobacillus rhamnosus GG
310286728 Bifidobacterium bifidum S17
366983953 Oenococcus kitaharae DSM 17330
33962508 Fructobacillus fructosus KCTC 3544
169823755 Finegoldia magna ATCC 29328
303229466 Veillonella atypica ACS-134-V-Col7a
320528778 Solobacterium moorei F0204
227824983 Acidaminococcus sp. D21
306821691 Eubacterium yurii ATCC 43715
291520705 Coprococcus catus GD-7
34762592 Fusobacterium nucleatum ATCC 49256
374307738 Filifactor alocis ATCC 35896
304438954 Peptoniphilus duerdenii ATCC BAA-1640
42525843 Treponema denticole ATCC 35405
315659848 Staphylococcus lugdunensis M23590
160915782 Eubacterium dolichum DSM 3991
116627542 Streptococcus thermophilus LMD-9
315149830 Enterococcus faecalis TX0012
238924075 Eubacterium rectale ATCC 33656
47458868 Mycoplasma mobile 163K
363542550 Mycoplasma ovipneumoniae SC01
284931710 Mycoplasma gallisepticum str. F
71894592 Mycoplasma synoviae 53
384393286 Mycoplasma canis PG 14

Figure. 1A

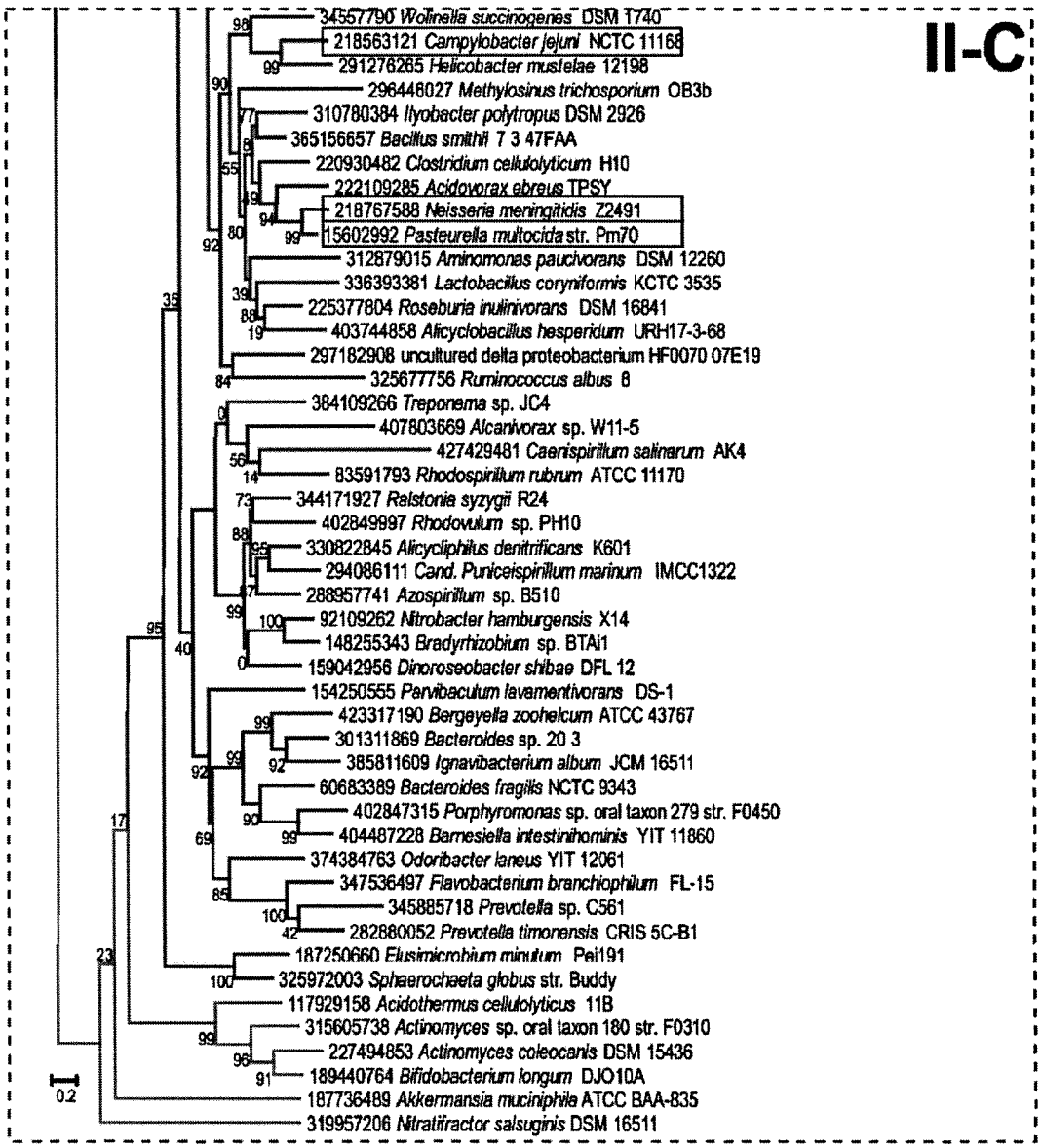


Figure. 1A
(Continued)

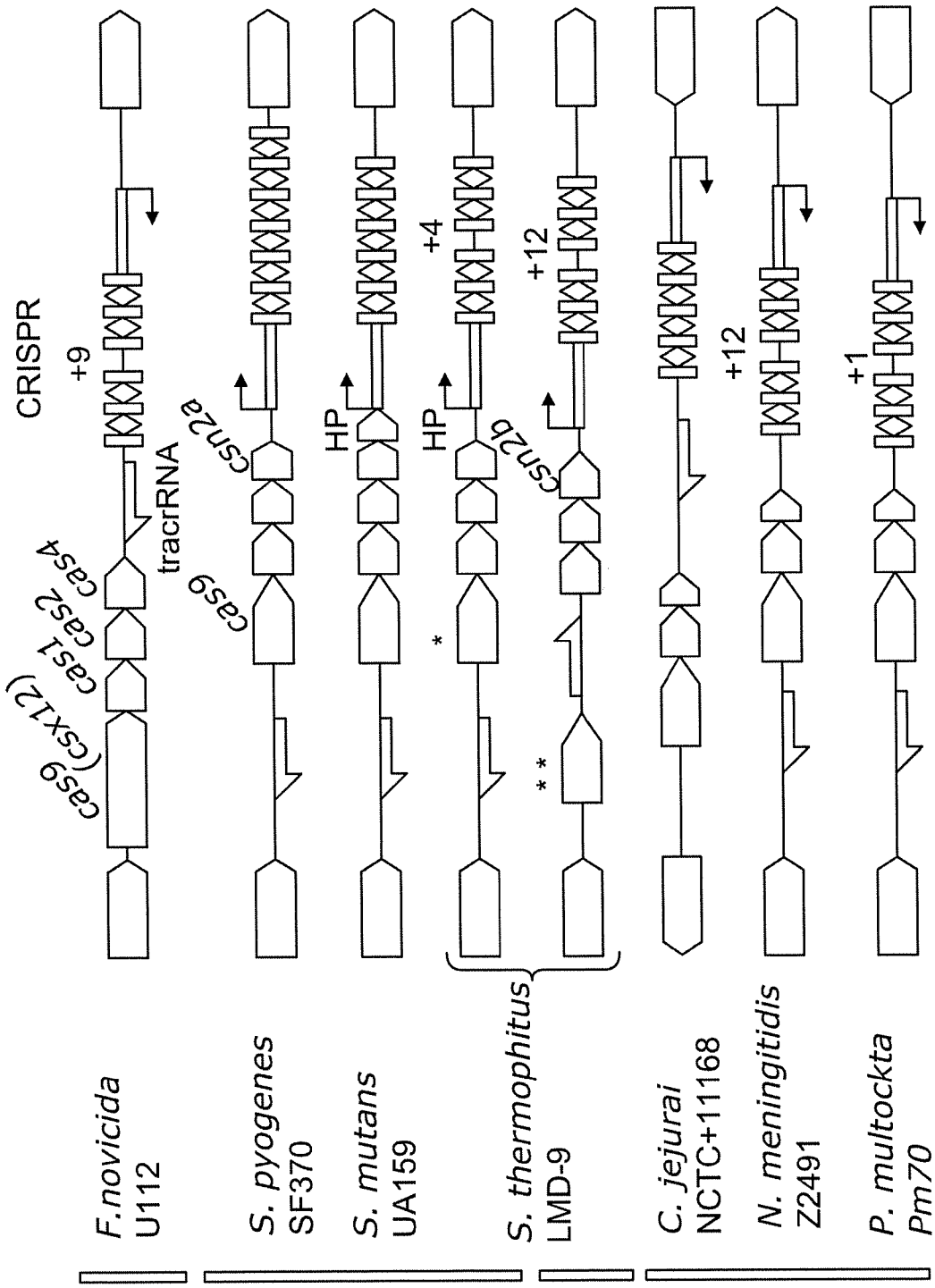

CRISPR
+9
tracrRNA
cas9 (csx12)
cas1
cas2
cas4
cas9
csn2a
HP
+4
csn2b
+12
+12
+1
F.novicida U112
S. pyogenes SF370
S. mutans UA159
S. thermophitus LMD-9
C. jejurai NCTC+11168
N. meningitidis Z2491
P. multockta Pm70

Figure 1B

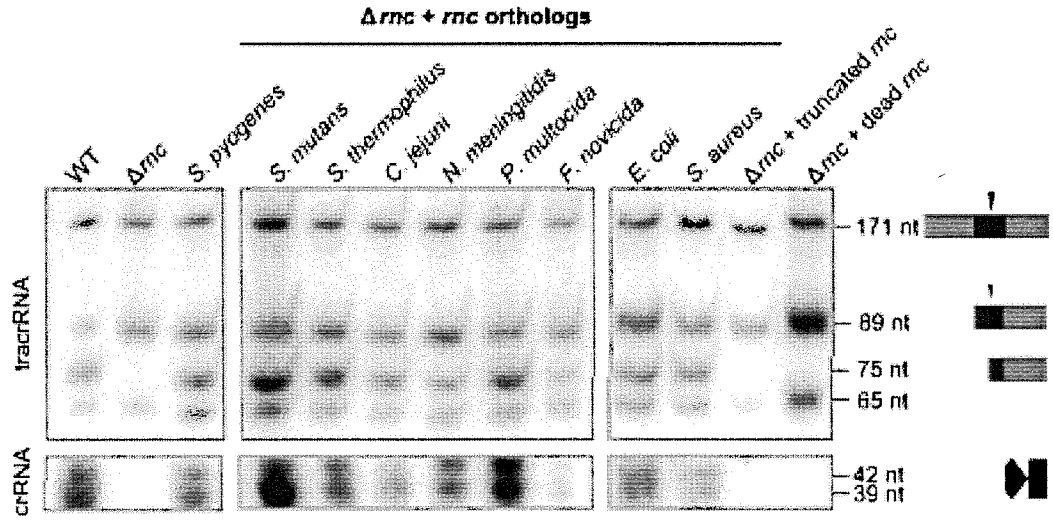

Δrnc + rnc orthologs
WT
Δrnc
S. pyogenes
S. mutans
S. thermophilus
C. jejuni
N. meningitidis
P. multocida
F. novicida
E. coli
S. aureus
Δrnc + truncated rnc
Δrnc + dead rnc
tracrRNA
crRNA
171 nt
89 nt
75 nt
65 nt
42 nt
39 nt

Figure 2

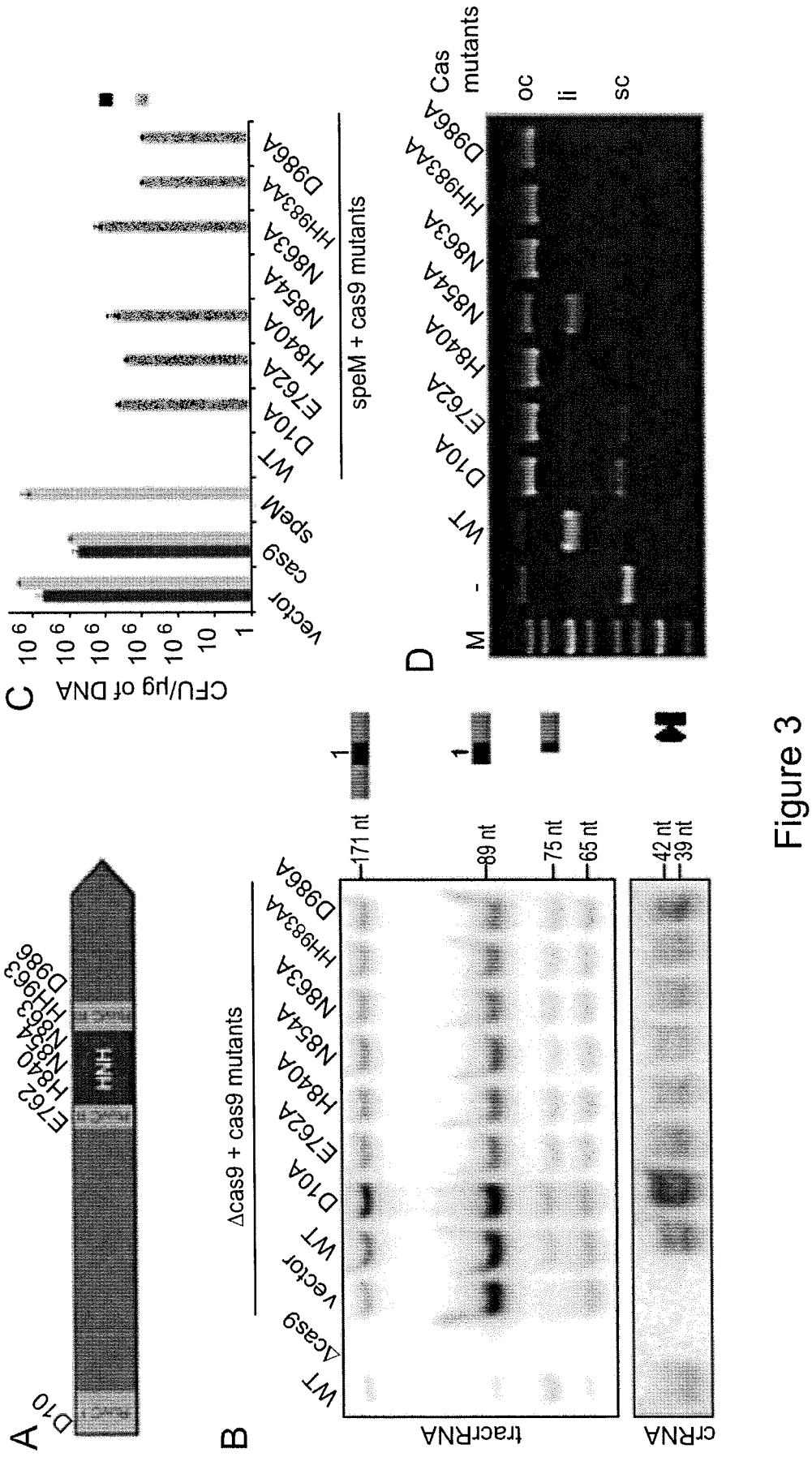

A
D10
HNH
E762 H840 N854 N863 H983 D986
B
ΔCas9 + cas9 mutants
WT
ΔCas9
vector
WT
D10A
E762A
H840A
N854A
N863A
HH983AA
D986A
171 nt
89 nt
75 nt
65 nt
tracrRNA
42 nt
39 nt
crRNA
C
CFU/µg of DNA
vector
cas9
speM
WT
D10A
E762A
H840A
N854A
N863A
HH983AA
D986A
speM + cas9 mutants
D
Cas mutants
oc
li
sc
M
-
WT
D10A
E762A
H840A
N854A
N863A
HH983AA
D986A

Figure 3

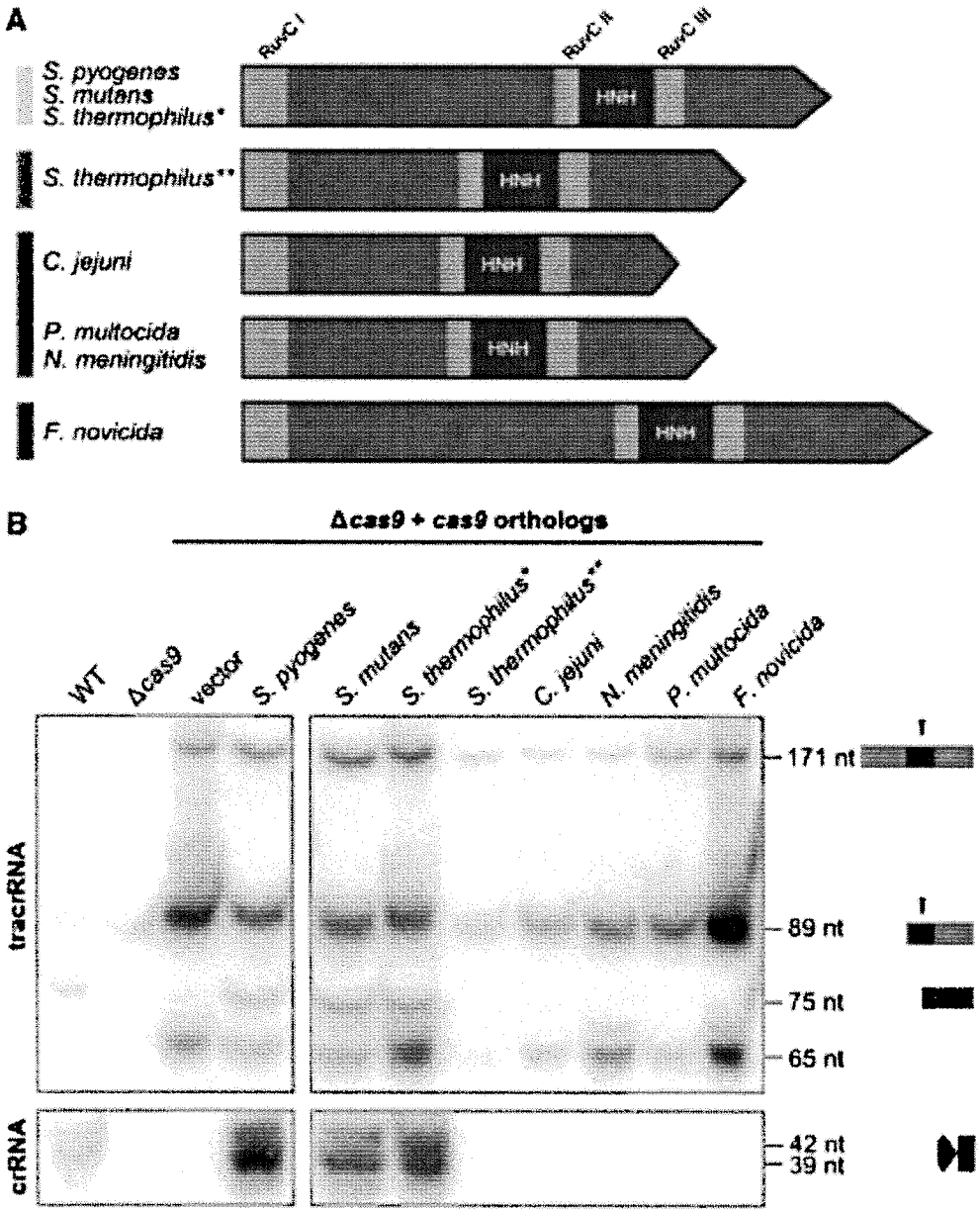
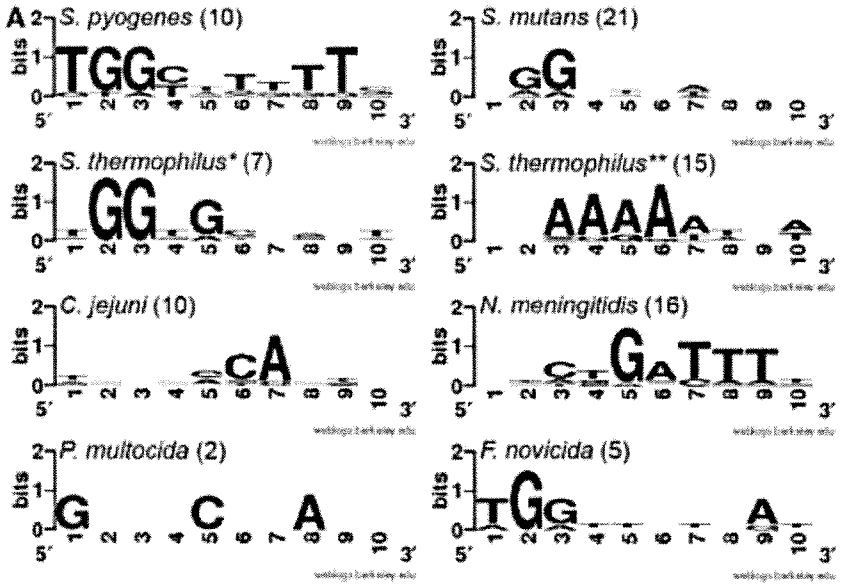

A
RuvC I
RuvC II
RuvC III
HNH
S. pyogenes
S. mutans
S. thermophilus*
S. thermophilus**
C. jejuni
P. multocida
N. meningitidis
F. novicida
B
Δcas9 + cas9 orthologs
WT
Δcas9
vector
S. pyogenes
S. mutans
S. thermophilus*
S. thermophilus**
C. jejuni
N. meningitidis
P. multocida
F. novicida
tracrRNA
crRNA
171 nt
89 nt
75 nt
65 nt
42 nt
39 nt

Figure 4



B

Species	substrates		PAM
	PAM–	PAM+	
S. pyogenes	GTTTGATTGG	TGGTATTGGG	NGG
S. mutans	GTTTGATTGG	TGGTATTGGG	NGG
*S. thermophilus**	GTTTGATTGG	TGGTGTTGGG	NGGNG
*S. thermophilus***	TGGTATTGGG	GGAAAATGGG	NNAAAAW
C. jejuni	TGGTATTGGG	AGAAACAGGG	NNNNACA
N. meningitidis	TGGTATTGGG	GGGTGATTGG	NNNNGATT
P. multocida	TGGTATTGGG	GGGTCATAGG	GNNNCNNA
F. novicida	GTTTGATTGG	TGGTATTGGG	NG

C

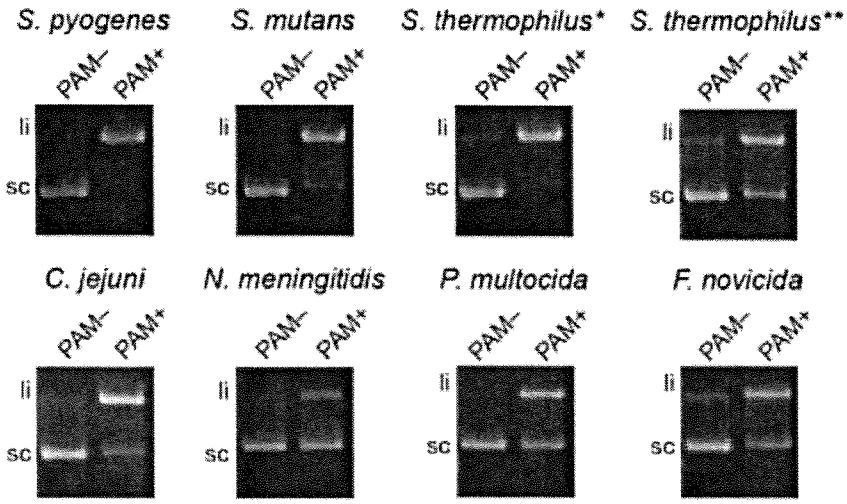


Figure 5

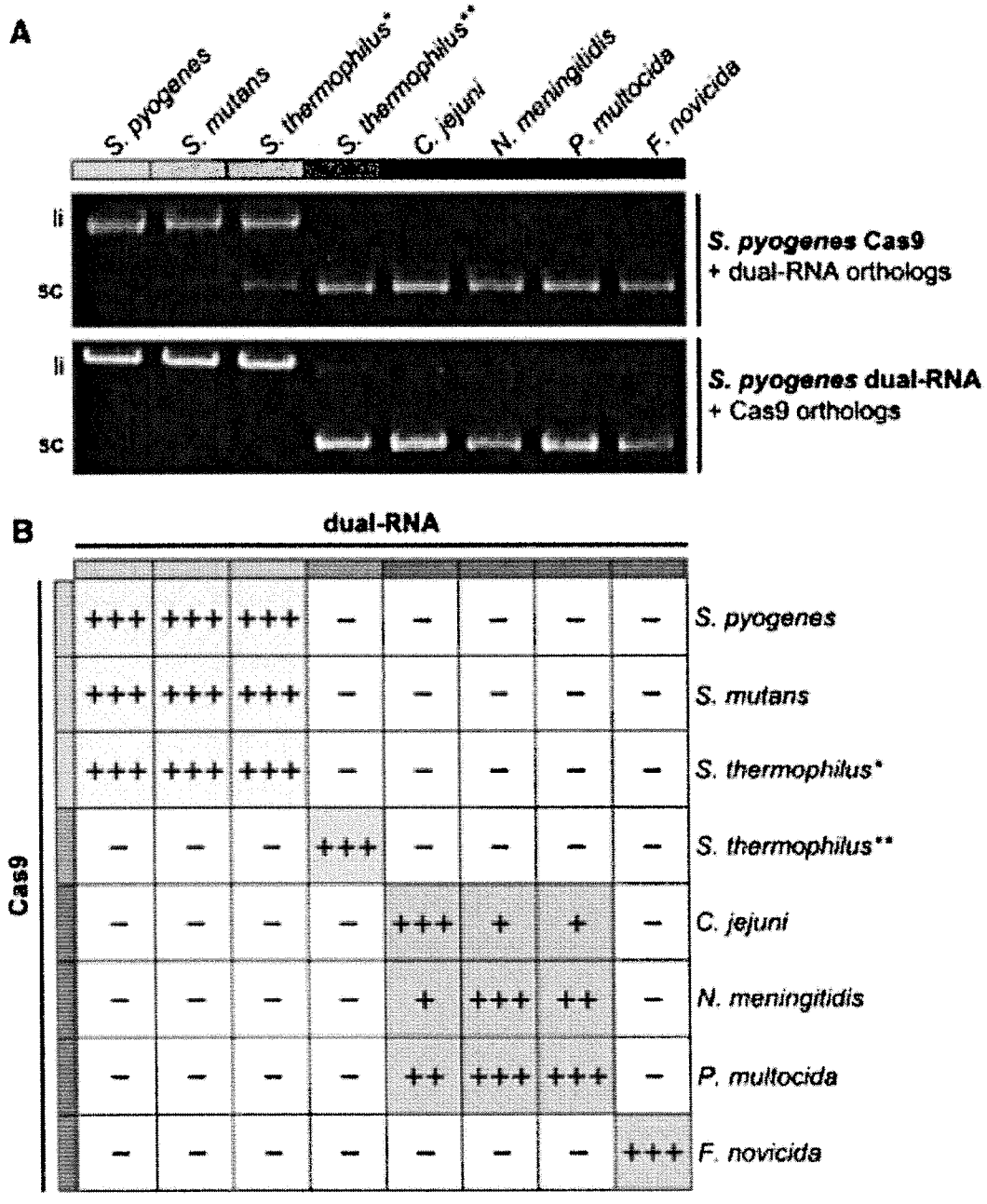

A
S. pyogenes
S. mutans
S. thermophilus*
S. thermophilus**
C. jejuni
N. meningitidis
P. multocida
F. novicida
li
sc
S. pyogenes Cas9
+ dual-RNA orthologs
li
sc
S. pyogenes dual-RNA
+ Cas9 orthologs
B
dual-RNA
Cas9
+++ +++ +++ – – – – – S. pyogenes
+++ +++ +++ – – – – – S. mutans
+++ +++ +++ – – – – – S. thermophilus*
– – – +++ – – – – S. thermophilus**
– – – – +++ + + – C. jejuni
– – – – + +++ ++ – N. meningitidis
– – – – ++ +++ +++ – P. multocida
– – – – – – – +++ F. novicida

Figure 6

A

Strain*	GI accession number[b]	Amino acids (including purification-tag)	Molecular weight [kDa] (including purification-tag)	ε [M^{-1} cm^{-1}]
S. pyogenes SF370	13622193	1368 (1392)	158.4 (161.2)	110470
S. mutans UA159	24379809	1345 (1369)	156.6 (159.4)	114190
*S. thermophilus** LMD-9	116628213	1388 (1412)	161.0 (163.8)	128390
*S. thermophilus*** LMD-9	116627542	1121 (1145)	129.4 (132.2)	103860
C. jejuni NCTC 11168	218563121	984 (1008)	114.9 (117.7)	78640
N. meningitidis A Z2491	218767586	1082 (1106)	124.3 (127.1)	101950
P. multocida Pm70	15602992	1056 (1080)	121.8 (124.6)	121460
F. novicida U112	118497352	1629 (1653)	190.4 (193.2)	189080

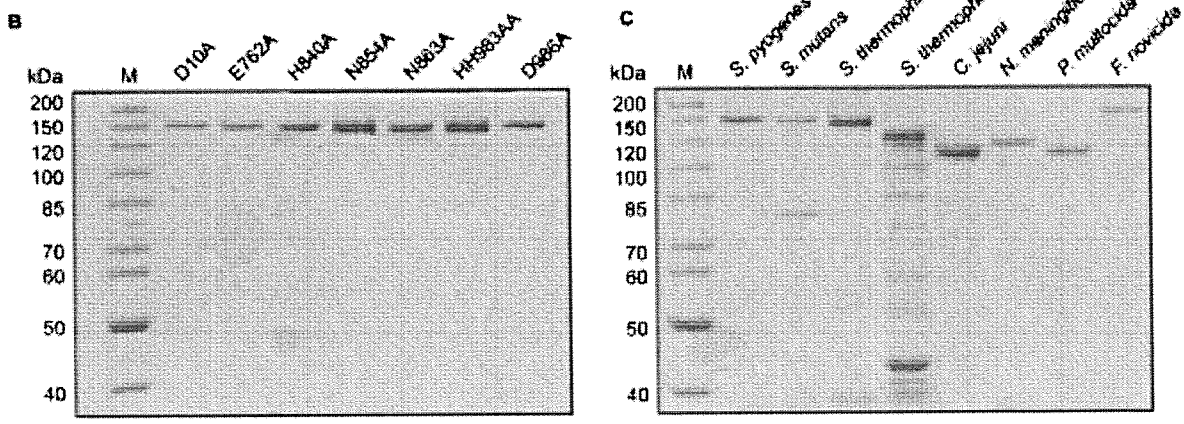


Figure S1

```
227494853  Actinomyces coleocanis DSM 15436#    NYRIGIDVGLNSIGFCAV----------D----------EVDQHD----------
328956315  Coriobacterium glomerans PW2         DYSIGLDMGTSSVGWAVT----------D----------ERGTLA----------
227824983  Acidaminococcus sp. D21              MYYLGLDIGTNSVGYAVT----------D----------PSYHLL----------
303229466  Veillonella atypica ACS-134-V-Col7a  DYFVGLDIGTNSVGWAVT----------N----------TSYELL----------
34762592   Fusobacterium nucleatum ATCC 49256   DYYLGFDIGTNSVGWCVT----------D----------LDYNVL----------
374307738  Filifactor alocis ATCC 35896         EYYLGLDVGTNSVGWAVT----------D----------SQYNLC----------
320528778  Solobacterium moorei F0204           NYYLGLDVGTSSVGWAVT----------D----------TDYNVL----------
291520705  Coprococcus catus GD-7               EYFLGLDMGTGSLGWAVT----------D----------STYQVM----------
42525843   Treponema denticola ATCC 35405       DYFLGLDVGTGSVGWAVT----------D----------TDYKLL----------
304438954  Peptoniphilus duerdenii ATCC BAA-1640 EYYIGLDIGTASVGWAVT----------D----------ESYNIP----------
224543312  Catenibacterium mitsuokai DSM 15897  DYCIGLDLGTGSVGWAVV----------D----------MNHRLM----------
116628213  Streptococcus thermophilus LMD-9     PYSIGLDIGTNSVGWAVT--TDNYKVPSKKMKVLGNTSK----------
24379809   Streptococcus mutans UA159           PYSIGLDIGTNSVGWAVV--TDDYKVPAKKMKVLGNTDK----------
13622193   Streptococcus pyogenes SF370         KYSIGLDIGTNSVGWAVI--TDEYKVPSKKFKVLGNTDR----------
310286728  Bifidobacterium bifidum S17          DYAISLDIGNASVGWSAF----------T----------PNYRLV----------
366983953  Oenococcus kitaharae DSM 17330       DYSVGLDIGTSSVGWAAI----------D----------NKYHLI----------
422884106  Streptococcus sanguinis SK49         NYSIGLDIGVSSVGYAVV----------T----------EDYRVPAFKFKVLGNTEK-
339625081  Fructobacillus fructosus KCTC 3544   VYDVGLDIGTGSVGWVAL----------D----------ENGKLA----------
306821691  Eubacterium yurii ATCC 43715         QYYIGLDVGTNSVGWAVT----------D----------TSYNLL----------
336394882  Lactobacillus farciminis KCTC 3681   PYNIGLDIGTSSVGWAVT----------N----------DNYDLL----------
323463801  Staphylococcus pseudintermedius ED99 PYTLSLDIGTGSVGYACM----------D----------KGFNVL----------
389815359  Planococcus antarcticus DSM 14505    NYTIGLDIGVASVGWVCI----------D----------ENYKIL----------
258509199  Lactobacillus rhamnosus GG           PYGIGLDIGSNSIGFAVV----------D----------ANSHLL----------
169823755  Finegoldia magna ATCC 29328          KYYIGLDVGTNSVGWAVT----------D----------EFYNIL----------
Jnet       227501312                            --EEEEE----EEEEE----------E----------EE----------
47458868   Mycoplasma mobile 163K               KVVLGLDLGIASVGWCLT----------D----------ISQKED----------
284931710  Mycoplasma gallisepticum str. F      EVTIGLDLGVGSVGWATV----------D----------DN----------
71894592   Mycoplasma synoviae 53               KIVIGFDLGVASVGWSIV----------NA----------
363542550  Mycoplasma ovipneumoniae SC01        NITIGFDLGIASIGWAII----------DS----------
384393286  Mycoplasma canis PG 14               KVTLGFDLGIASVGWAIV----------DS----------
238924075  Eubacterium rectale ATCC 33656       KYILALDIGIASVGWAIL----------DK----------
315149830  Enterococcus faecalis TX0012         MYSIGLDLGISSVGWSVI----------DE----------
116627542  Streptococcus thermophilus LMD-9     DLVLGLDIGIGSVGVGIL----------NK----------
31565848   Staphylococcus lugdunensis M23590    KFTIGLDIGITSVGYGLI----------DY----------
160915782  Eubacterium dolichum DSM 3991        RLVLGLDIGITSVGFGII----------DL----------
325677756  Ruminococcus albus 8                 NYYLGLDVGIGSIGWAVI----------NI----------
225377804  Roseburia inulinivorans DSM 16841    QYRIGLDIGITSVGWAVL----------QNNSQD----------
336393381  Lactobacillus coryniformis KCTC 3535 GYRIGLDVGITSTGYAVL----------KTDKNG----------
310780384  Ilyobacter polytropus DSM 2926       KYSIGLDIGIASVGWSVI----------NK----------
301311869  Bacteroides sp. 20 3                 KKIVGLDLGTNSIGWALI----------N----------AYINKE----------
385811609  Ignavibacterium album JCM 16511      KKVLGLDIGTNSIGCALL----------S----------LPKSIQDY----------
60683389   Bacteroides fragilis NCTC 9343       KRILGLDLGTNSIGWALV----------N----------EAENKD----------
319957206  Nitratifractor salsuginis DSM 16511  KKILGVDLGITSIGYAIL----------QET----------GK----------
187250660  Elusimicrobium minutum Pei191        PYRIGLDLGVGSIGFAIVS-MEE----------NDGNVL----------
325972003  Sphaerochaeta globus str. Buddy      PYSIGLDLGVGSIGVAVA----------A----------YDPIKK----------
296446027  Methylosinus trichosporium OB3b      MRVLGLDAGIASLGWALIE-IEE----------SNRGEL----------
```

Figure S2

Figure S2 (Continued)

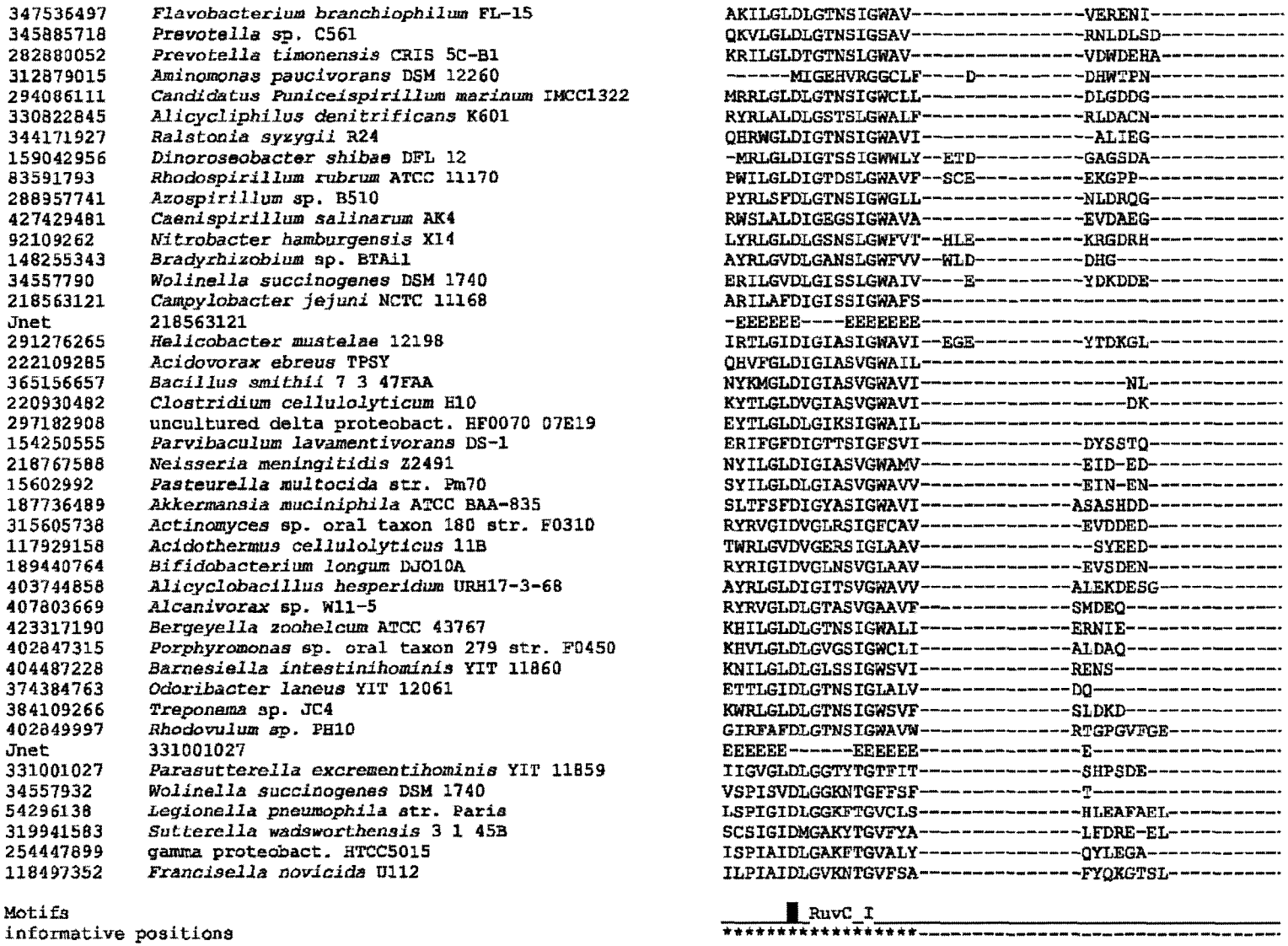


Figure S2 (Continued)

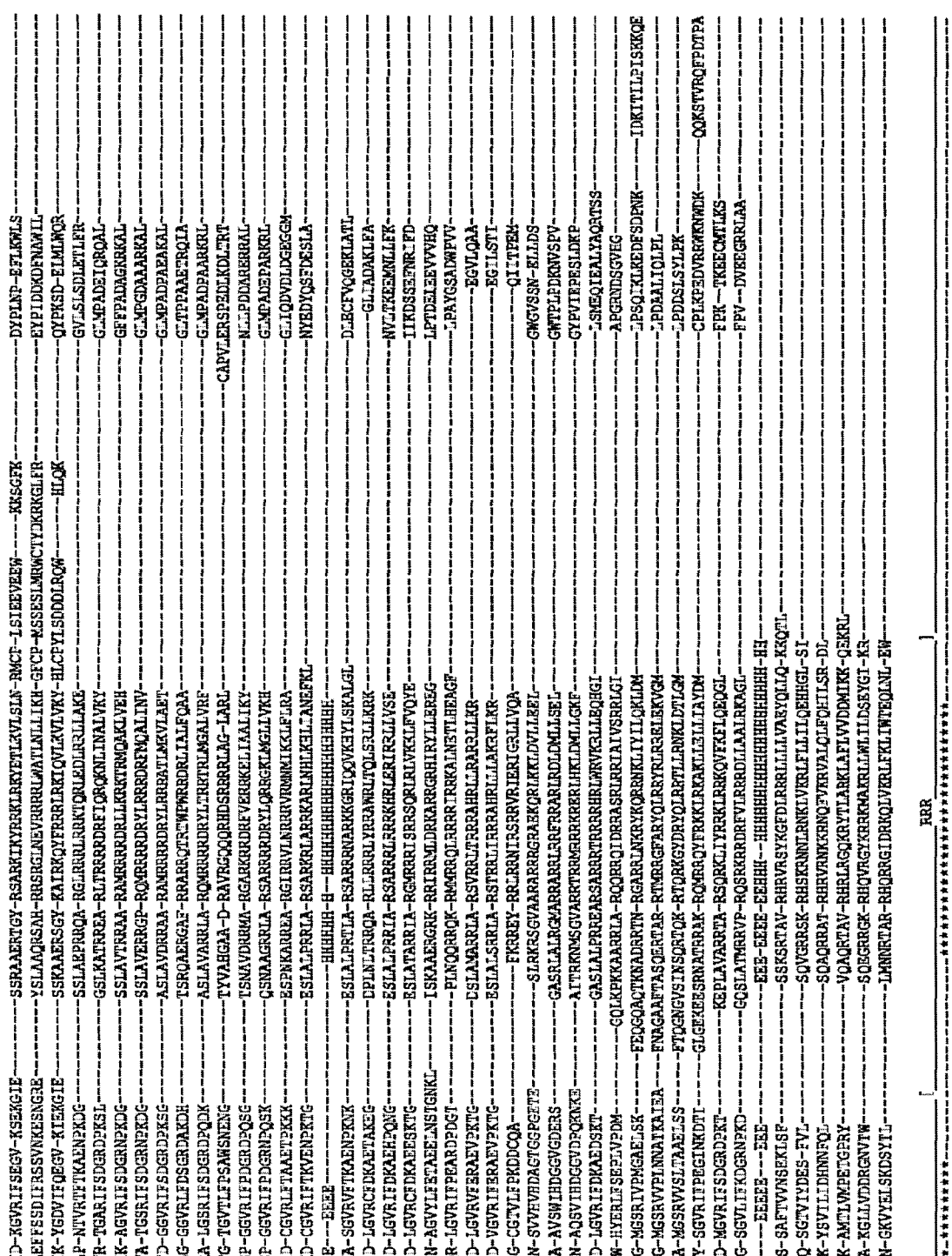


Figure S2 (Continued)

Figure S2 (Continued)

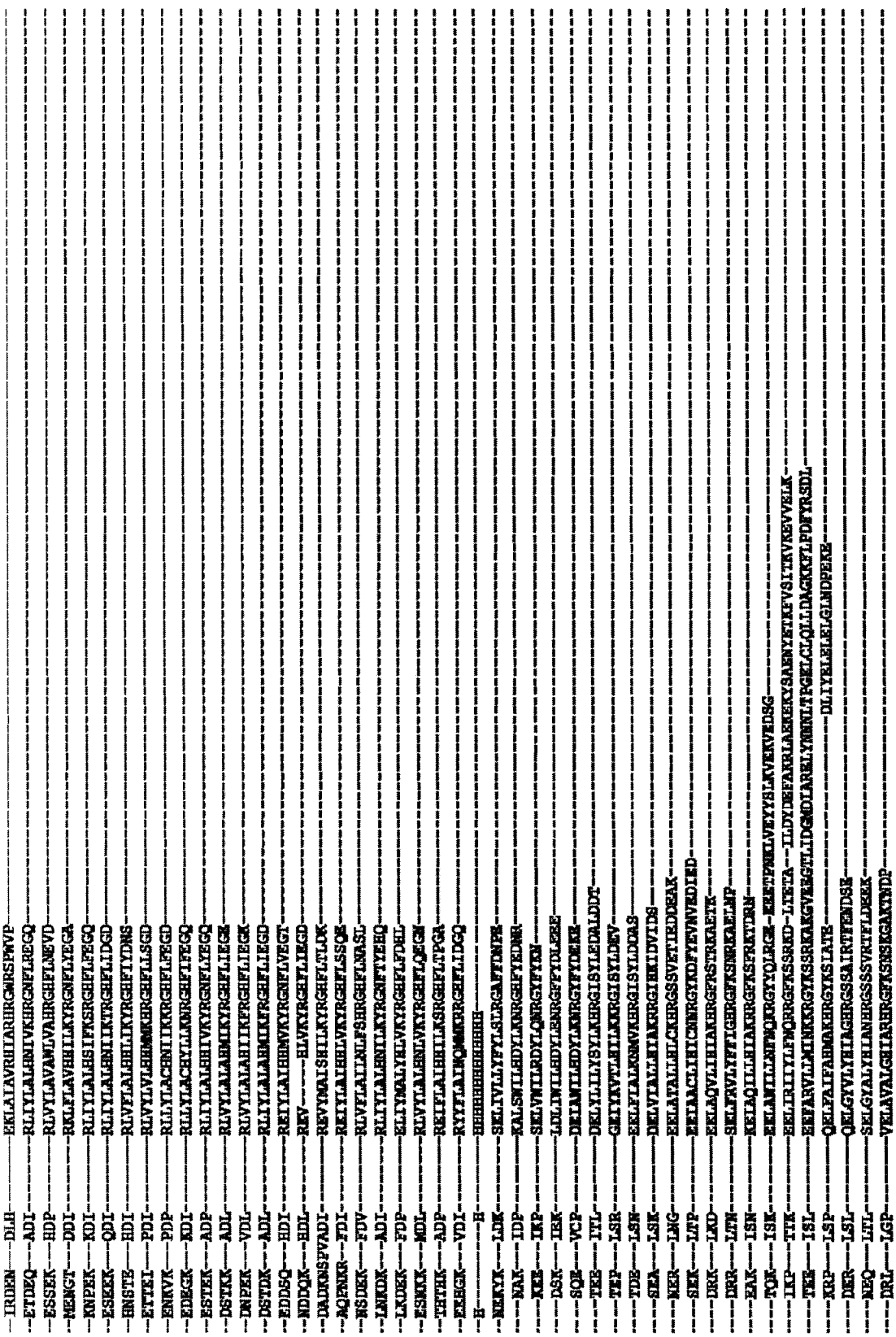


Figure S2 (Continued)

Figure S2 (Continued)

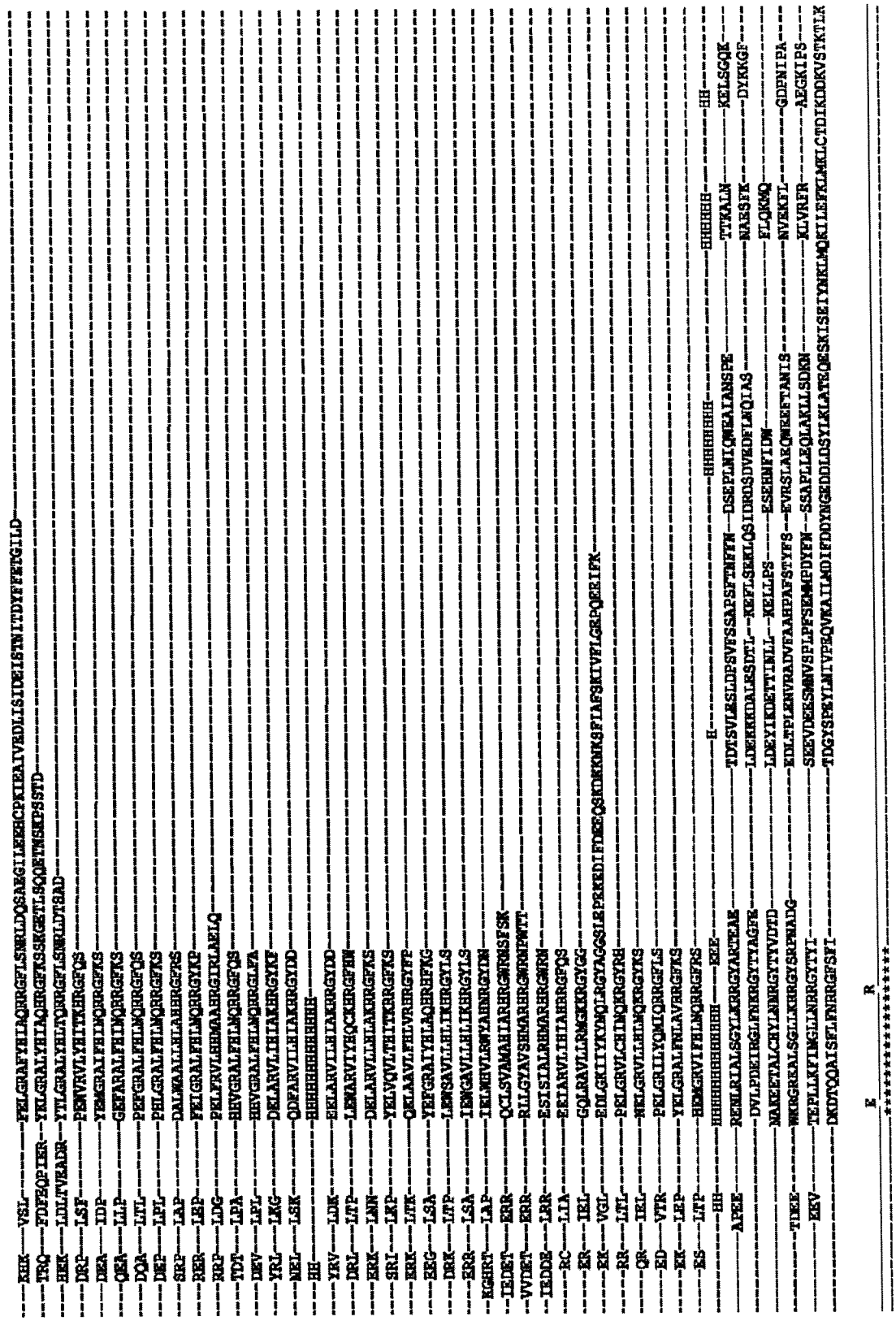


Figure S2 (Continued)

Figure S2 (Continued)

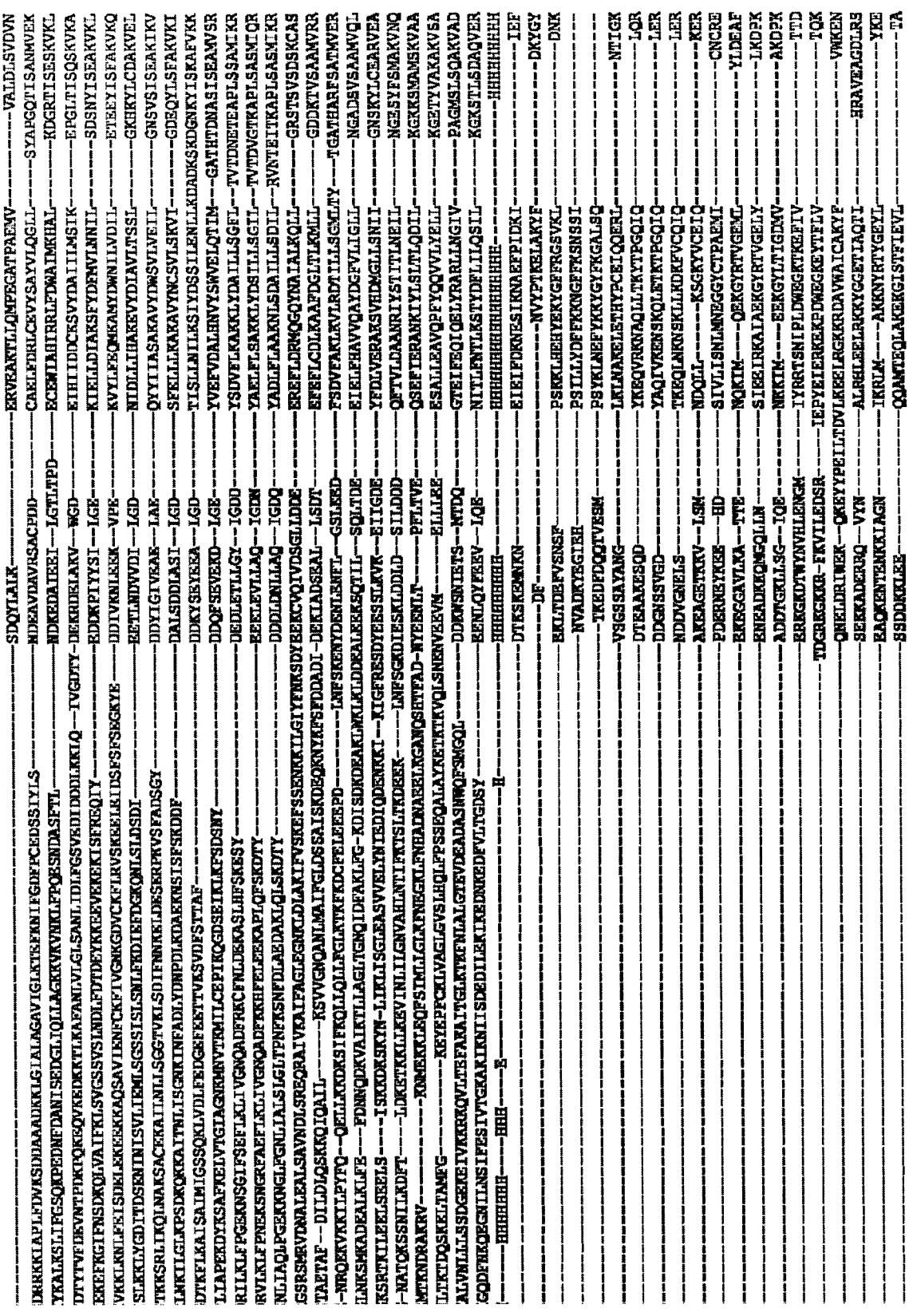


Figure S2 (Continued)

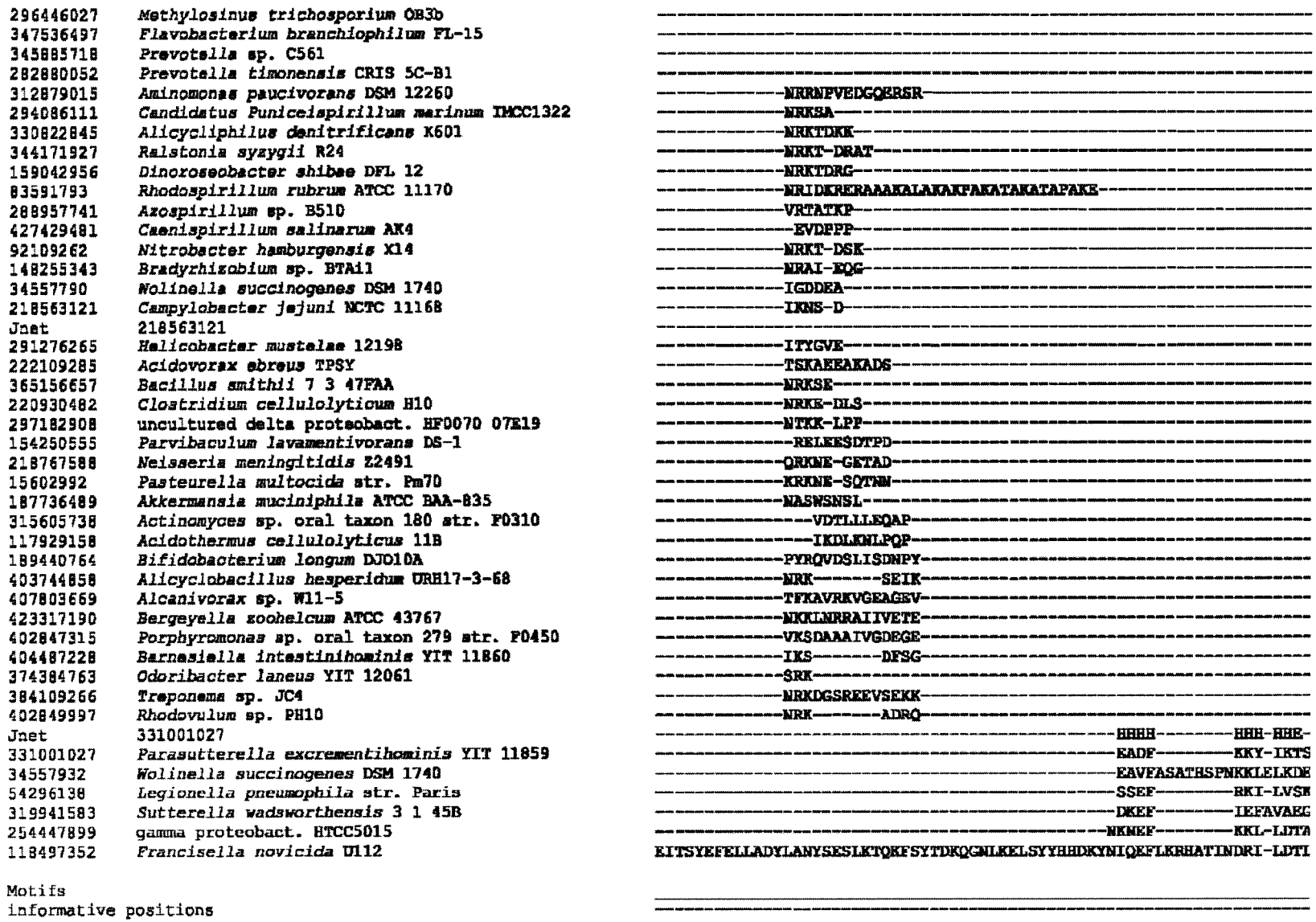


Figure S2 (Continued)

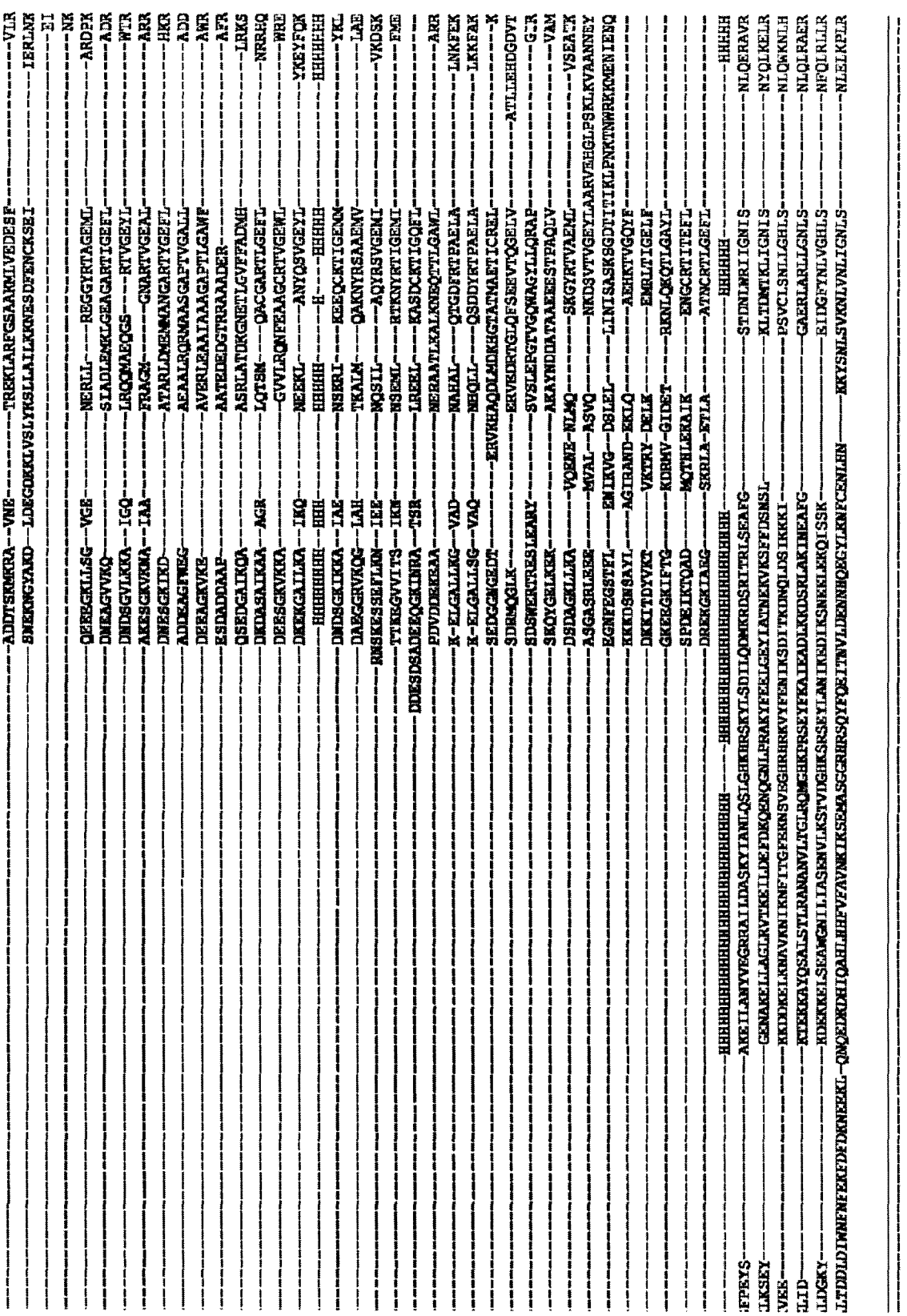


Figure S2 (Continued)

Figure S2 (Continued)

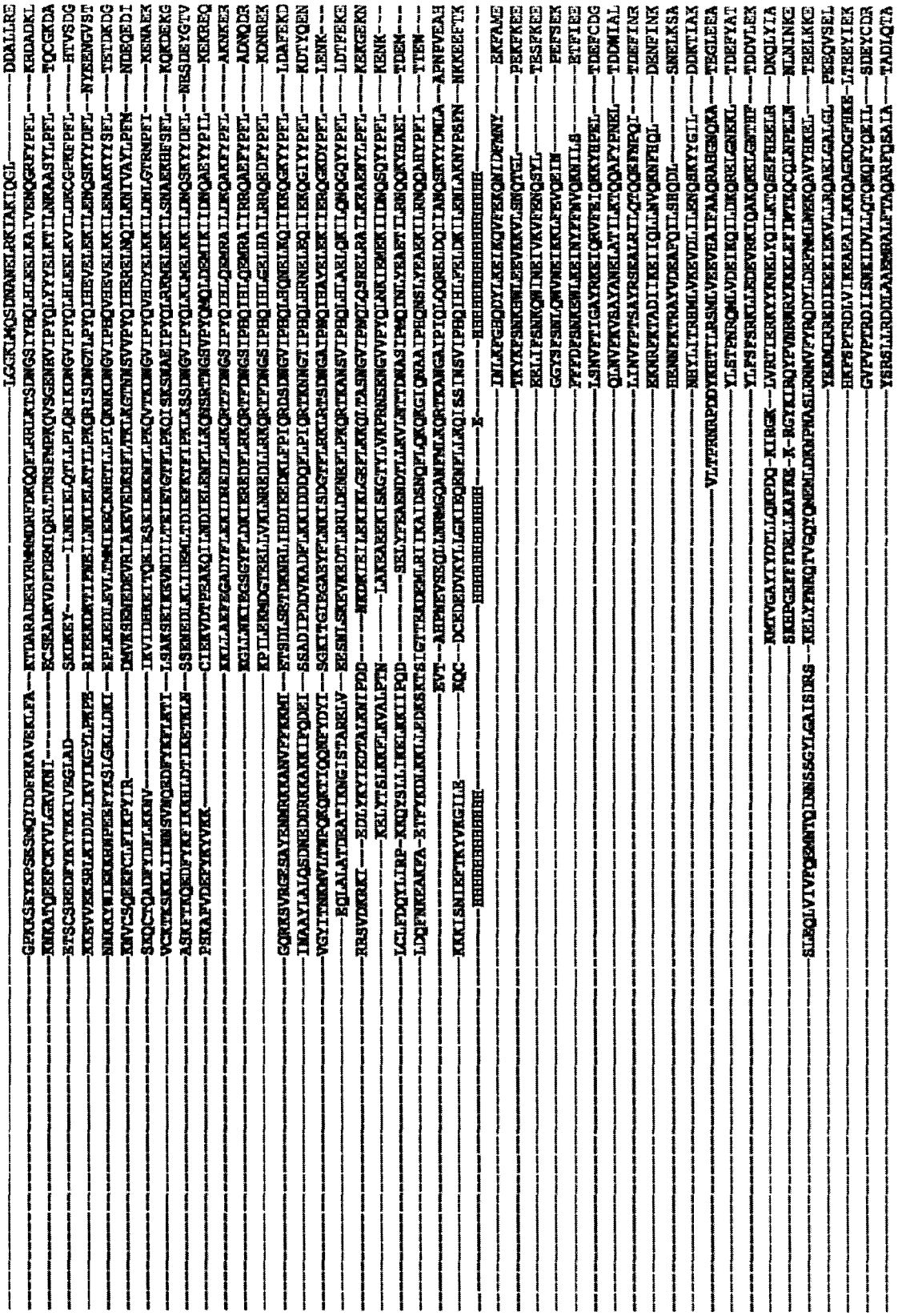


Figure S2 (Continued)

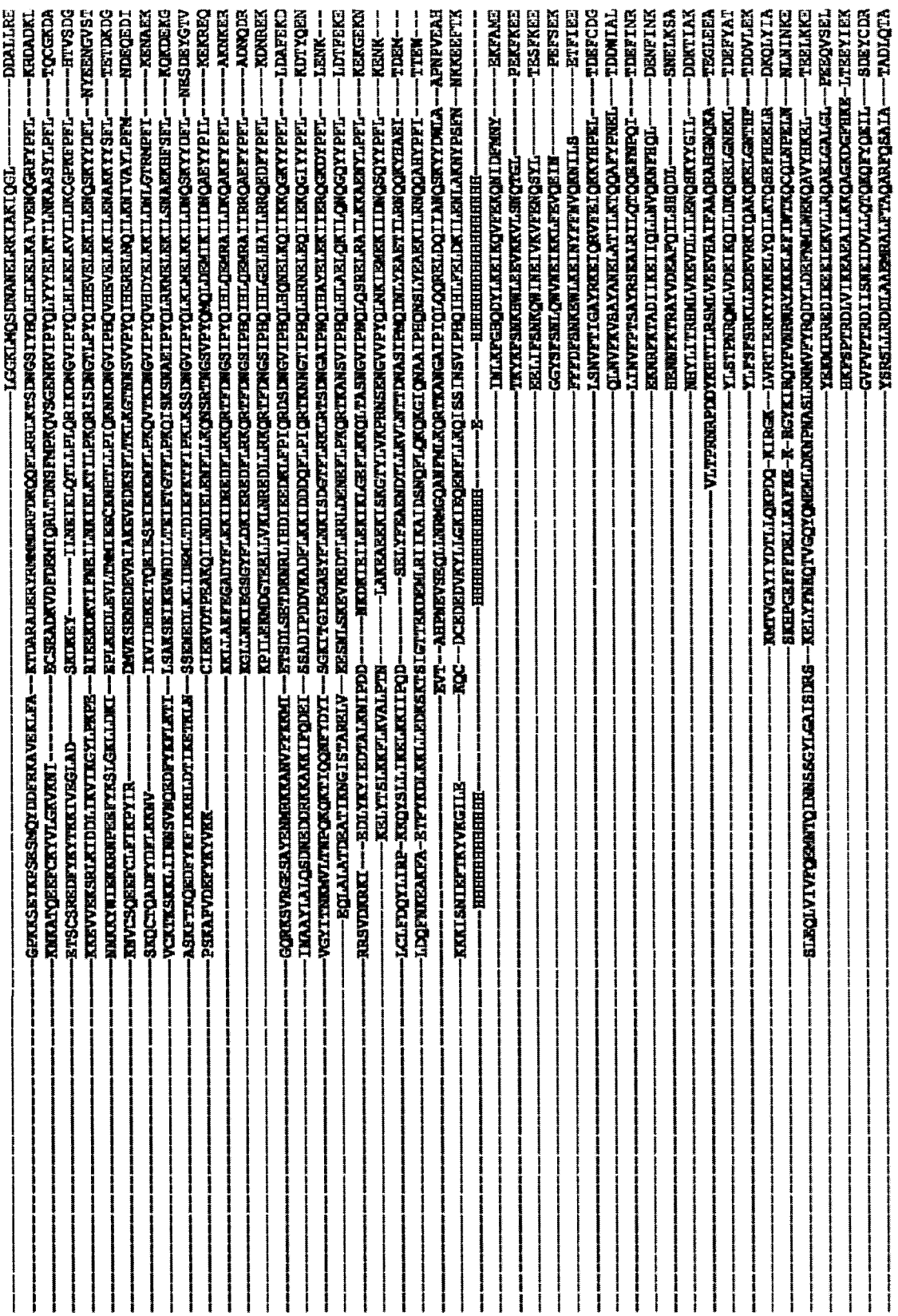


Figure S2 (Continued)

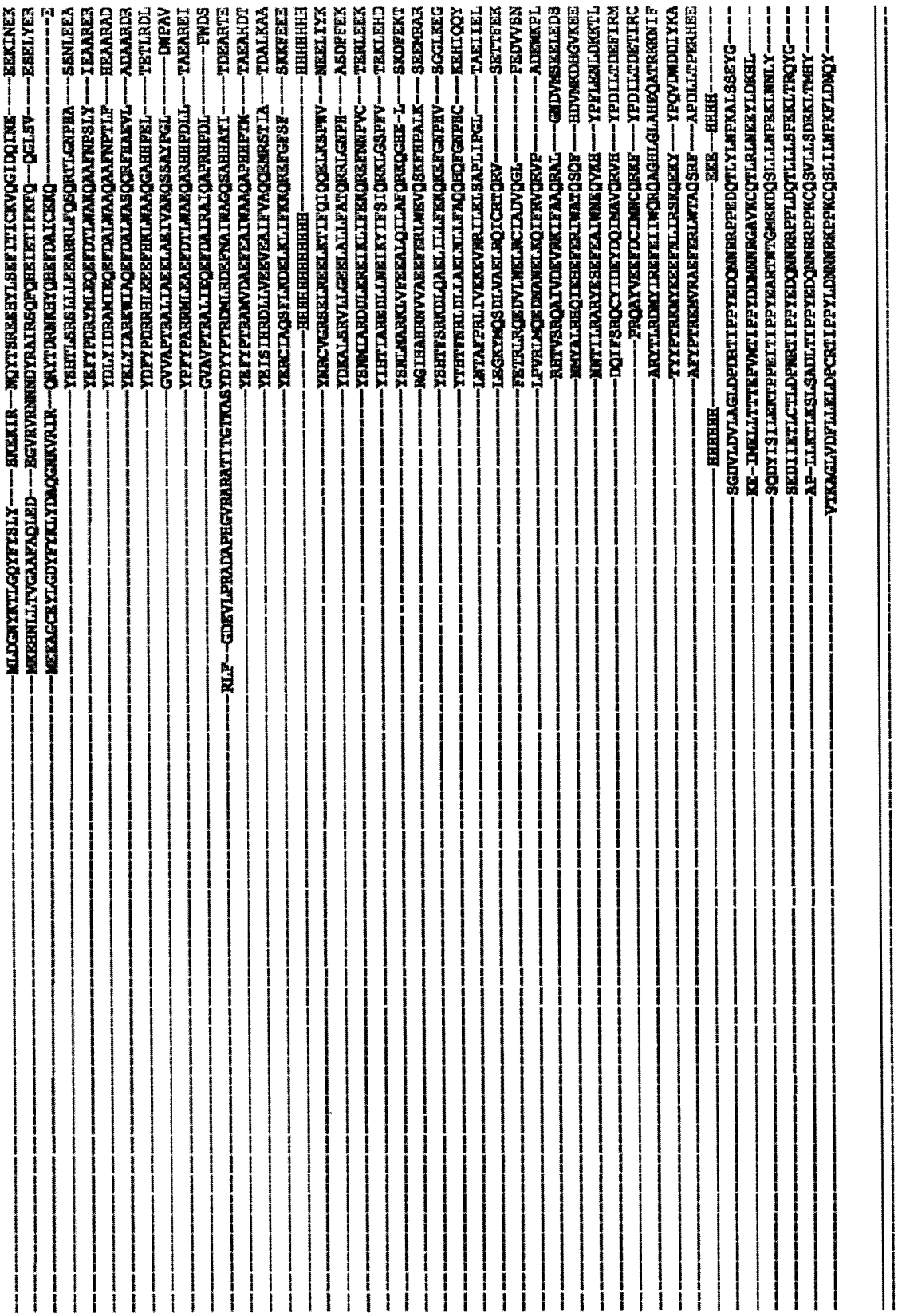


Figure S2 (Continued)

Figure S2 (Continued)

Figure S2 (Continued)

Figure S2 (Continued)

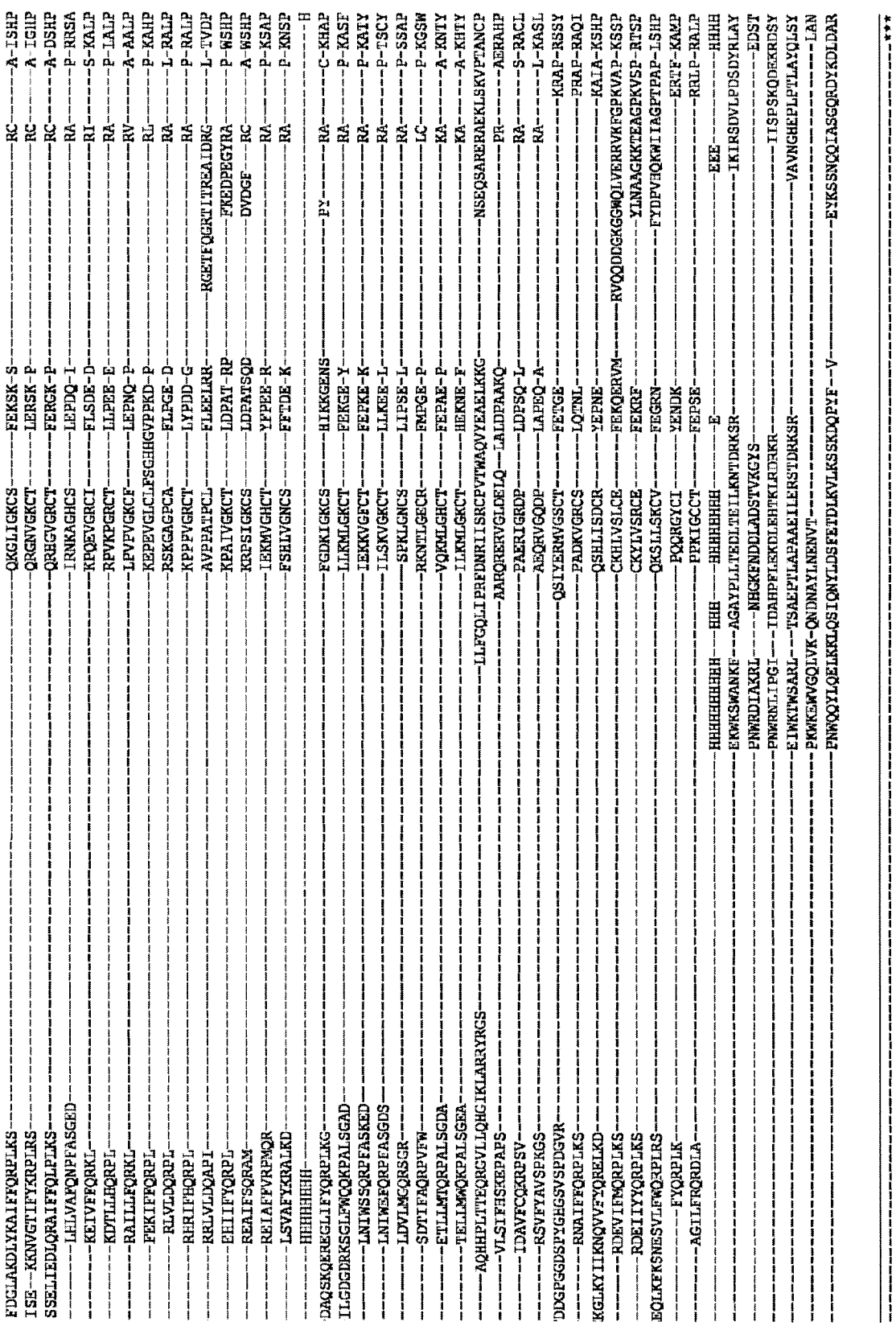


Figure S2 (Continued)

Figure S2 (Continued)

Figure S2 (Continued)

Figure S2 (Continued)

Figure S2 (Continued)

```
227494853 Actinomyces coleocanis DSM 15436#  ----------AEK----LTEGTAAEV----------EVAEFLQNL----------
328956315 Coriobacterium glomerans PW2       ----------IILW----NTLF-EDRK----------IISQRLKEEYGSR----LS----------
227824983 Acidaminococcus sp. D21            ----------IVER----MTYS-DDTK----------RVRLWLNNNYGT----LT----------
303229466 Veillonella atypica ACS-134-V-Col7a ----------IITD----ITIFGESKK----------MIRQTLRNKFGSQ----LN----------
34762592  Fusobacterium nucleatum ATCC 49256 ----------SILW----KCLYGDKK----------IFEKKIKNEYGDI----LN----------
374307738 Filifactor alocis ATCC 35896       ----------IIKW----ITIYGNDKE----------MLKRVIRANYSNQ----LT----------
320528778 Solobacterium moorei F0204         ----------IIKG----ILVYSDDRS----------MIRRWLKNNIKG----LS----------
291520705 Coprococcus catus GD-7             ----------IVLN----VVLFGDDKK----------LLKQRLSKMYPN----LT----------
42525843  Treponema denticola ATCC 35405      ----------IIRWAT--IYDEGEGKT----------IIKTKIKAEYGKY----CS----------
304438954 Peptoniphilus duerdenii ATCC BAA-1640 ----------IIKL----IVLYEDDKT----------YLKKKIKSAYKND----FT----------
224543312 Catenibacterium mitsuokai DSM 15897 ----------IIYD----LTVF-EDKK----------IMKRRLKKKYA----LP----------
116628213 Streptococcus thermophilus LMD-9    ----------IIHT----LTIF-EDRE----------MIKQRLSKFEN----IF----------
24379809  Streptococcus mutans UA159          ----------IVLT----LTLF-EDRE----------MIRKRLENYSD----LL----------
13622193  Streptococcus pyogenes SF370        ----------IVLT----LTLF-EDRE----------MIEERLKTYAH----LF----------
310286728 Bifidobacterium bifidum S17        ----------IVEL----QTVF-EDKE----------TLHQLRQLEG----IS----------
366983953 Oenococcus kitaharae DSM 17330     ----------ITEL----QTVF-EDKK----------VLRRQLDQLDG----LS----------
422884106 Streptococcus sanguinis SK49       ----------IIKV----ITVF-DDKK----------SIKNYLTKFFG--HLEILD----------
339625081 Fructobacillus fructosus KCTC 3544 ----------IIEI----QTVF-EDSK----------IASREL-SKLP----LD----------
306821691 Eubacterium yurii ATCC 43715       ----------IIFL----KTVYGDEKR----------FVKEEIVEKYG----DEID----------
336394882 Lactobacillus farciminis KCTC 3681 ----------LIEW----LTIF-EDKQ----------IINEKLHSSKY----SYT----------
323463801 Staphylococcus pseudintermedius ED99 ----------IIQW----TTIF-EDKK----------ILVQKLKECYP----ELT----------
389815359 Planococcus antarcticus DSM 14505  ----------LIYW----IAVF-EDRE----------ILHLKIQEKYP----SIT----------
25850199  Lactobacillus rhamnosus GG          ----------ITTW----STVF-EDHT----------IFETKLAEIEW----LD----------
169823755 Finegoldia magna ATCC 29328        ----------LIEK----ITIHTGNKK----------LLKKYIEETYPD----LS----------
Jnet      227501312                          ----------HHHH----HEE----HH----------HHHHHHHH----------
47458868  Mycoplasma mobile 163K             ----------I--SNI----LNKF-S----------TIQDRIILEGYPEPSNLK----------
284931710 Mycoplasma gallisepticum str. F    ----------VDSANL----KEFS-DSNK----------LFERILQKQKDGLFKLFEQ----------
71894592  Mycoplasma synoviae 53             ----------LDSIYLAISYSSDLKERNE----------WFKKLLKELYPKIKNNNLEIIENV
363542550 Mycoplasma ovipneumoniae SC01      ----------LDAICIILDREKS-RGQDE----------VIKKLTEKNITEVLKIDREKQLDFV
384393286 Mycoplasma canis PG 14             ----------LDNLYLFLIYQKESNNKDS----------SIDLFIAKNESLNIENLKLKLKEFL
238924075 Eubacterium rectale ATCC 33656     ----------IGYI----MTIN-TDKE----------AMMEAFQKSWID----LS----------
315149830 Enterococcus faecalis TX0012       ----------LAKV----LTLN-TERE----------GIENTLAFE----LP----------
116627542 Streptococcus thermophilus LMD-9   ----------LAYV----LTLN-TERE----------GIQEALEHE----FA----------
315659848 Staphylococcus lugdunensis M23590  ----------IAEI----LTIY-QDKD----------SIKSKLTELDI----LL----------
160915782 Eubacterium dolichum DSM 3991      ----------IAEI----LTKT-KDIE----------GRKKQISE----LS----------
325677756 Ruminococcus albus 8               ----------IGIV----LSQA-QTPK----------RRREKLKALNIG----LD----------
225377804 Roseburia inulinivorans DSM 16841  ----------IGEI----LTAY-KNDD----------SRSSRLKELG----LS----------
336393381 Lactobacillus coryniformis KCTC 3535 ----------IGTA----LTLY-SSDK----------RRRRYFAEELN----LP----------
310780384 Ilyobacter polytropus DSM 2926     ----------IIEI----LTFN-KSDK----------TIESNLKKLE----LS----------
301311869 Bacteroides sp. 20 3               ----------IWHI----LYSI-EDKQ----------EIEKALHSFANKNN----LS----------
385811609 Ignavibacterium album JCM 16511    ----------LWHS----DYSNDYADKE----------KTEKSILSSLGWKNRNGKWEKSKN
60683389  Bacteroides fragilis NCTC 9343     QPYYK----------LWHL----LYSF-EGDNTPTGNGRLIQKMTELYGF----------
319957206 Nitratifractor salsuginis DSM 16511 ----------LAEI----LQRS-KTPQ----------EALDRLRALMAGKG----ID----------
187250660 Elusimicrobium minutum Pei191      ----------FLYD----WNSC-PDEK----------LLTEKLSNEYH----LT----------
325972003 Sphaerochaeta globus str. Buddy    ----------FFSA----WTNT-PDDK----------RLSKYLMKELL----LT----------
```

Figure S2 (Continued)

Figure S2 (Continued)

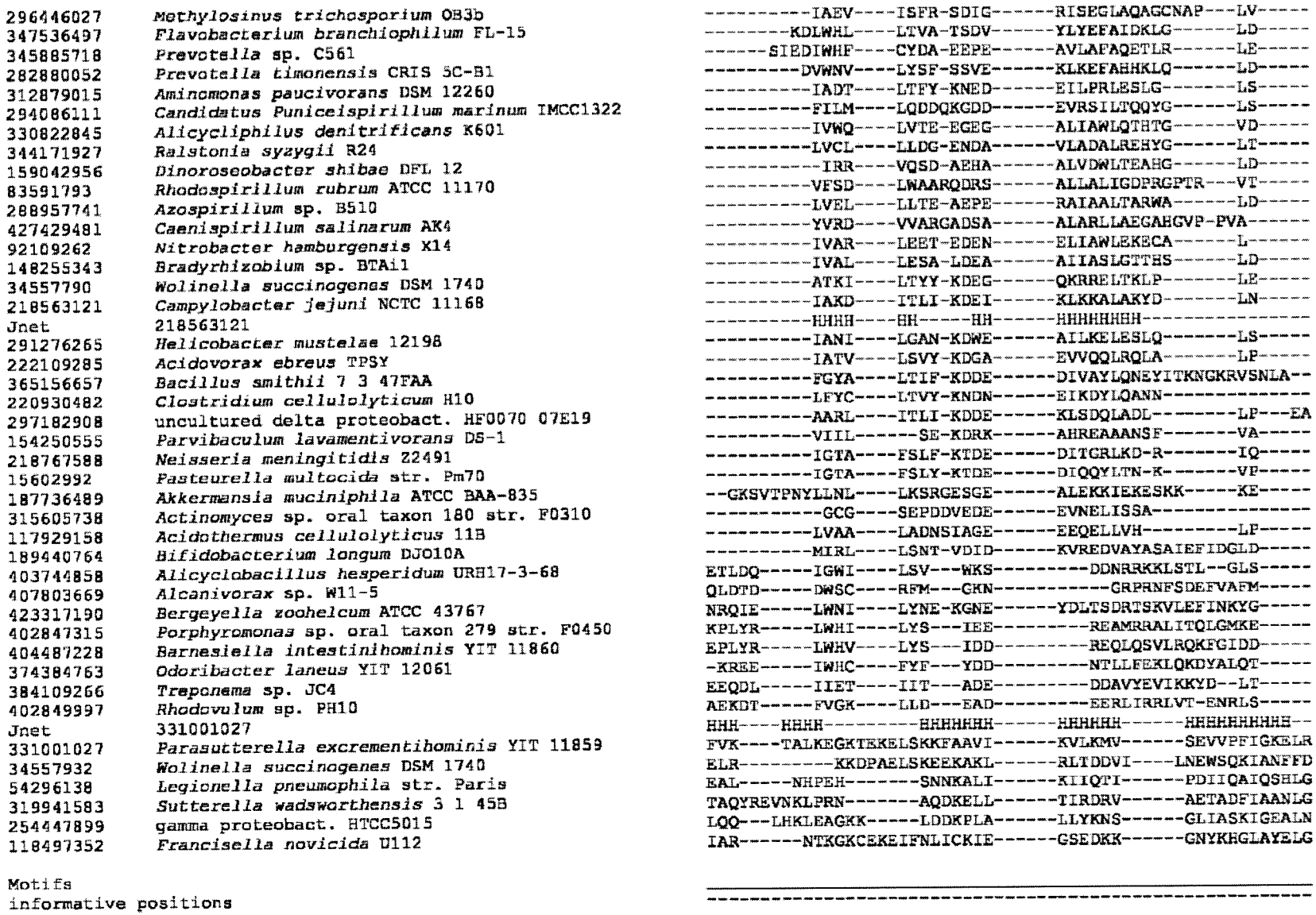


Figure S2 (Continued)

Figure S2 (Continued)

Figure S2 (Continued)

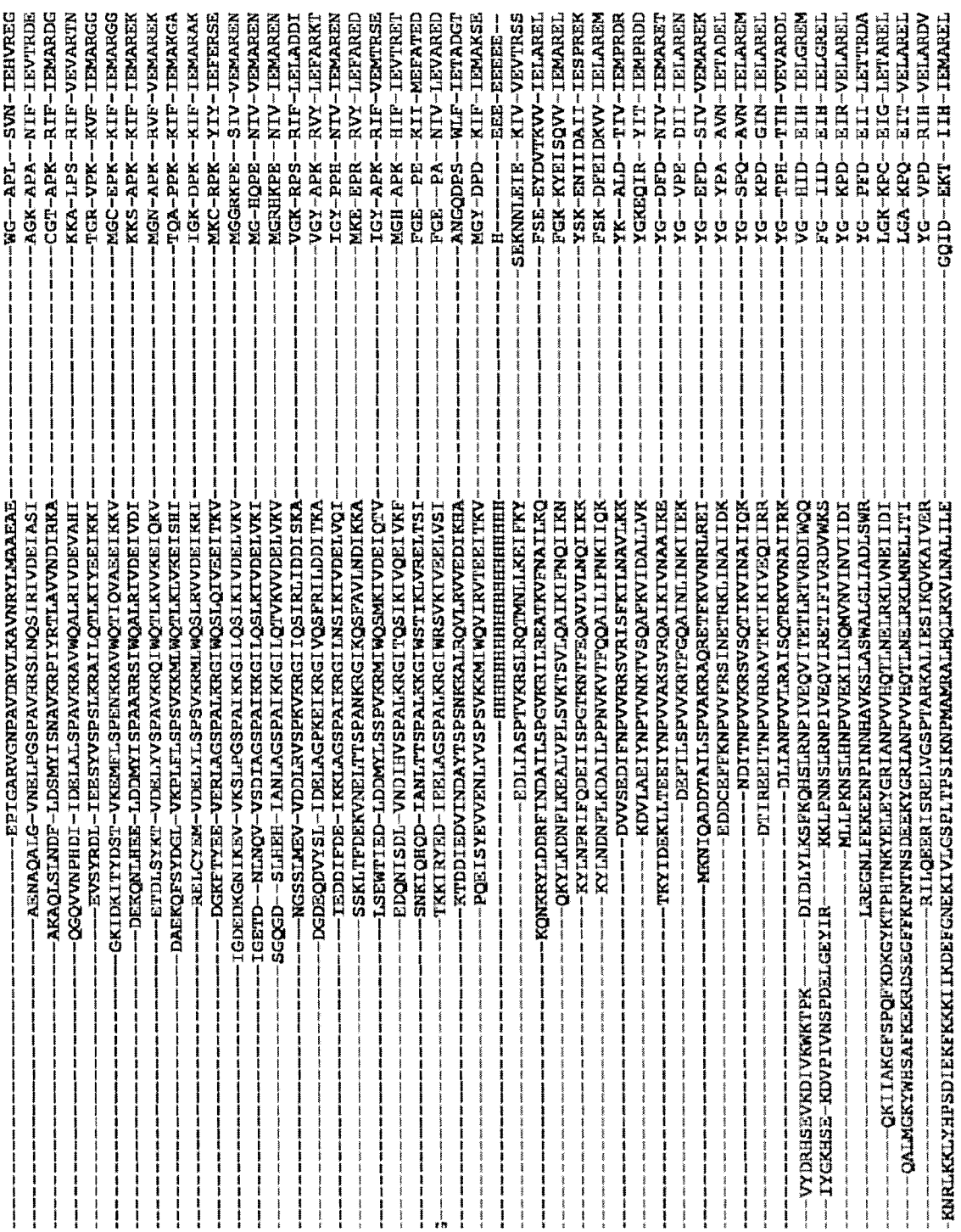


Figure S2 (Continued)

Figure S2 (Continued)

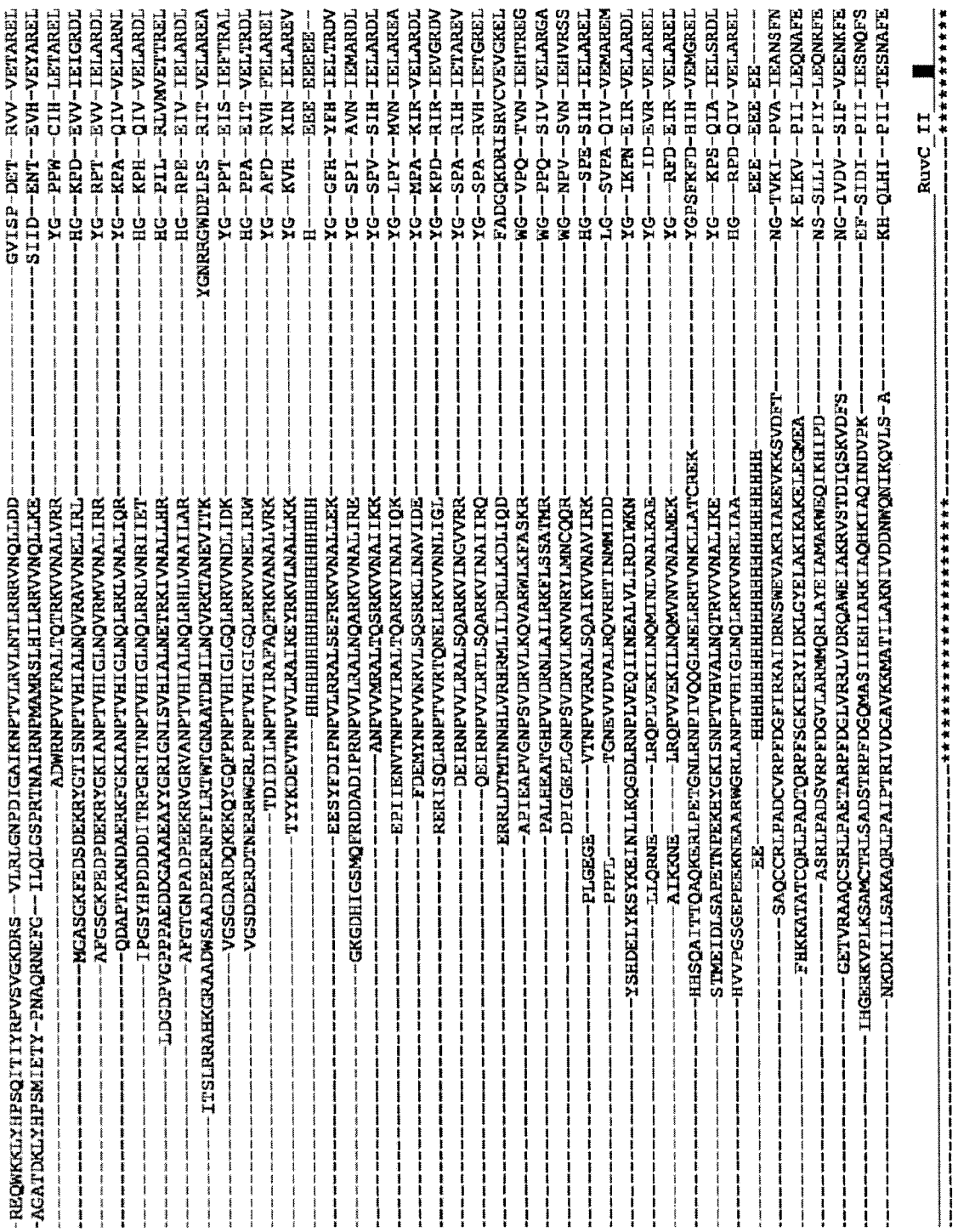


Figure S2 (Continued)

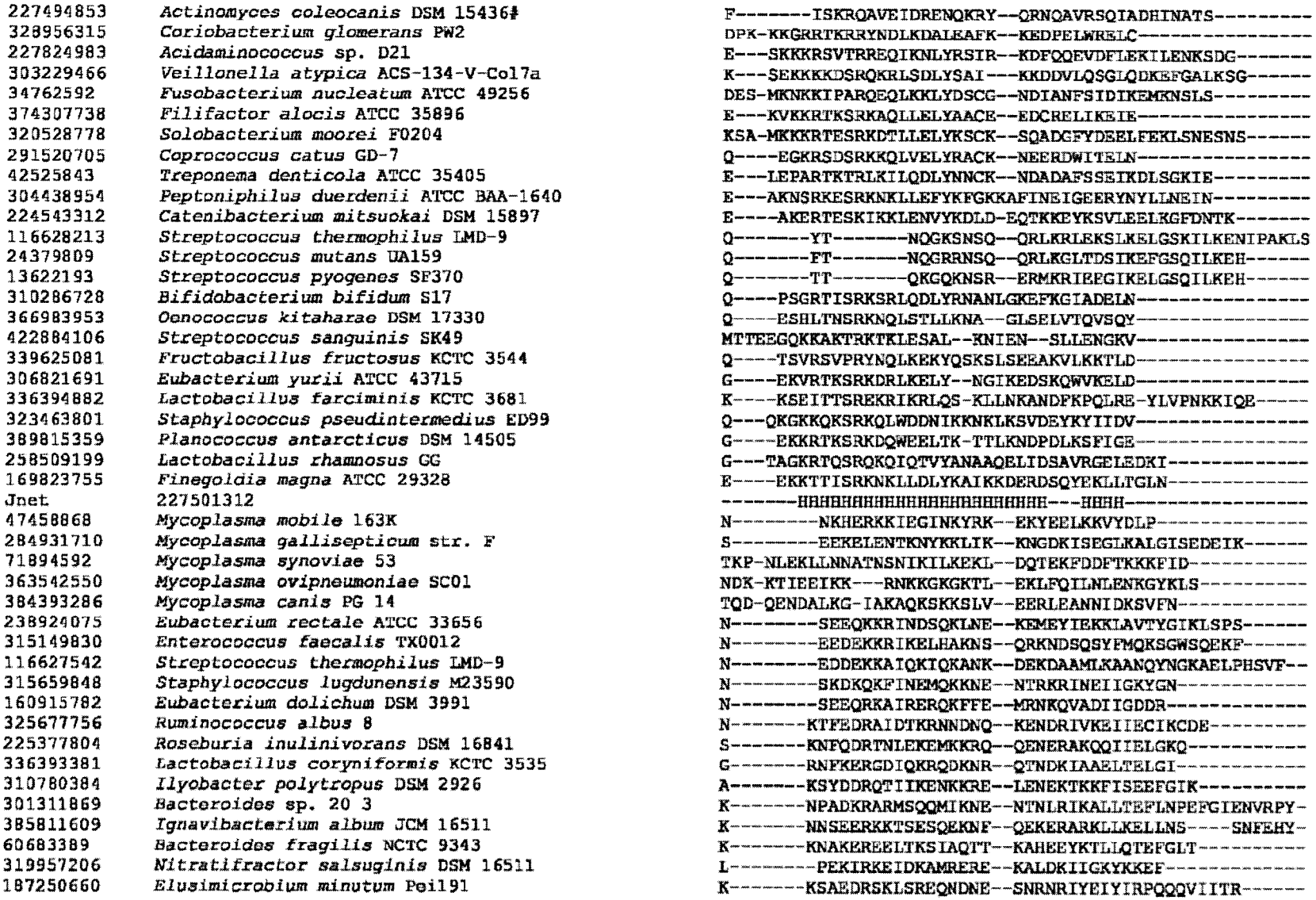


Figure S2 (Continued)

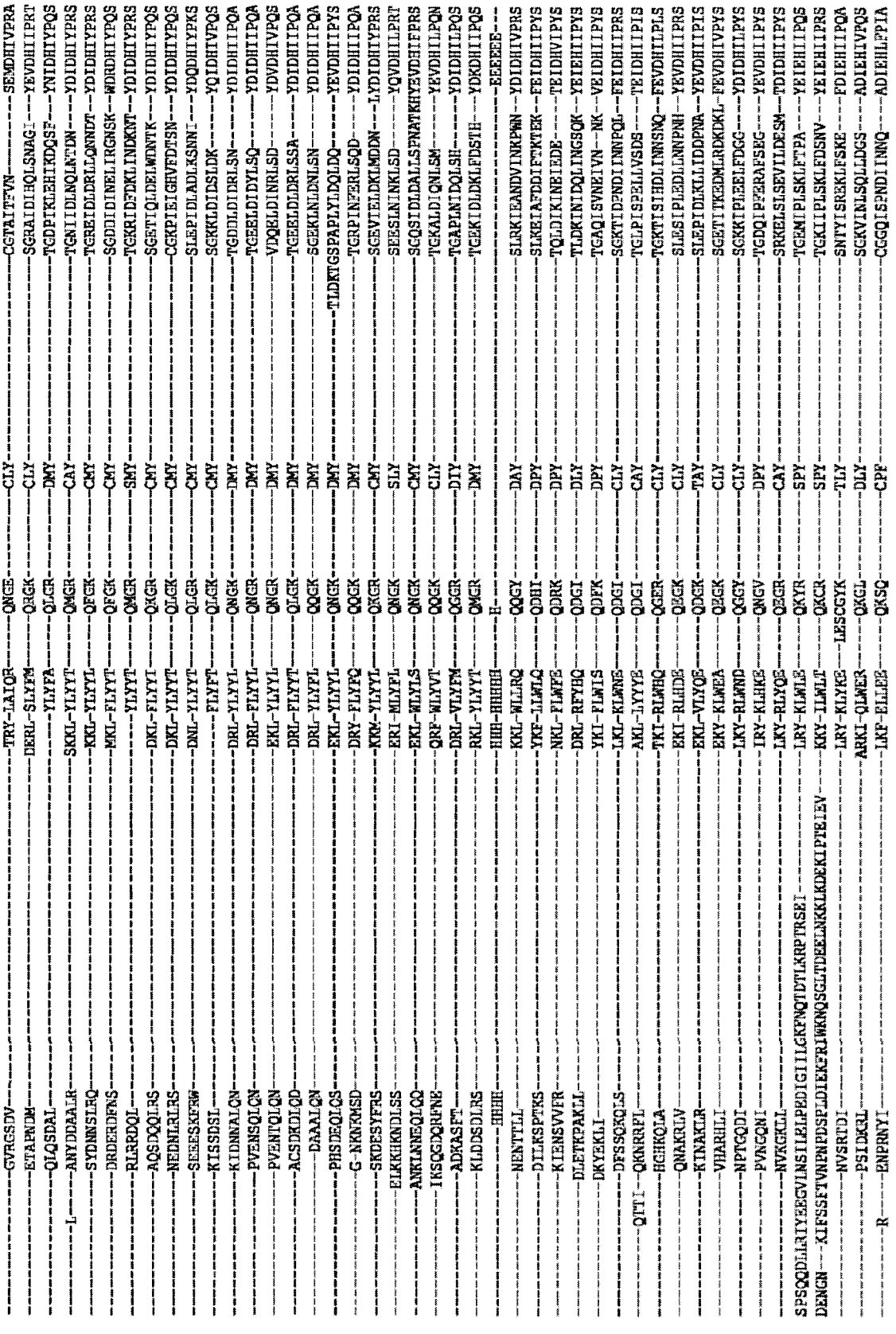


Figure S2 (Continued)

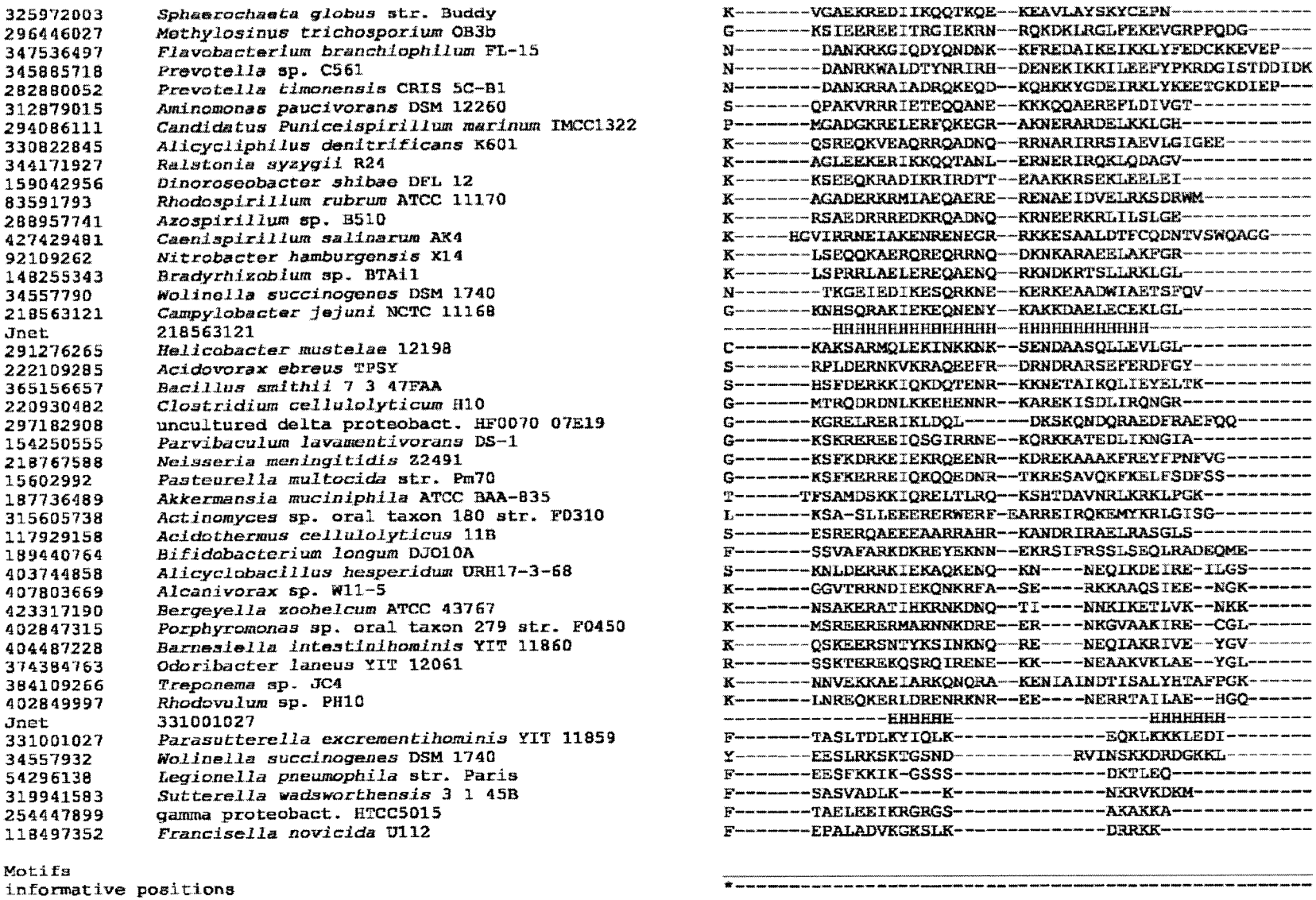


Figure S2 (Continued)

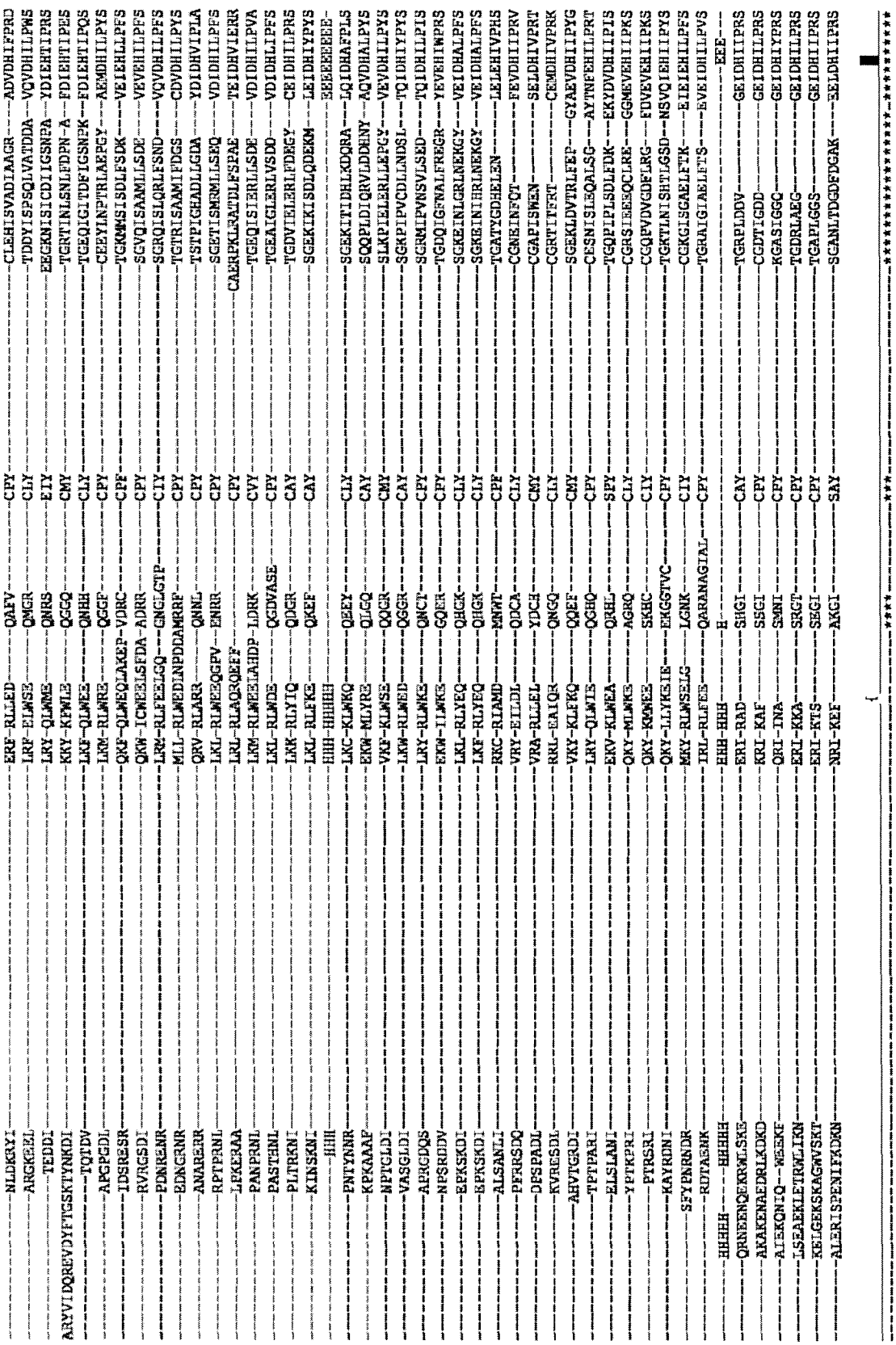


Figure S2 (Continued)

Figure S2 (Continued)

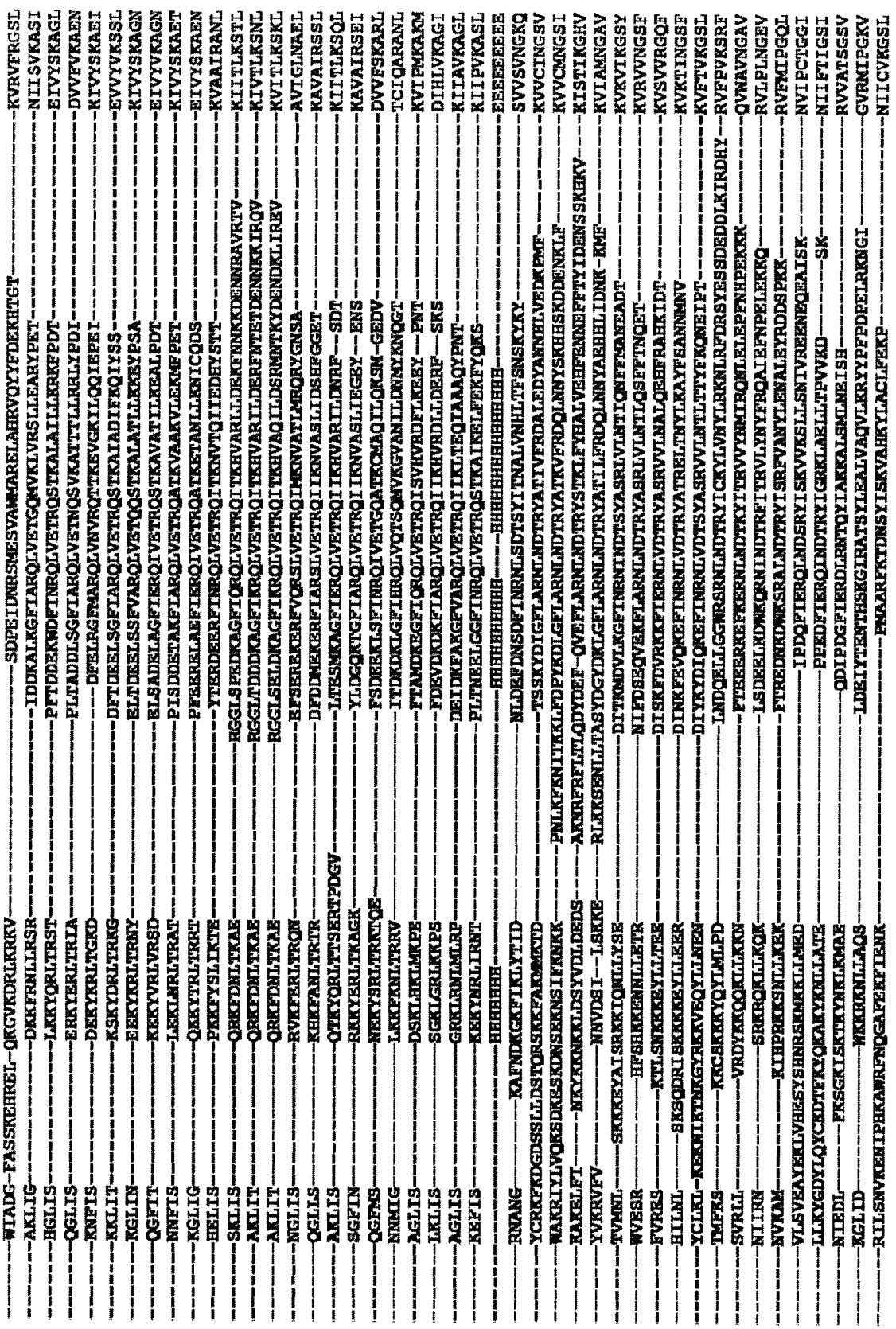


Figure S2 (Continued)

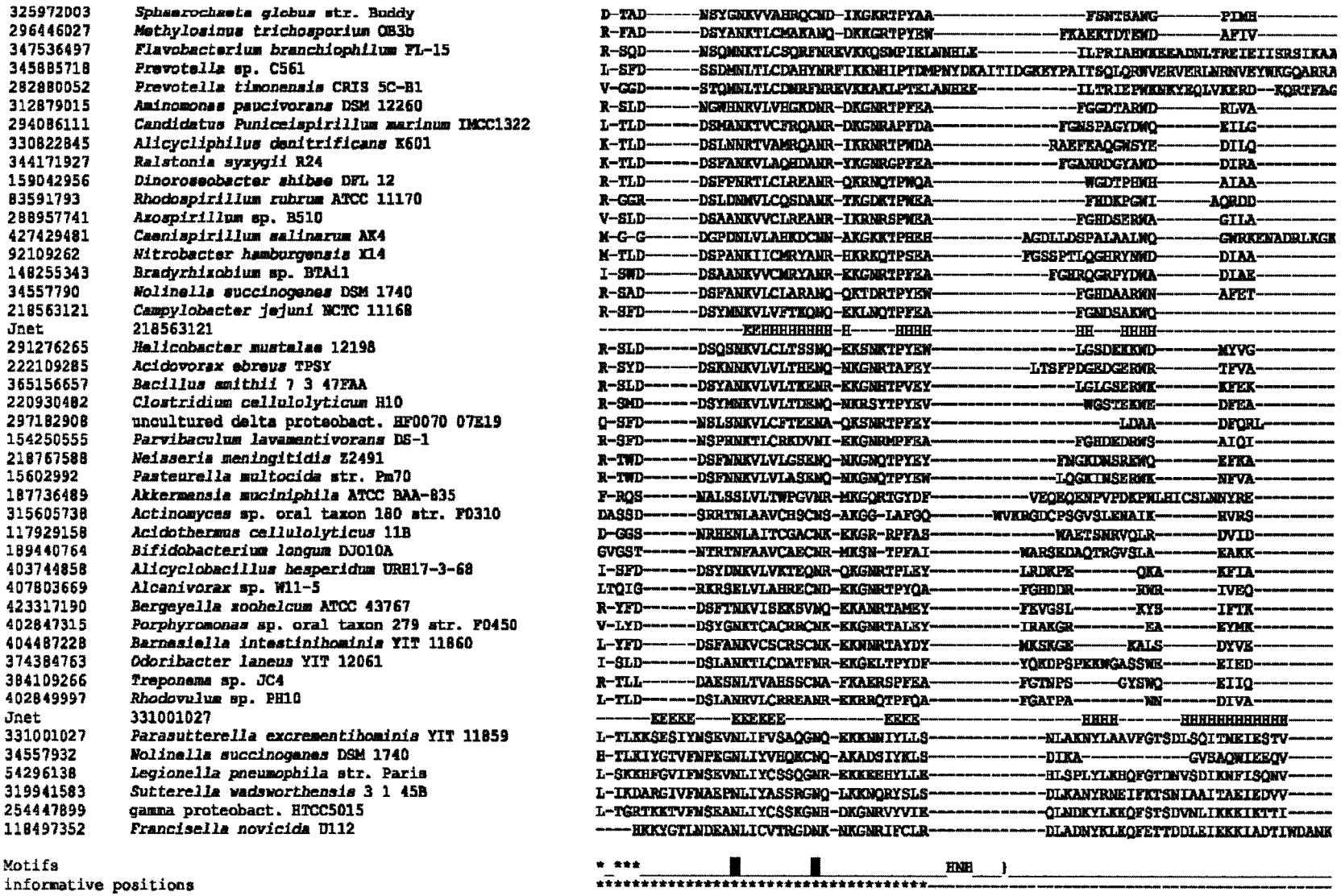


Figure S2 (Continued)

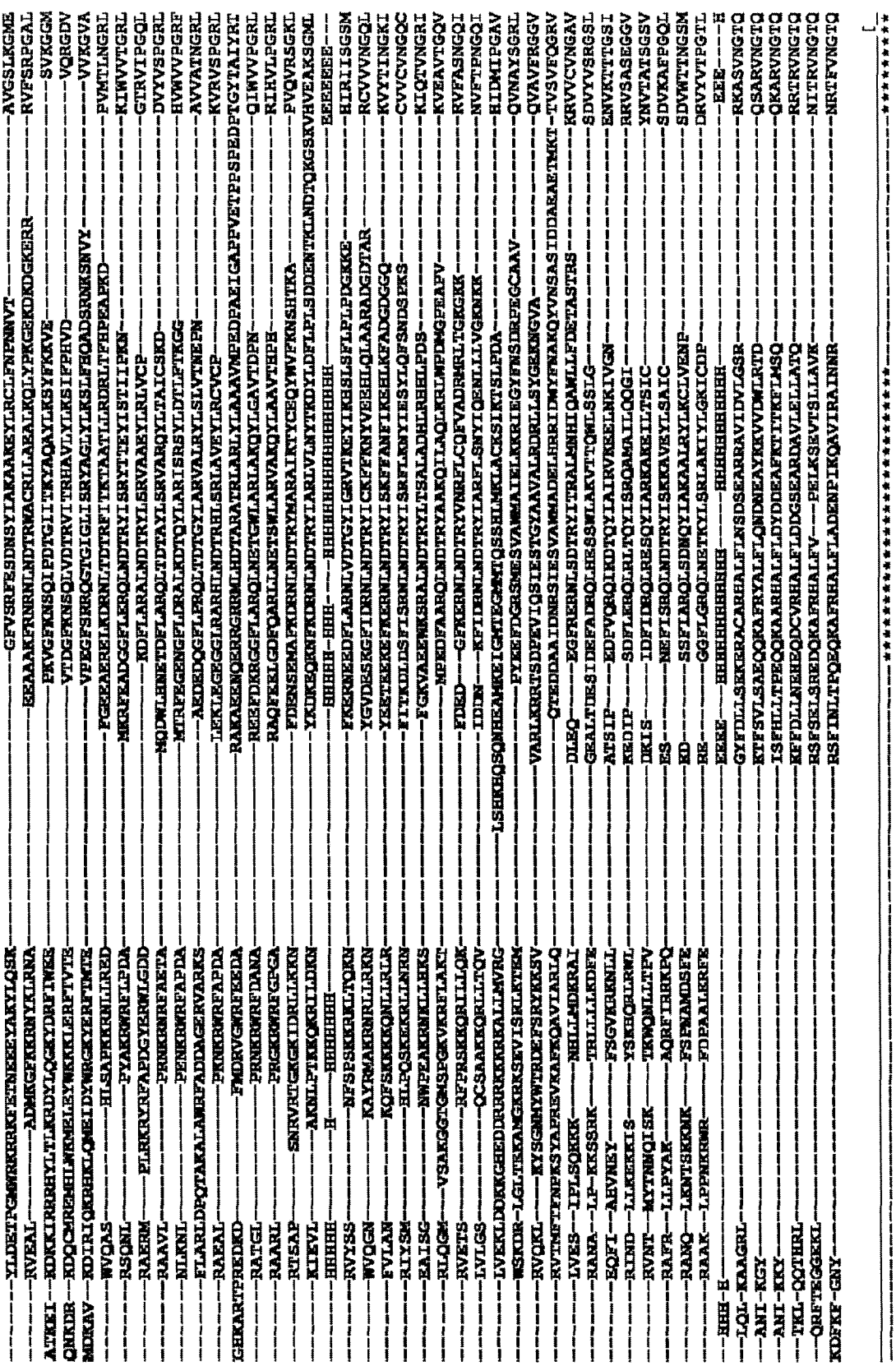


Figure S2 (Continued)

```
227494853  Actinomyces coleocanis DSM 15436#  TSAARKAS----------GFS----------
328956315  Coriobacterium glomerans PW2       SHDLRTAA----------ELV----------
227824983  Acidaminococcus sp. D21            SSDFRHEF----------GLV----------
303229466  Veillonella atypica ACS-134-V-Col7a VSDFRHNN----------NFI----------
34762592   Fusobacterium nucleatum ATCC 49256 ASSFREMF----------DFI----------
374307738  Filifactor alocis ATCC 35896       VSDFRKKPL----------NYL----------
320528778  Solobacterium moorei F0204         VSEFRNRKD----------KELP---------
291520705  Coprococcus catus GD-7             VSNFRQTY----------ELL----------
42525843   Treponema denticola ATCC 35405     VSMFRNKF----------DIV----------
304438954  Peptoniphilus duerdenii ATCC BAA-1640 ASRFRQEF----------DII----------
224543312  Catenibacterium mitsuokai DSM 15897 SHEFRVKN----------HIY----------
116628213  Streptococcus thermophilus LMD-9   VSQFRKDF----------ELY----------
24379809   Streptococcus mutans UA159         VSNFRKEF----------ELY----------
13622193   Streptococcus pyogenes SF370       VSDFRKDF----------QFY----------
310286728  Bifidobacterium bifidum S17        TKEMHRYL----------GFS----------
366983953  Oenococcus kitaharae DSM 17330     TADMRRYV----------DIP----------
422884106  Streptococcus sanguinis SK49       ITNFRNTF----------HIA----------
339625081  Fructobacillus fructosus KCTC 3544 TADMRLLV----------GIK----------
306821691  Eubacterium yurii ATCC 43715       VSEFRHKF----------ELF----------
336394882  Lactobacillus farciminis KCTC 3681 STAFRKALSGQDDTYHFKHP----------ELV----------
323463801  Staphylococcus pseudintermedius ED99 VSEFRKKF----------DIP----------
389815359  Planococcus antarcticus DSM 14505  VSKFRRFS----------EIP----------
258509199  Lactobacillus rhamnosus GG         SHQLREEL----------DFP----------
169823755  Finegoldia magma ATCC 29328        ASDLRKDM----------NTL----------
Jnet       227501312                          EHHHHHH----------
47458868   Mycoplasma mobile 163K             TSNLRNQIAFVGIKNNKETEREW-KRPEGFK----------
284931710  Mycoplasma gallisepticum str. F    TSFLRKNFD----------DSSYA----------
71894592   Mycoplasma synoviae 53             TSFLRKSMW----------RKN----------LNNNK----------
363542550  Mycoplasma ovipneumoniae SC01      TKYFRAKPVQKNNGPNEN----------SYDN----------
384393286  Mycoplasma canis PG 14             TSFIRKNM----------KLD----------
238924075  Eubacterium rectale ATCC 33656     THQMRCNL----------GAD----------
315149830  Enterococcus faecalis TX0012       THTLRKKW----------GIE----------
116627542  Streptococcus thermophilus LMD-9   TSQLRRHW----------KFK----------
315659848  Staphylococcus lugdunensis M23590  TDYLRKVW----------NLK----------
160915782  Eubacterium dolichum DSM 3991      TNAFRRKI----------LN----------
325677756  Ruminococcus albus 8               TSMFRRWW----------GLM----------
225377804  Roseburia inulinivorans DSM 16841  TSYLRKRW----------GFL----------
336393381  Lactobacillus coryniformis KCTC 3535 TSKIRSRW----------RLN----------
310780384  Ilyobacter polytropus DSM 2926     TAQLRARW----------GIN----------
301311869  Bacteroides sp. 20 3               TDRLKKDW----------GLN----------
385811609  Ignavibacterium album JCM 16511    TSELKITW----------QLI----------
60683389   Bacteroides fragilis NCTC 9343     TDKLREDW----------GIK----------
319957206  Nitratifractor salsuginis DSM 16511 TSKTRSLL----------GLQ----------
187250660  Elusimicrobium minutum Pei191      TAQLRMAW----------NLQ----------
325972003  Sphaerochaeta globus str. Buddy    TSILRKAW----------GLQ----------
296446027  Methylosinus trichosporium OB3b    TDRLRRAW----------
```

Figure S2 (Continued)

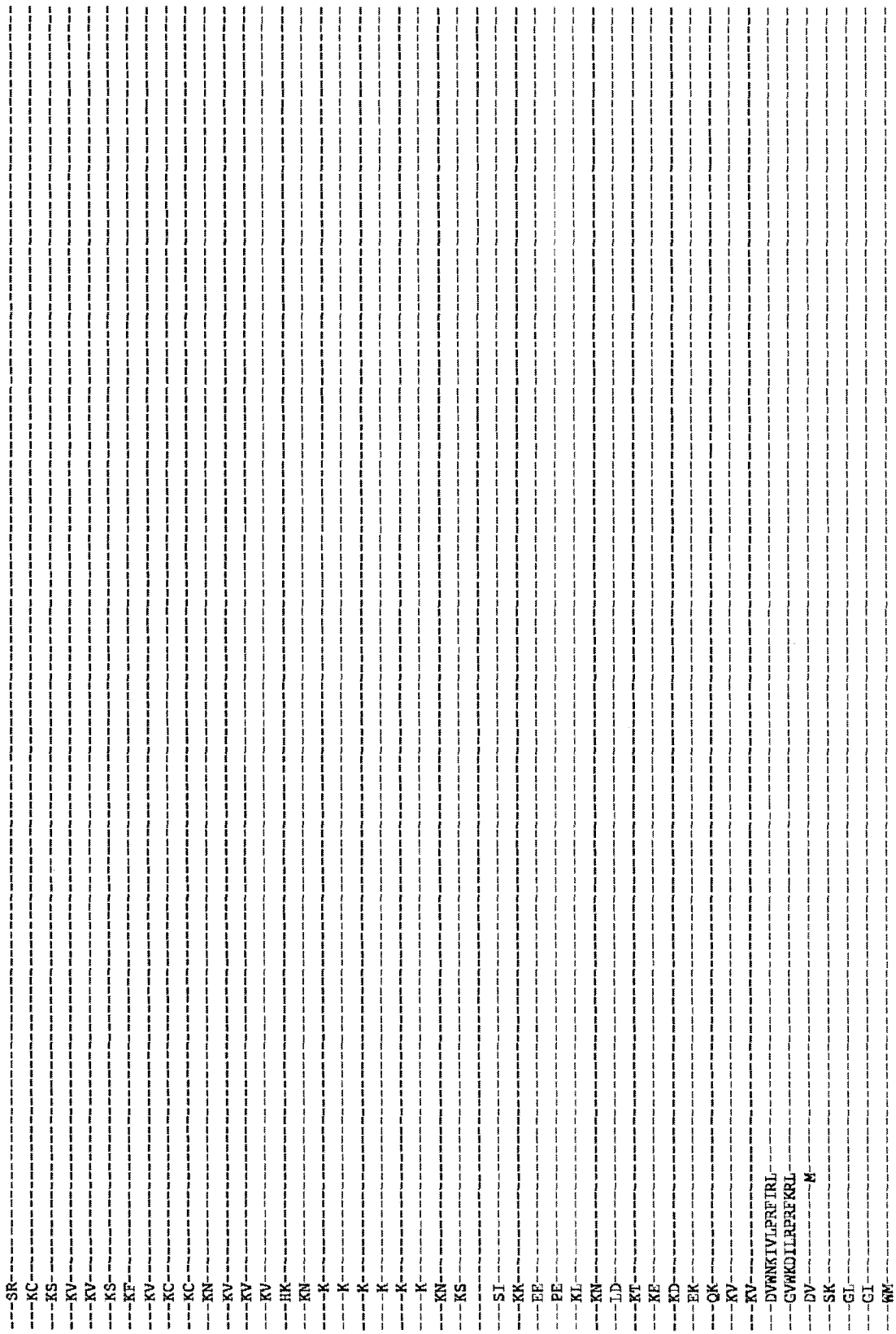


Figure S2 (Continued)

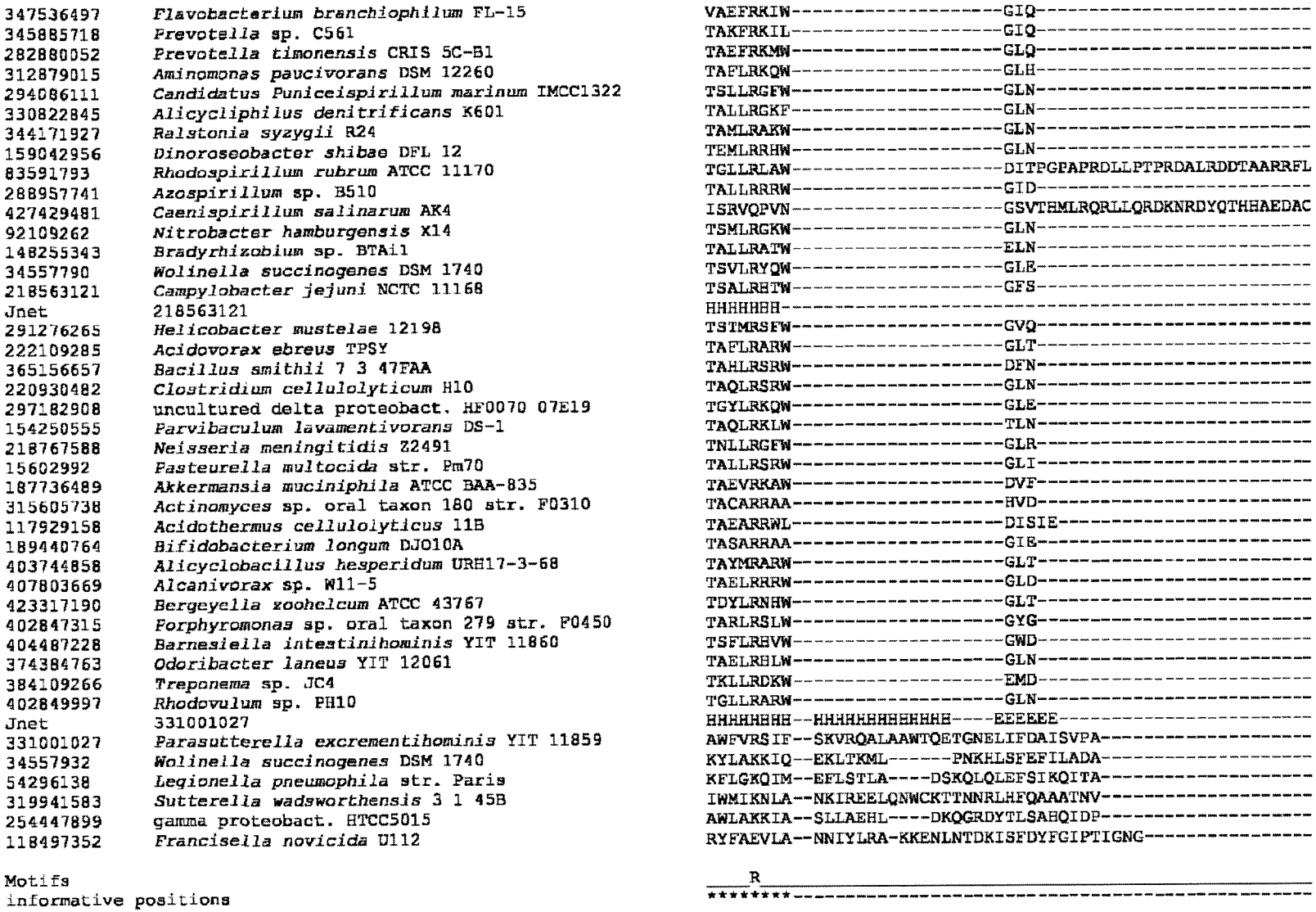


Figure S2 (Continued)

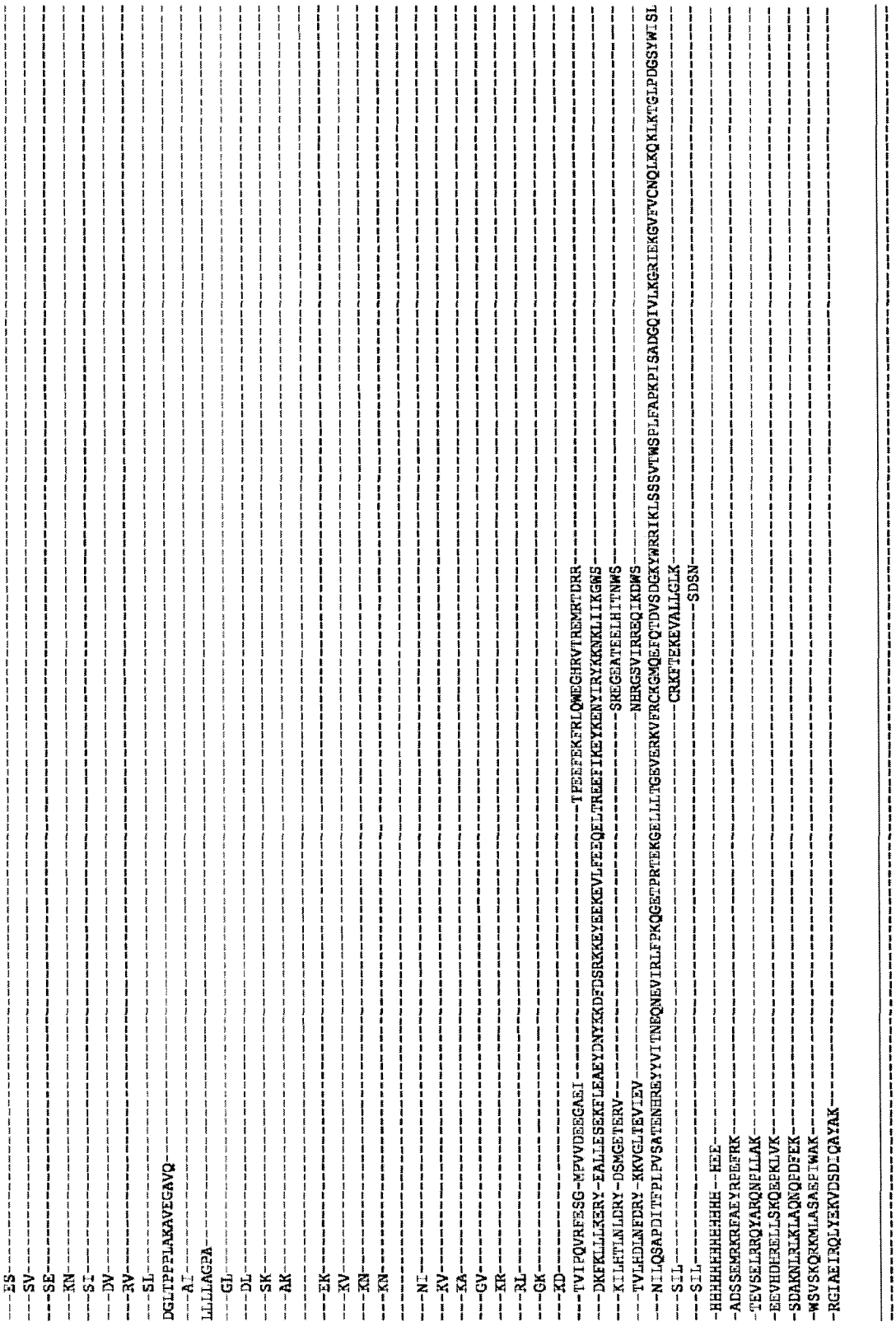


Figure S2 (Continued)

227494853	*Actinomyces coleocanis* DSM 15436
328956315	*Coriobacterium glomerans* PW2
227824983	*Acidaminococcus* sp. D21
303229466	*Veillonella atypica* ACS-134-V-Col7a
34762592	*Fusobacterium nucleatum* ATCC 49256
374307738	*Filifactor alocis* ATCC 35896
320528778	*Solobacterium moorei* F0204
291520705	*Coprococcus catus* GD-7
42525843	*Treponema denticola* ATCC 35405
304438954	*Peptoniphilus duerdenii* ATCC BAA-1640
224543312	*Catenibacterium mitsuokai* DSM 15897
116628213	*Streptococcus thermophilus* LMD-9
24379809	*Streptococcus mutans* UA159
13622193	*Streptococcus pyogenes* SF370
310286728	*Bifidobacterium bifidum* S17
366983953	*Oenococcus kitaharae* DSM 17330
422884106	*Streptococcus sanguinis* SK49
339625081	*Fructobacillus fructosus* KCTC 3544
306821691	*Eubacterium yurii* ATCC 43715
336394882	*Lactobacillus farciminis* KCTC 3681
323463801	*Staphylococcus pseudintermedius* ED99
389815359	*Planococcus antarcticus* DSM 14505
258509199	*Lactobacillus rhamnosus* GG
169823755	*Finegoldia magna* ATCC 29328
Jnet	227501312
47458868	*Mycoplasma mobile* 163K
28493171O	*Mycoplasma gallisepticum* str. F
71894592	*Mycoplasma synoviae* 53
363542550	*Mycoplasma ovipneumoniae* SC01
384393286	*Mycoplasma canis* PG 14
238926075	*Eubacterium rectale* ATCC 33656
315149830	*Enterococcus faecalis* TX0012
116627542	*Streptococcus thermophilus* LMD-9
315659848	*Staphylococcus lugdunensis* M23590
160915782	*Eubacterium dolichum* DSM 3991
325677756	*Ruminococcus albus* 8
225377804	*Roseburia inulinivorans* DSM 16841
336393381	*Lactobacillus coryniformis* KCTC 3535
310780384	*Ilyobacter polytropus* DSM 2926
301311869	*Bacteroides* sp. 20 3
385811609	*Ignavibacterium album* JCM 16511
60683389	*Bacteroides fragilis* NCTC 9343
319957206	*Nitratifractor salsuginis* DSM 16511
187250660	*Elusimicrobium minutum* Pei191

Figure S2 (Continued)

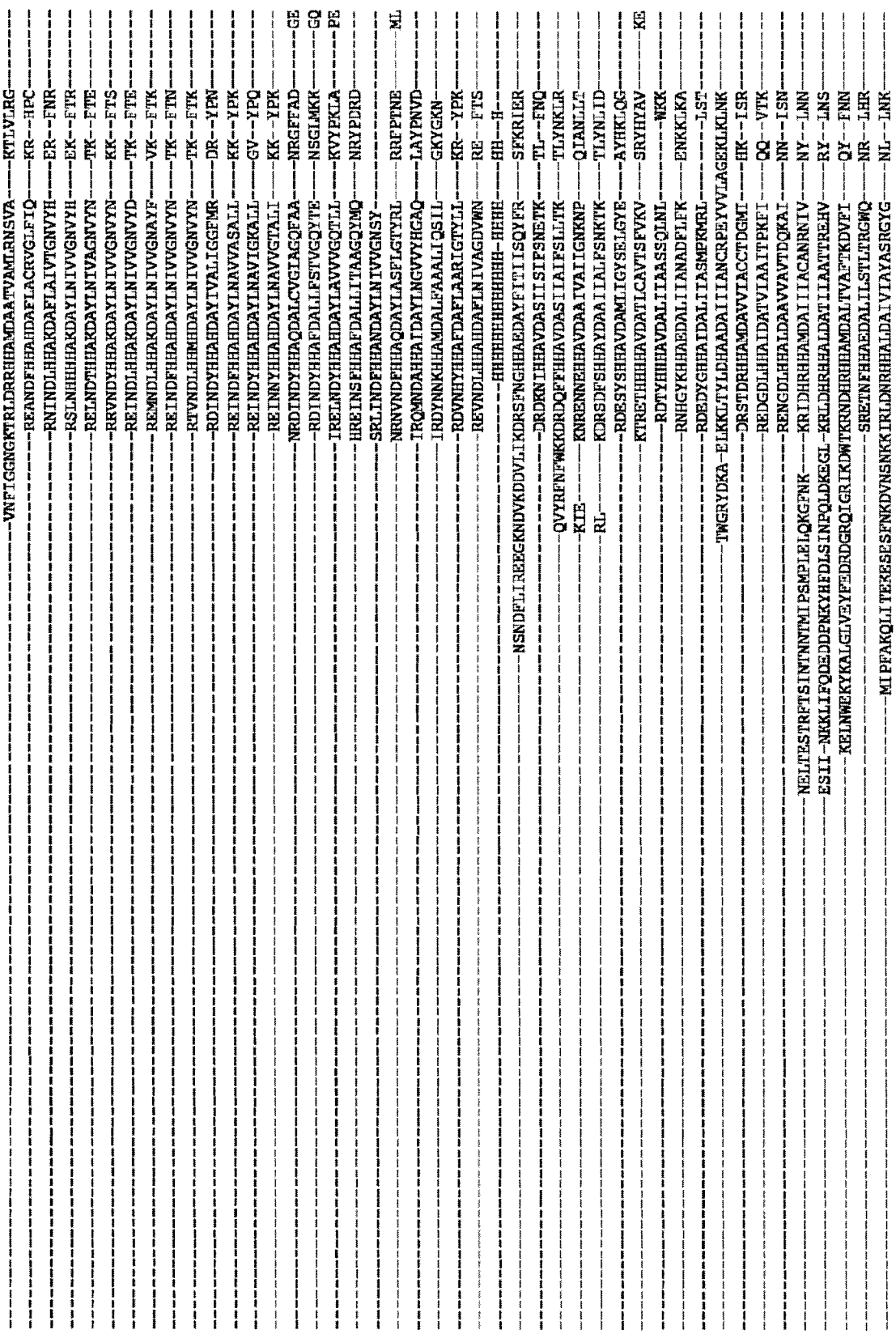


Figure S2 (Continued)

```
325972003  Sphaerochaeta globus str. Buddy
296446027  Methylosinus trichosporium OB3b
347536437  Flavobacterium branchiophilum FL-15
345885718  Prevotella sp. C561
282880052  Prevotella timonensis CRIS 5C-B1
312879015  Aminomonas paucivorans DSM 12260
294086111  Candidatus Puniceispirillum marinum IMCC1322
330822845  Alicycliphilus denitrificans K601
344171927  Ralstonia syzygii R24
159042956  Dinoroseobacter shibae DFL 12
83591793   Rhodospirillum rubrum ATCC 11170
288957741  Azospirillum sp. B510
427429481  Caenispirillum salinarum AK4
92109262   Nitrobacter hamburgensis X14
148255343  Bradyrhizobium sp. BTAi1
34557790   Wolinella succinogenes DSM 1740
218563121  Campylobacter jejuni NCTC 11168
Jnet
291276265  Helicobacter mustelae 12198
222109285  Acidovorax ebreus TPSY
365156657  Bacillus smithii 7 3 47FAA
220930482  Clostridium cellulolyticum H10
297182908  uncultured delta proteobact. HF0070 07E19
154250535  Parvibaculum lavamentivorans DS-1
218767588  Neisseria meningitidis Z2491
15602992   Pasteurella multocida str. Pm70
187736489  Akkermansia muciniphila ATCC BAA-835
315605738  Actinomyces sp. oral taxon 180 str. F0310
117929158  Acidothermus cellulolyticus 11B
189440764  Bifidobacterium longum DJO10A
403744858  Alicyclobacillus hesperidum URH17-3-68
407803669  Alcanivorax sp. W11-5
423317190  Bergeyella zoohelcum ATCC 43767
402847315  Porphyromonas sp. oral taxon 279 str. F0450
404487228  Barnesiella intestinihominis YIT 11860
374384763  Odoribacter laneus YIT 12061
384109266  Treponema sp. JC4
402849997  Rhodovulum sp. PH10
Jnet
331001027  Parasutterella excrementihominis YIT 11859
34557932   Wolinella succinogenes DSM 1740
54296138   Legionella pneumophila str. Paris
319941583  Sutterella wadsworthensis 3 1 45B
254447899  gamma proteobact. HTCC5015
118497352  Francisella novicida U112

Motifs
Informative positions
```

Figure S2 (Continued)

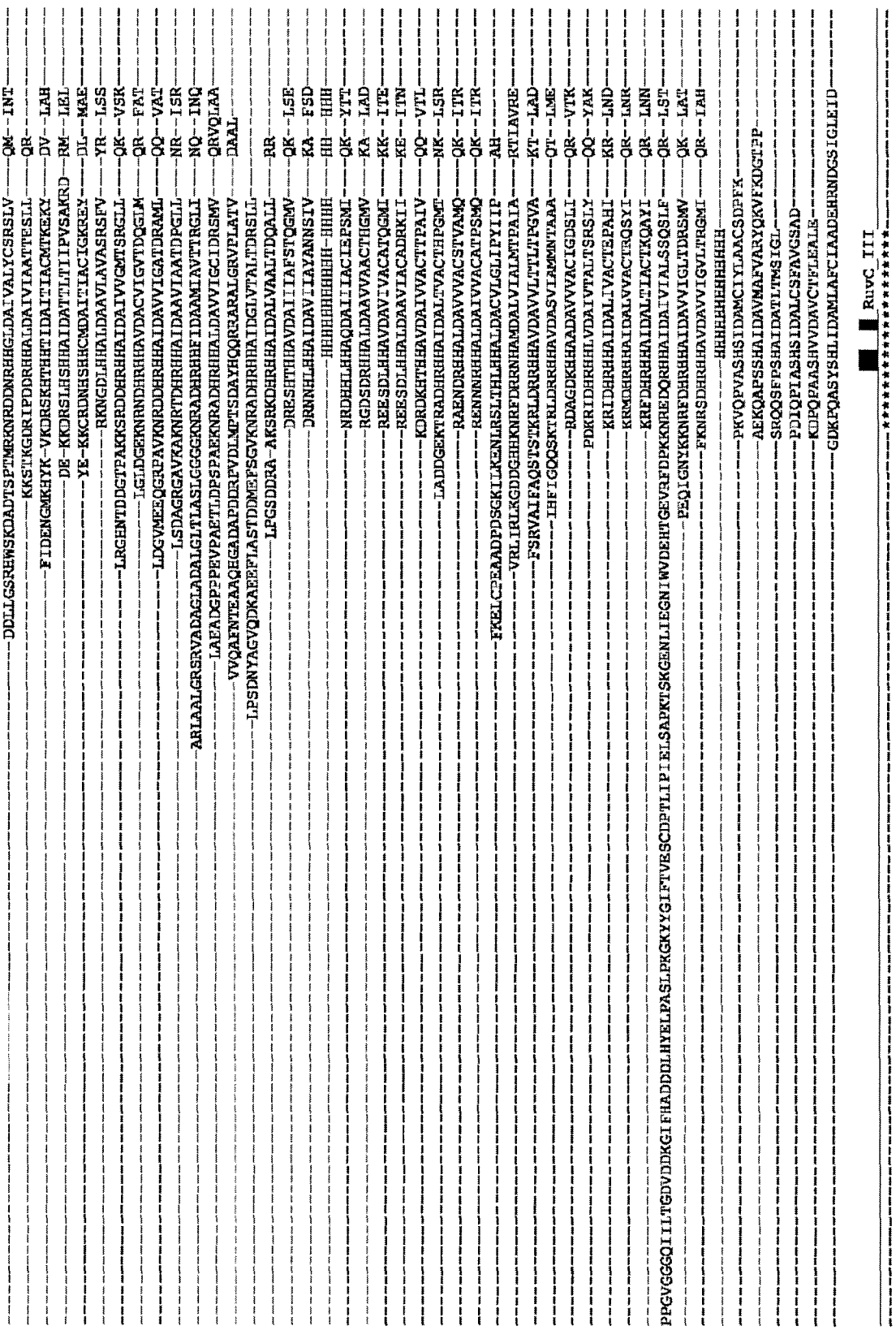


Figure S2 (Continued)

Figure S2 (Continued)

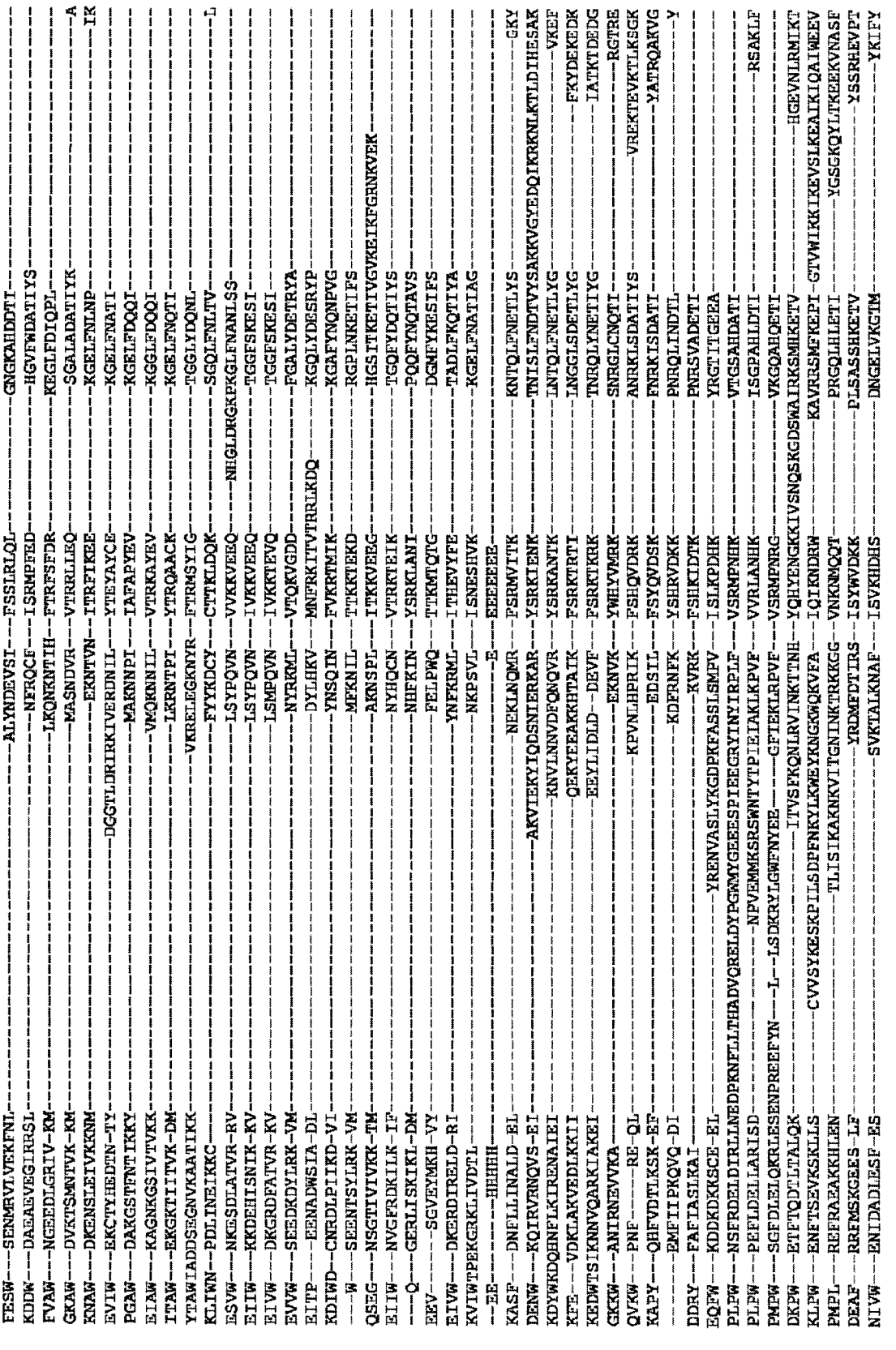


Figure S2 (Continued)

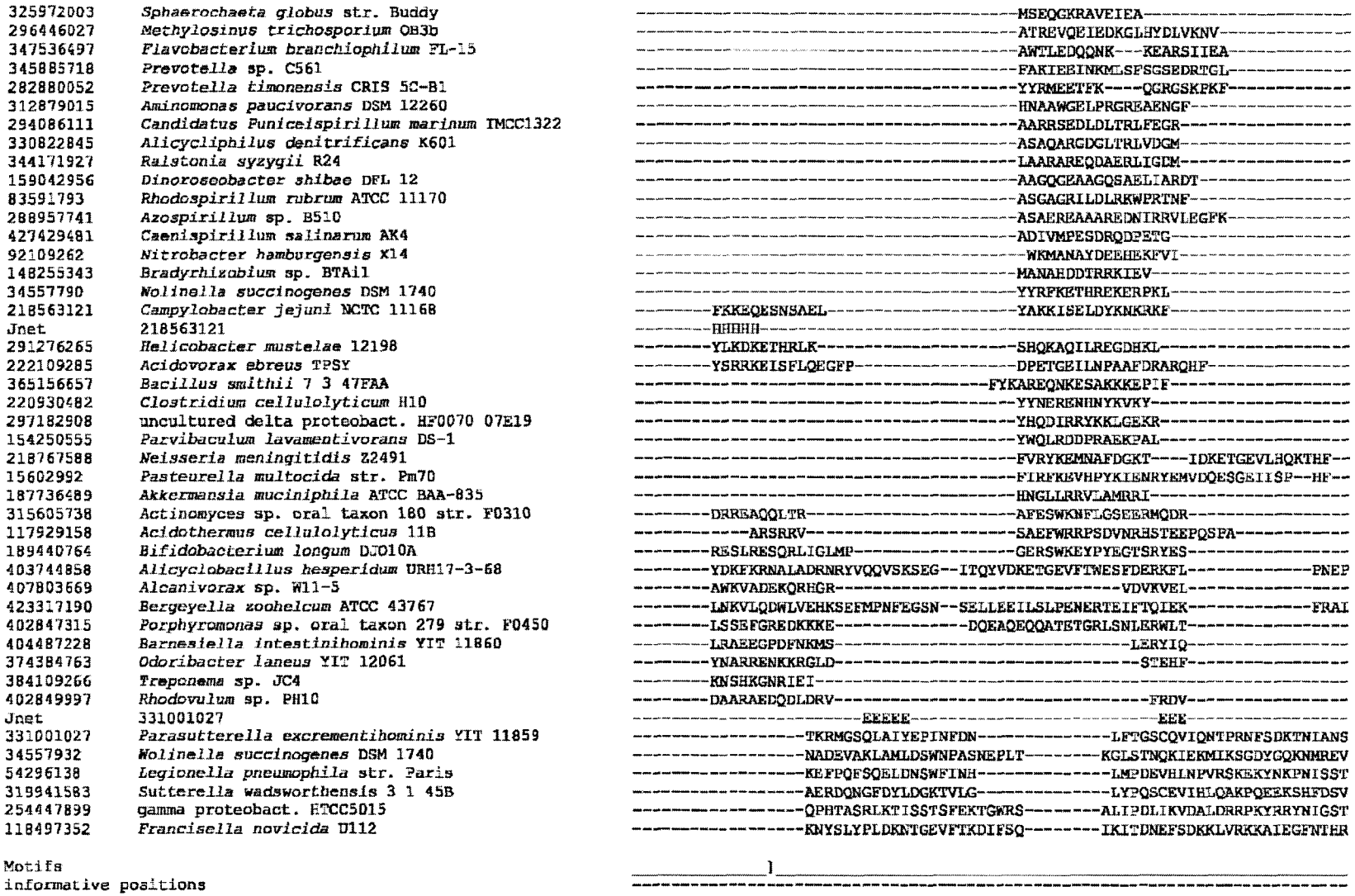


Figure S2 (Continued)

Figure S2 (Continued)

```
227494853  Actinomyces coleocanis DSM 15436#  -TKLQMHKVGDAWSLTEIDRAST------PALWCALTR-QPDFTWKDGL----------
328956315  Coriobacterium glomerans PW2       ----------PRAKKTAALPLKQGLN------PSRYGSFSHEQFAYFFIYKARNPRK----------
227824983  Acidaminococcus sp. D21            ----------KASTGLVTPRKAGLD------VVKYGGYDKSTAAYYLLVRFTLED----------
303229466  Veillonella atypica ACS-134-V-Col7a SVAAKAKDGAYIGMKTKYSVFAD------VTKYGGWTKIKNAYSIIVQTTGKK----------
34762592   Fusobacterium nucleatum ATCC 49256 KGETSNEIISIKPKLNYDGKDWKL------NEKYGTYTSLKAAYFIYVEHEKKNK----------
374307738  Filifactor alocis ATCC 35896       ----------QMKNGNSTVSLKKGLD------VKKYGGYFSANTSYFSLIEF----------
320528778  Solobacterium moorei F0204         ----------VPKGKGQFPIKQGKD------IEKYGGYNKLSSAFLFAVEYKGKKAR----------
291520705  Coprococcus catus GD-7              -----MKKGKGQVPIKGNDERLAD------IEKYGGYNKQAGTYFMLVKSLDKKGK----------
42525843   Treponema denticola ATCC 35405      -----MKKGLGQMPLKKEGPFSN------ISKYGGYNKVSAAYYTLIEYEEK----------
304438954  Peptoniphilus duerdenii ATCC BAA-1640 -----MRKGKGQLPQKENTNKSN------IEKYGGYNKASSAYFALIESDGKAG----------
224543312  Catenibacterium mitsuokai DSM 15897 SNDAHADKGVTKAVVPVNKNRSD------VEKYGGFSGLQYTIVAIEGQKKKGK----------
116628213  Streptococcus thermophilus LMD-9    ----------KPKPNSNENLVGAKEYLD------PKKYGGYAGISNSFTVIVKGTIEKG----------
24379809   Streptococcus mutans UA159         ----------LPKGNSDKLIPRKTKKFTWD------TKKYGGFDSPTVAYSILVIADIEKG----------
13622193   Streptococcus pyogenes SF370       ----------LPKRNSDKLIARKKDWD------PKKYGGFDSPTVAYSVLVVAKVEKG----------
310286728  Bifidobacterium bifidum S17        -----ATDPKGIKGIPFDGAKQD------TSLYGGFSSAKPAYAVLIESKG----------
366983953  Oenococcus kitaharae DSM 17330     ---SVLHDAKSKASINFDKEKP------VDLYGGFSSAKPAYAALIKFKN----------
422884106  Streptococcus sanguinis SK49       ---KTNKQIAANNRYPLKTKALCLD---TSIYGGYGPMNSALSIIIIAERFNE----------
339625081  Fructobacillus fructosus KCTC 3544 -----PKS--GKKLIPLNSKRSD------TALYGGYSNVYSAYMTLVRANG----------
306821691  Eubacterium yurii ATCC 43715       -ADKTPKKPNLQAYRPIKTS-DERLCN-ILRYGGRTSISISGYCLVEYVK----------
336394882  Lactobacillus farciminis KCTC 3681 -----PKNPKYKKLIAQKKD-MD------PNIYGGFSGDNKSSTITVKIDN----------
323463801  Staphylococcus pseudintermedius ED99 -----PKTAELKYESNKSN------EVVYKGLTPYQTVVVAIKSVNK----------
389815359  Planococcus antarcticus DSM 14505  -----PKVKQAKYVSPKTS------KFVHDEVKNHSICLVEFTMK----------
258509199  Lactobacillus rhamnosus GG         --AKDSKERGGSKQLIPKKQGYP------TQVYGGYTQESGSYNAIVRVAEA----------
169823755  Finegoldia magna ATCC 29328        ---KKDYKKGKIYLPLKKDDRLQD------VSKYGGYKAINGAFFFIVEHTKS----------
           227501312                          ----------------------------EEEE-----EEEEEEEE-EE----------
           Jnet                               DKGKNTIKKVEKLNLLDNRTDKI------KKIEEFFDEDKLKENELTKLHIFNH----------
47458668   Mycoplasma mobile 163K             ENKNSKVKRQFVVRKGVNVSLIN------NDKLADLFAEKEDTIMYRANPWVINLASQIFNEYTE----------
284931710  Mycoplasma gallisepticum str. F    ENNFYKLEKVNLFSRKDLDKFIL------EDLNEESEKNKKNENGSRKRILT----------
71894592   Mycoplasma synoviae 53             ---YTKIIKKKLVTSK----------NEELKKYFENPFGKKADGKSEYTVIMAQSHLSEFNKLKEIFE-
363542550  Mycoplasma ovipneumoniae SC01      ITNYYKKEKFSILDDKDIYIRLL------RE-REKFVINQSNPE----------VIDQIIEIIESYGKENNIPSRDEA
384393286  Mycoplasma canis PG 14             YDGKQYKINKLDIRTKEGIKVFA------KLAFSKKDSDRERLLVYLN----------
238924075  Eubacterium rectale ATCC 33656     QKITTDEYTIGKIKDIYTLDGNE------AFKKKQDKLLAKDIL----------
315149830  Enterococcus faecalis TX0012       KDKADETVVLGKIKDIYTQDGYD------AFMKIYKKDKSKFLMYRH----------
116627542  Streptococcus thermophilus LMD-9   STKKDNSTYIVQTIKDIYAKDN------TTLKKQFDKSPEKFLMYQH----------
315659848  Staphylococcus lugdunensis M23590  --YSTKVIDGKEKVVKKYKDIY------DPKFTALAEDIINNAYQEKYLMALH----------
160915782  Eubacterium dolichum DSM 3991      IRVKEIDGKLIKLKRKSISEITA------ESINSIYTD----------
325677756  Ruminococcus albus 8               ----------RSARDYETAGWI------TKVPLTDLKLNKDNEIEGTYDKDS----------
225377804  Roseburia inulinivorans DSM 16841  DEKGIVLSRVSITKLKINKKGQV------ATGDGIYDPENSNNG----------
336393381  Lactobacillus coryniformis KCTC 3535 ----------RSSKKISNQIAVS------KKPLNSIKLKDLEKMQGRDT----------
310780384  Ilyobacter polytropus DSM 2926      VSFNEALKKPQAIVEMDLKKKIL------AMLEIGYDTKRIKNYFENKDTWQDINPS----------
301311869  Bacteroides sp. 20 3                KNDP-VRKKKKKYIYDDYAQKVI------AKIVQELGLSSSMRKQDDEKLNKFINEAKVSAGVNKNLATTNKTI----
385811609  Ignavibacterium album JCM 16511     DMRKIGTVSKSAYRDALLKRLYENNDPKKAFAGKNSLDKQPIWLDKEQMRKV----------
60683389   Bacteroides fragilis NCTC 9343     LRKNILEAFDSLNVIKDRHKLTT------EEPMKRYDKEIRQKLWLHRIGNT----------
319957206  Nitratifractor salsuginis DSM 16511
```

Figure S2 (Continued)

Figure S2 (Continued)

Figure S2 (Continued)

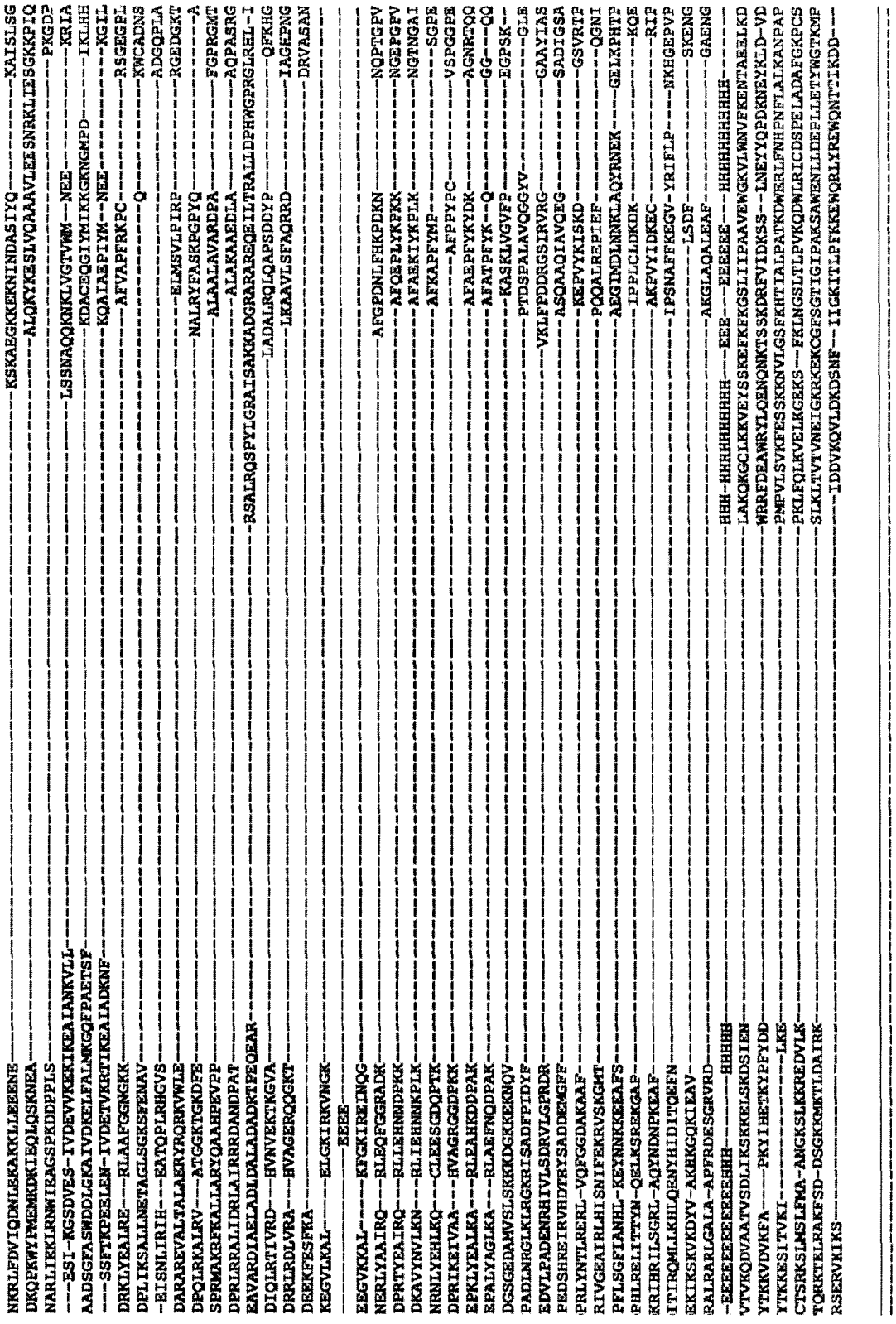


Figure S2 (Continued)

Figure S2 (Continued)

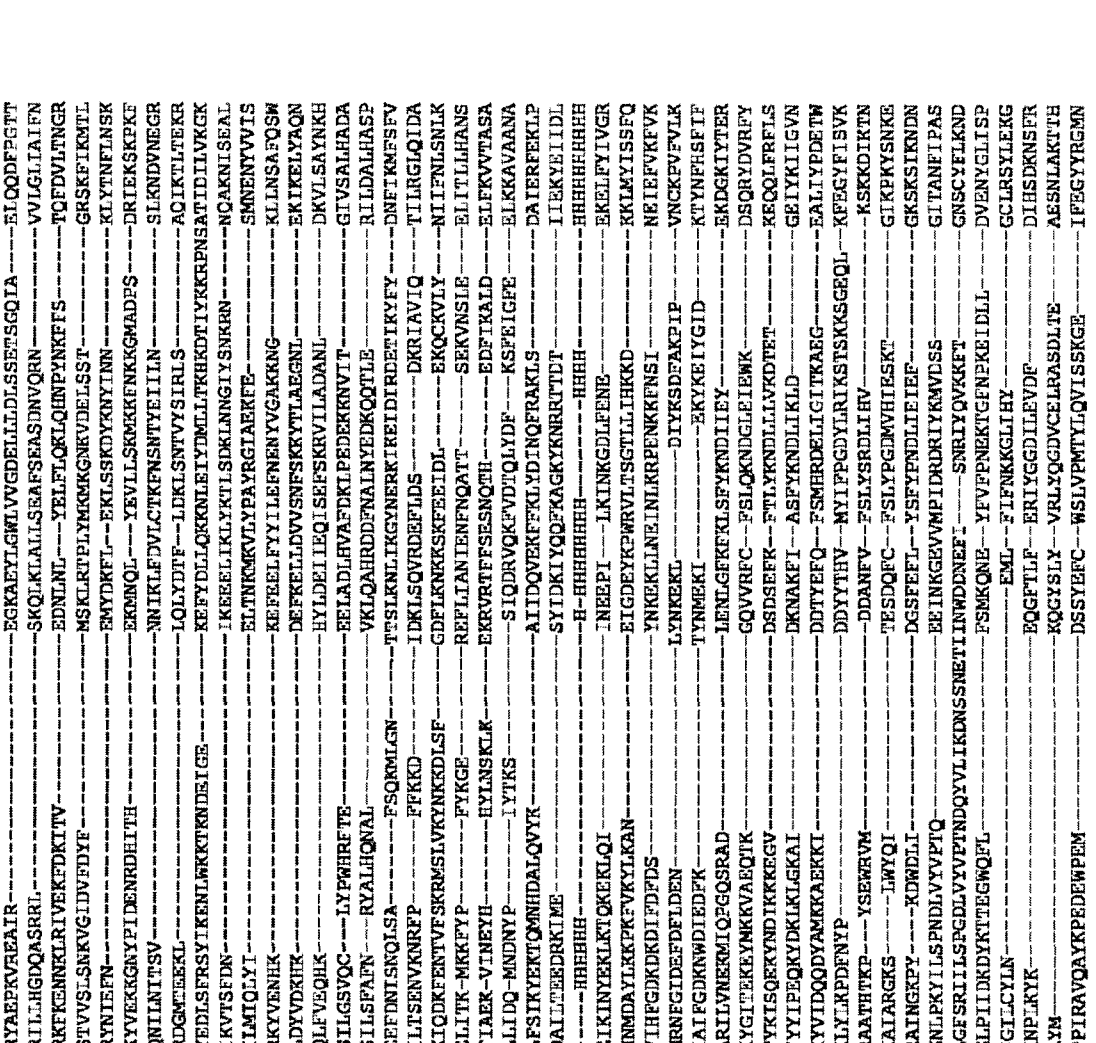


Figure S2 (Continued)

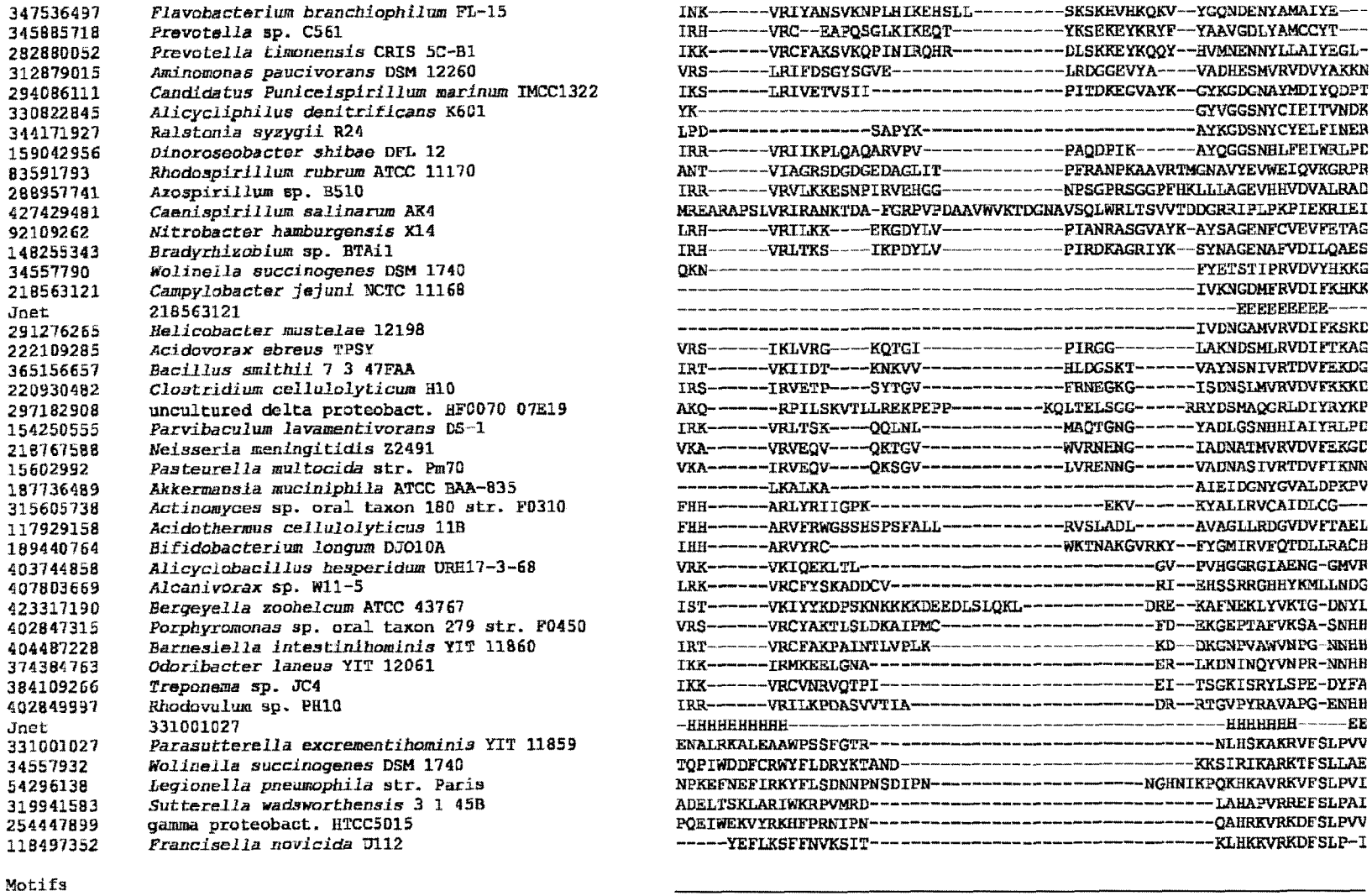


Figure S2 (Continued)

SUBSTITUTE SHEET (RULE 26)

Figure S2 (Continued)

```
227494853  Actinomyces coleocanis DSM 15436#  HFTVAGFFSPSR----------
328956315  Coriobacterium glomerans PW2       G---------STNM----------
227824983  Acidaminococcus sp. D21            STFTKLSPEEQVQTL----------
303229466  Veillonella atypica ACS-134-V-Col7a ----------EEQS----------
34762592   Fusobacterium nucleatum ATCC 49256 EKEKKLKLWEKS----------
374307738  Filifactor alocis ATCC 35896       IKLEDLIDKINV----------
320528778  Solobacterium moorei F0204         EKFSELDILEQC----------
291520705  Coprococcus catus GD-7             AKFIGLSNEDQC----------
42525843   Treponema denticola ATCC 35405     EKFKSLIIENQF----------
304438954  Peptoniphilus duerdenii ATCC BAA-1640 DKFKEISIEEKI----------
224543312  Catenibacterium mitsuokai DSM 15897 KEEKANTIKQML----------
116628213  Streptococcus thermophilus LMD-9   QNHSIDELCSSFIGPTGSEKK----------
24379809   Streptococcus mutans UA159         NGEDLKELASSF----------
13622193   Streptococcus pyogenes SF370       RDKPIREQAENI----------
310286728  Bifidobacterium bifidum S17        KTANLSIVGMTG----------
366983953  Oenococcus kitaharae DSM 17330     ASSDLKKINLSS----------
422884106  Streptococcus sanguinis SK49       KSGAPKDINDFFDNKCTV----------
339625081  Fructobacillus fructosus KCTC 3544 AYQAPVKIISK-KVSDW----------
306821691  Eubacterium yurii ATCC 43715       E---VDLSDIGGSKST----------
336394882  Lactobacillus farciminis KCTC 3681 T-SAHLIFNNI-EKKAFG----------
323463801  Staphylococcus pseudintermedius ED99 TKSD----------KIGSRK----------
389815359  Planococcus antarcticus DSM 14505  Q----------RSDTFG----------
258509199  Lactobacillus rhamnosus GG         INTDGNKIGKTE----------
169823755  Finegoldia magna ATCC 29328        DTLDNKQLYQLL----------
Jnet       227501312                          ----------HHHHH----------
47458868   Mycoplasma mobile 163K             DE----------KPQK----------
284931710  Mycoplasma gallisepticum str. F    N----------LNDV----------
71894592   Mycoplasma synoviae 53             PGALLLNFENQQ----------
363542550  Mycoplasma ovipneumoniae SC01      KGSILKKKSLDIDDFKETKETEEGN-
384393286  Mycoplasma canis PG 14             PGTILLDKQNKE----------
238924075  Eubacterium rectale ATCC 33656     LVSRTMPKQRNY----------
315149830  Enterococcus faecalis TX0012       NFQ----------SANS----------
116627542  Streptococcus thermophilus LMD-9   RTMPK----------QKEY----------
315659848  Staphylococcus lugdunensis M23590  SD----------TRNM----------
160915782  Eubacterium dolichum DSM 3991      ENFNFTFHAGET----------
325677756  Ruminococcus albus 8               NV----------NENS----------
225377804  Roseburia inulinivorans DSM 16841  LVNGGLLLQ----------
336393381  Lactobacillus coryniformis KCTC 3535 NNTSVVPIKNFY----------
310780384  Ilyobacter polytropus DSM 2926     KLTKTEIPEVNL----------
301311869  Bacteroides sp. 20 3               TANLIFALPKAT----------
385811609  Ignavibacterium album JCM 16511    IASLILSYSASN----------
60683389   Bacteroides fragilis NCTC 9343     NLFRVQKFSLKN----------
319957206  Nitratifractor salsuginis DSM 16511 QGSKYIALFNFR----------
187250660  Elusimicrobium minutum Pei191      NNTGSAPENKVF----------
```

Figure S2 (Continued)

Figure S2 (Continued)

```
325972003  Sphaerochaeta globus str. Buddy                VRLPNAKPGRTF----------
296446027  Methylosinus trichosporium OB3b                R-------SVGA----------
347536497  Flavobacterium branchiophilum FL-15            RIYIIEGLSIQRIV-RPSGKVDEYGV
345885718  Prevotella sp. C561                            RLYKINRFESQ----------SN
282880052  Prevotella timonensis CRIS 5C-B1               RLYKVVGIEKD----------GR
312879015  Aminomonas paucivorans DSM 12260               R-------GDGR----------
294086111  Candidatus Puniceispirillum marinum IMCC1322   L-------SGSK----------
330822845  Alicycliphilus denitrificans K601              IS------GSNGQ---------
344171927  Ralstonia syzygii R24                          M-------SEMK----------
159042956  Dinoroseobacter shibae DFL 12                  IANG----------------
83591793   Rhodospirillum rubrum ATCC 11170               VDGRLALMPARL----------
288957741  Azospirillum sp. B510                          LEP-----SSNS----------
427429481  Caenispirillum salinarum AK4                   ----------------------
92109262   Nitrobacter hamburgensis X14                   LDA-----AAGR----------
148255343  Bradyrhizobium sp. BTAi1                       LSP-----SNNL----------
34557790   Wolinella succinogenes DSM 1740                GANMVLSSINNSPCEGFTCT----P
218563121  Campylobacter jejuni MCTC 11168                SSTVSLIVSKSD----------
Jnet                                                      ---EEEEEEE-----------
291276265  Helicobacter mustelae 12198                    KSTNSAKATIKL----------
222109285  Acidovorax ebreus TPSY                         R-------STGA----------
365156657  Bacillus smithii 7 3 47FAA                     -S------TNEE----------
220930482  Clostridium cellulolyticum H10                 R-------GTGS----------
297182908  uncultured delta proteobact. HF0070 07E19      I-------ANGQ----------
154250555  Parvibaculum lavamentivorans DS-1              VN------ASGQ----------
218767588  Neisseria meningitidis Z2491                   R-------GTGN----------
15602992   Pasteurella multocida str. Pm70                R-------ATGN----------
187736489  Akkermansia muciniphila ATCC BAA-835           DP------KHAG----------
315605738  Actinomyces sp. oral taxon 180 str. F0310      SWRVSALDTPSK----------
117929158  Acidothermus cellulolyticus 11B                SAEQAEVLRTER----------
189440764  Bifidobacterium longum DJ010A                  LRIRPRYLAASG----------
403744858  Alicyclobacillus hesperidum URH17-3-68         RIM-------------YFVSAN
407803669  Alcanivorax sp. W11-5                          ----------------QILGDG
423317190  Bergeyella zoohelcum ATCC 43767                GKNIYVVQKFSKKQ------IYFIKHT
402847315  Porphyromonas sp. oral taxon 279 str. F0450    SEHLYRVQKMSSSY------YVFRYH
404487228  Barnesiella intestinihominis YIT 11860         NKYLYRVQKLSKSD------YSFRYH
374384763  Odoribacter laneus YIT 12061                   SKHLYRVQKLSGMY------YTFRHH
384109266  Treponema sp. JC4                              KLDIRPV---------------YAVSYC
402849997  Rhodovulum sp. PH10                            NRV-----------------RLSPHN
Jnet                                                      --EEEEE---HHHHHHHH----------
331001027  Parasutterella excrementihominis YIT 11859     SRRYLRVEMPGEKFLAWFGR---------
345577932  Wolinella succinogenes DSM 1740                PISAIDKSAKRMIFRALFEKEKGKRK
54296138   Legionella pneumophila str. Paris              TRRYIRITQSLADFIKTIDE---------
319941583  Sutterella wadsworthensis 3 1 45B              GRRYVRVETTFIQASHWFEQSVENWA
254447899  gamma proteobact. HTCC5015                     NRFYIRFTVDSDQFKRHF-----------
118497352  Francisella novicida U112                      SRPKVRVKLDYVIDDDS------------

Motifs                                                    ----------------------------
informative positions
```

Figure S2 (Continued)

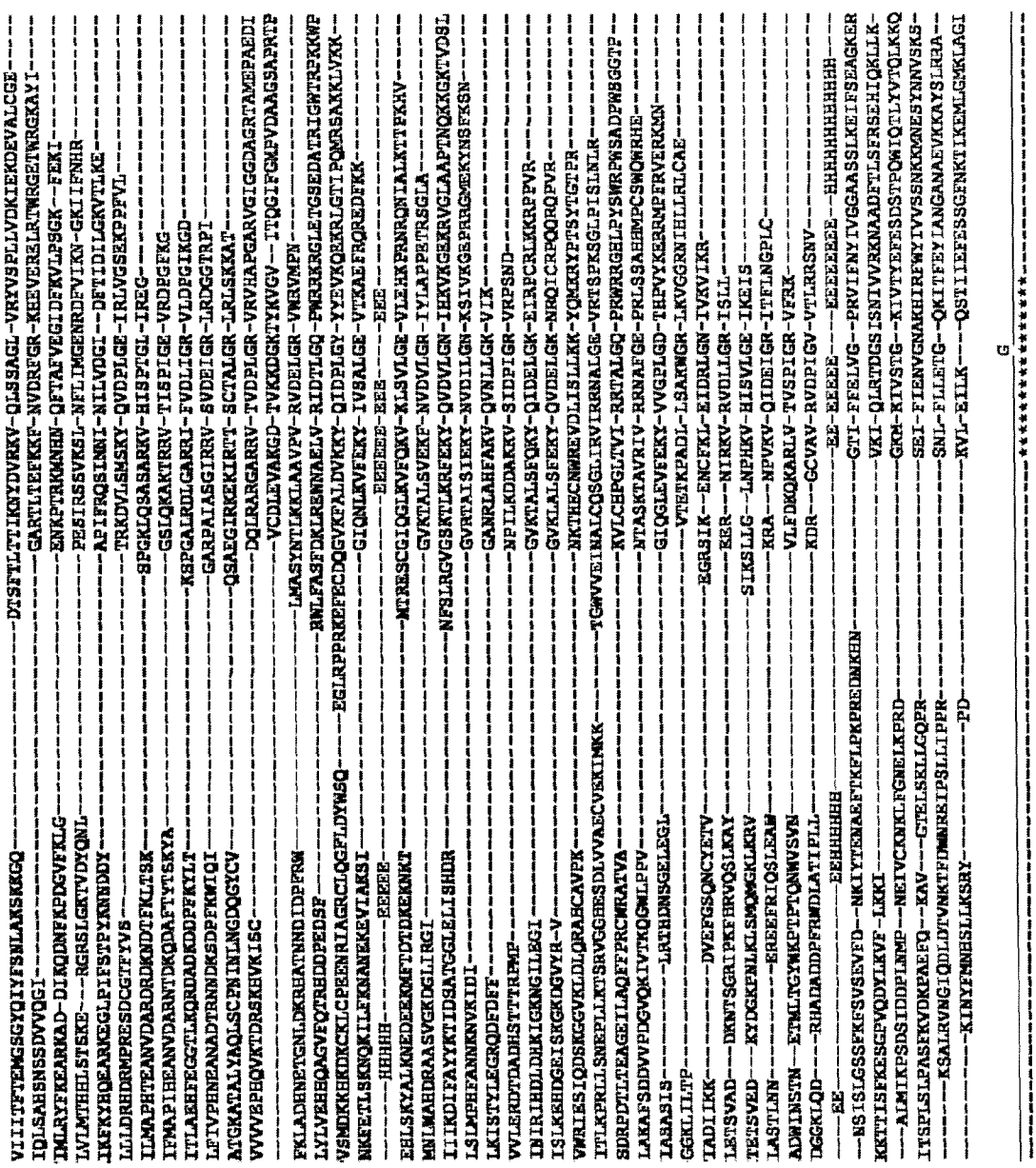


Figure S2 (Continued)

```
16130662  Escherichia coli str. K-12 substr. MG1655
11500011  Archaeoglobus fulgidus DSM 4304
18977490  Pyrococcus furiosus DSM 3638
46447796  Desulfovibrio vulgaris str. Hildenborough
54296139  Legionella pneumophila str. Paris
331001028 Parasutterella excrementihominis YIT 11859
319941582 Sutterella wadsworthensis 3 1 45B
118497353 Francisella novicida U112
34557933  Wolinella succinogenes DSM 1740
187736488 Akkermansia muciniphila ATCC BAA-835
407803668 Alcanivorax sp. W11-5
325972002 Sphaerochaeta globus str. Buddy
187250661 Elusimicrobium minutum Pei191
47458867  Mycoplasma mobile 163K
363542551 Mycoplasma ovipneumoniae SC01
71894593  Mycoplasma synoviae 53
384393287 Mycoplasma canis PG 14
284931711 Mycoplasma gallisepticum str. F
117929157 Acidothermus cellulolyticus 11B
317482065 Bifidobacterium sp. 12 1 47BFAA
315605739 Actinomyces sp. oral taxon 180 str. F0310
294086112 Candidatus Puniceispirillum marinum IMCC1322
282880053 Prevotella timonensis CRIS 5C-B1
374384762 Odoribacter laneus YIT 12061
423317188 Bergeyella zoohelcum ATCC 43767
385811610 Ignavibacterium album JCM 16511
402847305 Porphyromonas sp. oral taxon 279 str. F0450
345885719 Prevotella sp. C561
404487227 Barnesiella intestinihominis YIT 11860
301311870 Bacteroides sp. 20 3
60683388  Bacteroides fragilis NCTC 9343
402849996 Rhodovulum sp. PH10
154250554 Parvibaculum lavamentivorans DS-1
159042957 Dinoroseobacter shibae DFL 12
288957740 Azospirillum sp. B510
330822846 Alicycliphilus denitrificans K601
344171926 Ralstonia syzygii R24
319957207 Nitratifractor salsuginis DSM 16511
```

Figure S3

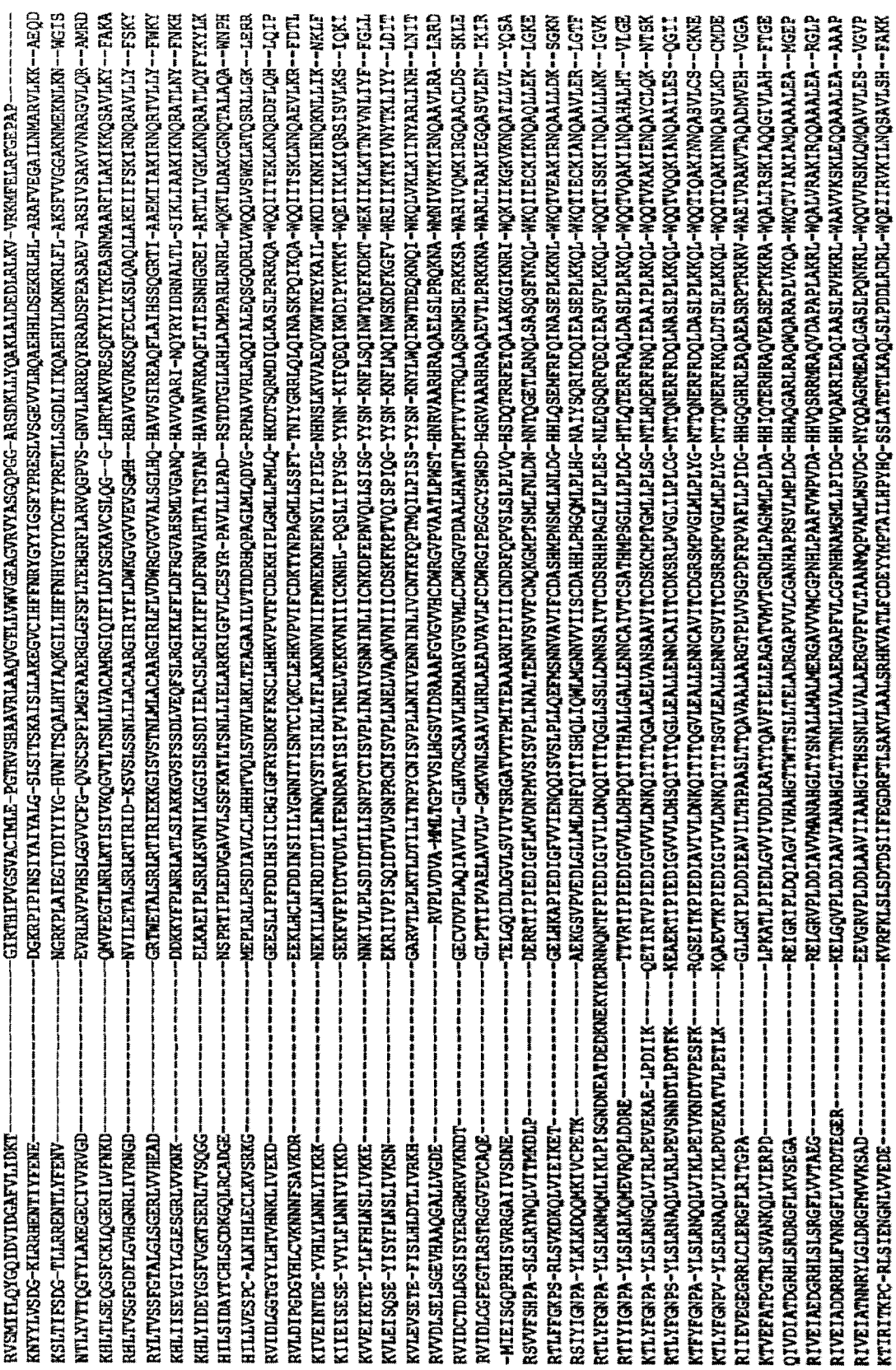


Figure S3 (Continued)

```
315149831  Enterococcus faecalis TX0012                    ----------MLLGF----------
116627543  Streptococcus thermophilus LMD-9                ----------MTW----------
238924076  Eubacterium rectale ATCC 33656                  ----------MAF----------
218563120  Campylobacter jejuni NCTC 11168                 ----------MSYDEAF----------
291276264  Helicobacter mustelae 12198                     ----------MFDEAF----------
325677757  Ruminococcus albus 8                            ----------MW----------
296446028  Methylosinus trichosporium OB3b                 ----------MAW----------
34557789   Nolinella succinogenes DSM 1740                 ----------MAAW----------
222109284  Acidovorax abreus TPSY                          ----------MVW----------
336393380  Lactobacillus coryniformis KCTC 3535            ----------MAF----------
218767587  Neisseria meningitidis Z2491                    ----------MTW----------
15602991   Pasteurella multocida Pm70                      MMSSVKIDKFADLSKDNLCVKLYHPDFLYIGNRDFLGVIMSW----------
310780383  Ilyobacter polytropus DSM 2926                  ----------MGW----------
312879014  Aminomonas paucivorans DSM 12260                ----------MGW----------
403744859  Alicyclobacillus hesperidum URH17-3-68          ----------MSH----------
421874296  Brevibacillus laterosporus GI-9                 ----------
220930481  Clostridium cellulolyticum H10                  ----------MGW----------
225377803  Roseburia inulinivorans DSM 16841               ----------MGF----------
315659847  Staphylococcus lugdunensis M23590               ----------MSW----------
160915783  Eubacterium dolichum DSM 3991                   ----------MGF----------
323463802  Staphylococcus pseudintermedius ED99            ----------MSF----------
389815358  Planococcus antarcticus DSM 14505               ----------MSW----------
328956316  Coriobacterium glomerans PW2                    ----------MAF----------
422884107  Streptococcus sanguinis SK49                    ----------MGW----------
116628212  Streptococcus thermophilus LMD-9                ----------MAGW----------
24379808   Streptococcus mutans UA159                      ----------MGW----------
13622194   Streptococcus pyogenes SF370                    ----------MAGW----------
227824982  Acidaminococcus sp. D21                         ----------MSY----------
169823756  Finegoldia magna ATCC 29328                     ----------MGW----------
320528779  Solobacterium moorei F0204                      ----------MGW----------
303229394  Veillonella atypica ACS-134-V-Col7a             ----------MSW----------
304438953  Peptoniphilus duerdenii ATCC BAA-1640           ----------MTW----------
374307737  Filifactor alocis ATCC 35896                    ----------MSW----------
256845020  Fusobacterium sp. 3 1 36A2                      ----------MSGW----------
306821690  Eubacterium yurii ATCC 43715                    ----------MSW----------
42525844   Treponema denticola ATCC 35405                  ----------MSW----------
291520706  Coprococcus catus GD-7                          ----------MSW----------
258509198  Lactobacillus rhamnosus GG                      ----------MGW----------
336394883  Lactobacillus farciminis KCTC 3681              ----------MGW----------
339625080  Fructobacillus fructosus KCTC 3544              ----------MAW----------
310286727  Bifidobacterium bifidum S17                     ----------MAW----------
366983954  Oenococcus kitaharae DSM 17330                  ----------MAW----------

informative positions
```

Figure S3 (Continued)

Figure S3 (Continued)

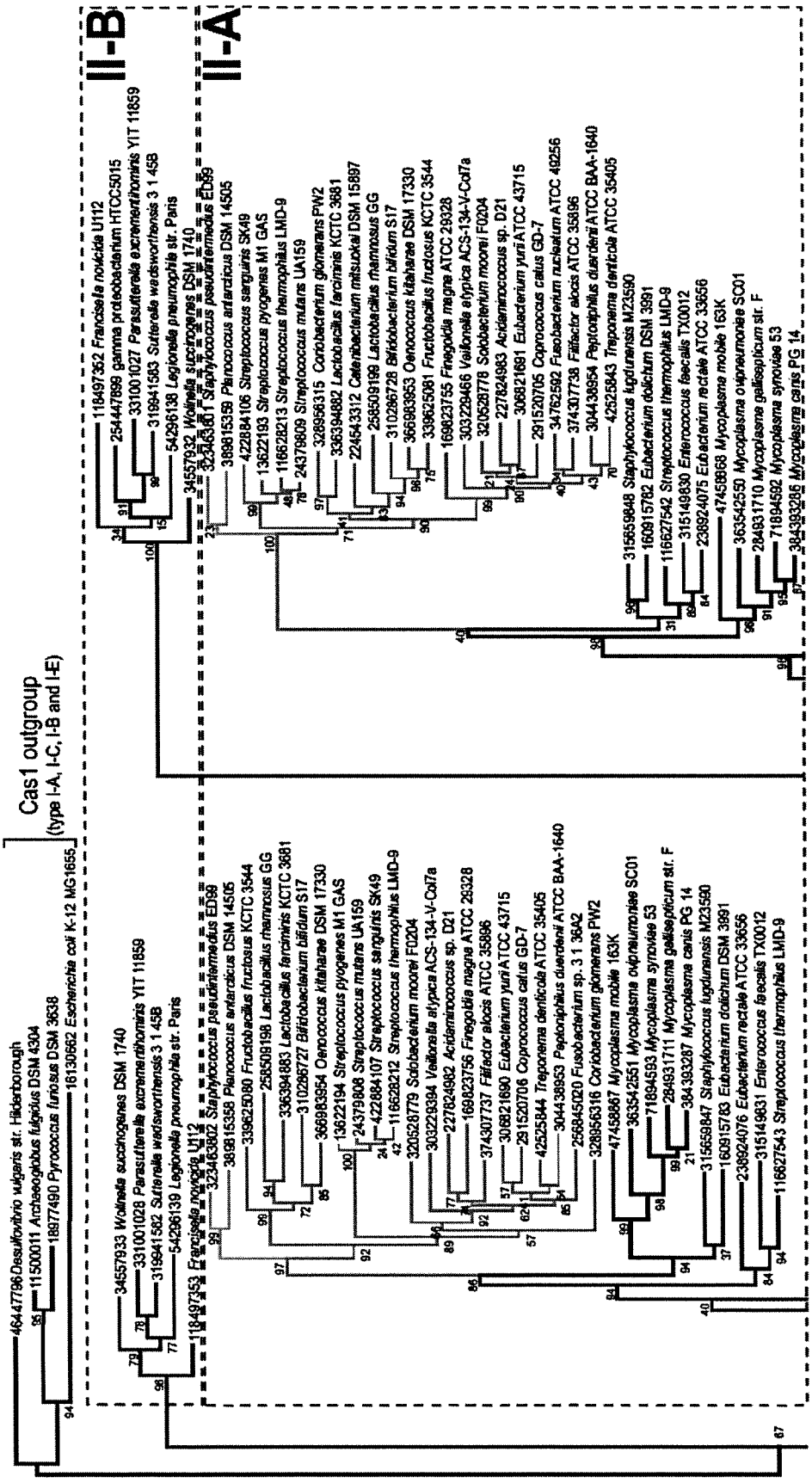

Cas1
Cas9
Cas1 outgroup
(type I-A, I-C, I-B and I-E)
II-B
II-A
46447796 Desulfovibrio vulgaris str. Hildenborough
11500011 Archaeoglobus fulgidus DSM 4304
18977490 Pyrococcus furiosus DSM 3638
18130662 Escherichia coli K-12 MG1655
34557933 Wolinella succinogenes DSM 1740
331001028 Parasutterella excrementihominis YIT 11859
319941582 Sutterella wadsworthensis 3 1 45B
54296139 Legionella pneumophila str. Paris
118497352 Francisella noviicida U112
323463802 Staphylococcus pseudintermedius ED99
389815358 Planococcus antarcticus DSM 14505
339625080 Fructobacillus fructosus KCTC 3544
258509198 Lactobacillus rhamnosus GG
336394883 Lactobacillus farciminis KCTC 3681
310286727 Bifidobacterium bifidum S17
366983954 Oenococcus kitaharae DSM 17330
13622194 Streptococcus pyogenes M1 GAS
24379808 Streptococcus mutans UA159
422884107 Streptococcus sanguinis SK49
116628212 Streptococcus thermophilus LMD-9
320528779 Solobacterium moorei F0204
303229394 Veillonella atypica ACS-134-V-Col7a
227824982 Acidaminococcus sp. D21
169823756 Finegoldia magna ATCC 29328
374307737 Filifactor alocis ATCC 35896
306821690 Eubacterium yurii ATCC 43715
291520706 Coprococcus catus GD-7
42525844 Treponema denticola ATCC 35405
304438953 Peptoniphilus duerdenii ATCC BAA-1640
256845020 Fusobacterium sp. 3 1 36A2
328956316 Coriobacterium glomerans PW2
47458867 Mycoplasma mobile 163K
363542551 Mycoplasma ovipneumoniae SC01
71894593 Mycoplasma synoviae 53
284931711 Mycoplasma gallisepticum str. F
384393287 Mycoplasma canis PG 14
315659847 Staphylococcus lugdunensis M23590
160915783 Eubacterium dolichum DSM 3991
238924076 Eubacterium rectale ATCC 33656
315149831 Enterococcus faecalis TX0012
116627543 Streptococcus thermophilus LMD-9
118497352 Francisella noviicida U112
254447899 gamma proteobacterium HTCC5015
331001027 Parasutterella excrementihominis YIT 11859
319941583 Sutterella wadsworthensis 3 1 45B
54296138 Legionella pneumophila str. Paris
34557932 Wolinella succinogenes DSM 1740
323463801 Staphylococcus pseudintermedius ED99
389815359 Planococcus antarcticus DSM 14505
422884106 Streptococcus sanguinis SK49
13622193 Streptococcus pyogenes M1 GAS
116628213 Streptococcus thermophilus LMD-9
24379809 Streptococcus mutans UA159
328956315 Coriobacterium glomerans PW2
336394882 Lactobacillus farciminis KCTC 3681
224543312 Catenibacterium mitsuokai DSM 15897
258509199 Lactobacillus rhamnosus GG
310286728 Bifidobacterium bifidum S17
366983953 Oenococcus kitaharae DSM 17330
339625081 Fructobacillus fructosus KCTC 3544
169823755 Finegoldia magna ATCC 29328
303229466 Veillonella atypica ACS-134-V-Col7a
320528778 Solobacterium moorei F0204
227824983 Acidaminococcus sp. D21
306821691 Eubacterium yurii ATCC 43715
291520705 Coprococcus catus GD-7
34762592 Fusobacterium nucleatum ATCC 49256
374307738 Filifactor alocis ATCC 35896
304438954 Peptoniphilus duerdenii ATCC BAA-1640
42525843 Treponema denticola ATCC 35405
315659848 Staphylococcus lugdunensis M23590
160915782 Eubacterium dolichum DSM 3991
116627542 Streptococcus thermophilus LMD-9
315149830 Enterococcus faecalis TX0012
238924075 Eubacterium rectale ATCC 33656
47458868 Mycoplasma mobile 163K
363542550 Mycoplasma ovipneumoniae SC01
284931710 Mycoplasma gallisepticum str. F
71894592 Mycoplasma synoviae 53
384393285 Mycoplasma canis PG 14

Figure. S4

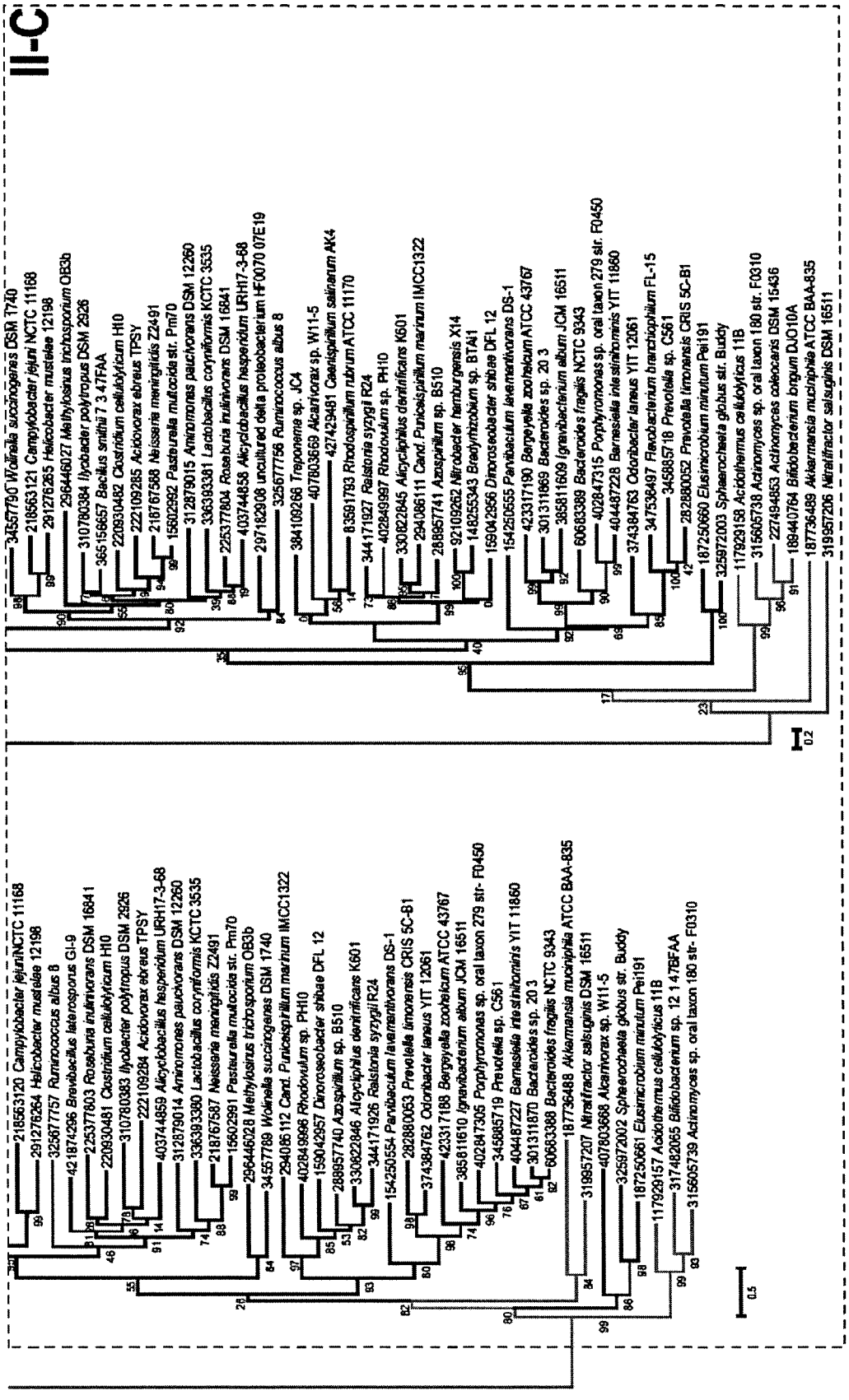


Figure. S4
(Continued)

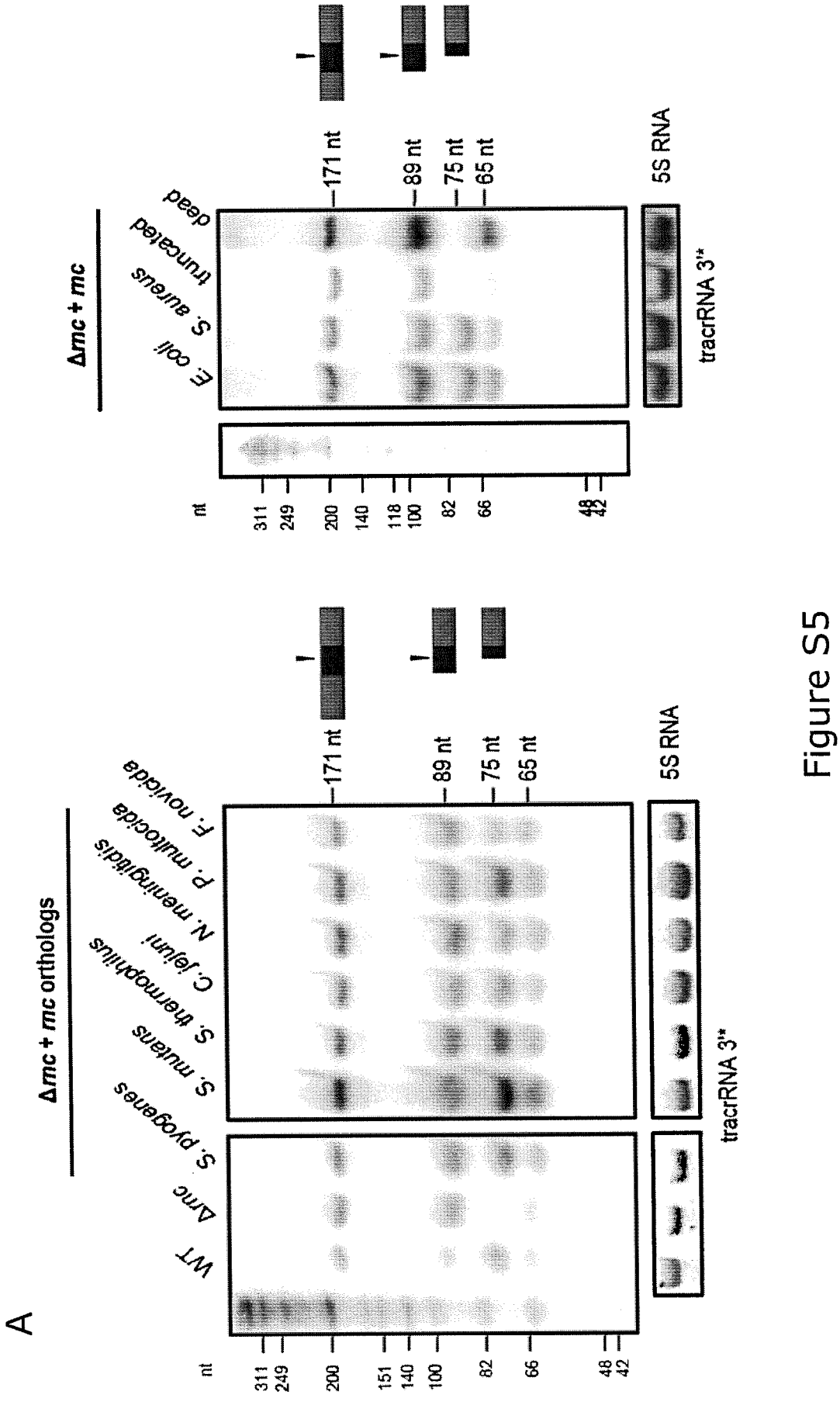

A
Δmc + mc orthologs
WT
Δmc
S. pyogenes
S. mutans
S. thermophilus
C. jejuni
N. meningitidis
P. multocida
F. novicida
171 nt
89 nt
75 nt
65 nt
5S RNA
tracrRNA 3'*
nt
311
249
200
151
140
100
82
66
48
42
Δmc + mc
E. coli
S. aureus
truncated
dead
171 nt
89 nt
75 nt
65 nt
5S RNA
tracrRNA 3'*
nt
311
249
200
140
118
100
82
66
48
42

Figure S5

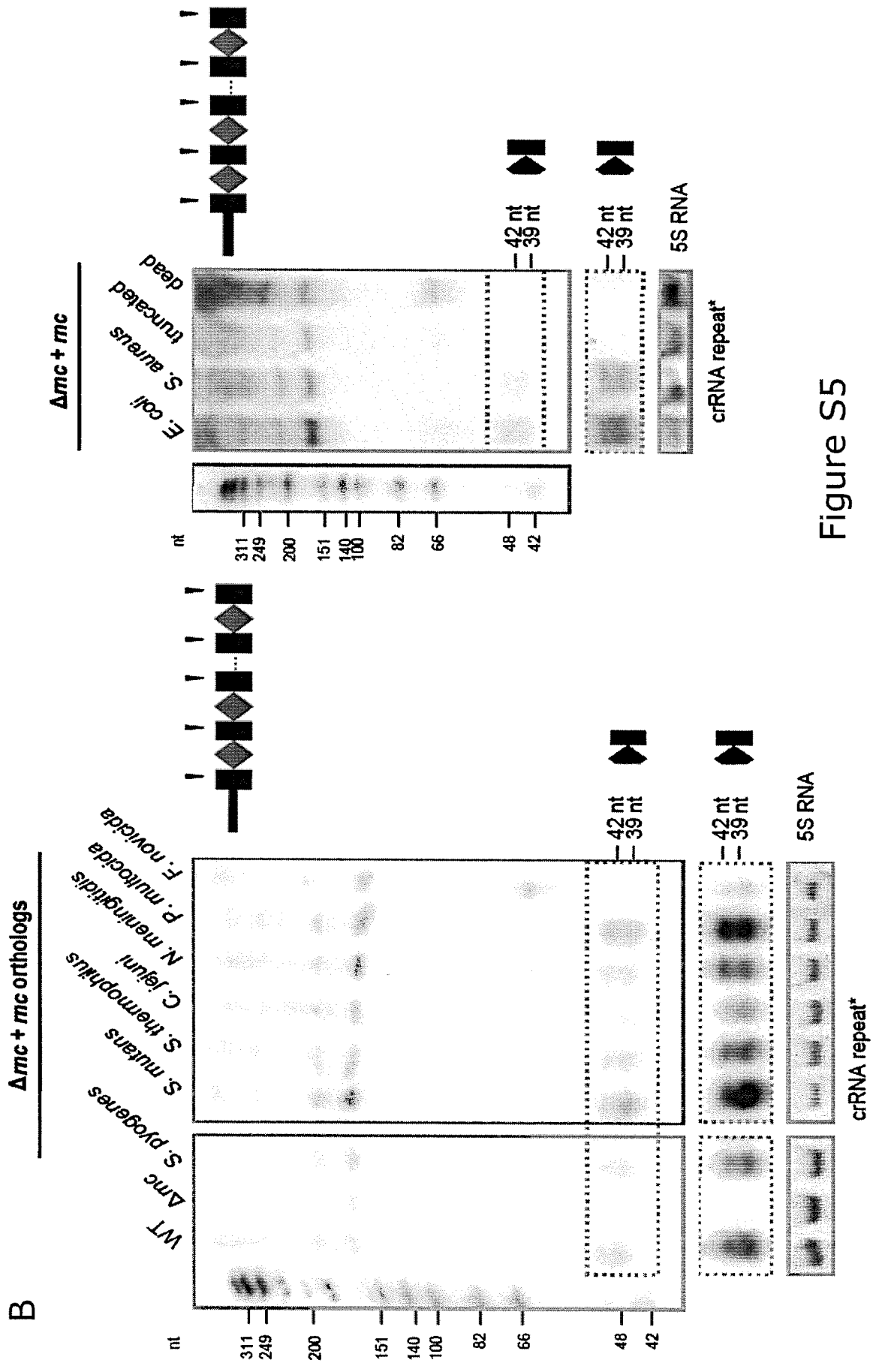

B
Δmc + mc orthologs
WT
Δmc
S. pyogenes
S. mutans
S. thermophilus
C. jejuni
N. meningitidis
P. multocida
F. novicida
nt
311
249
200
151
140
100
82
66
48
42
42 nt
39 nt
5S RNA
crRNA repeat*
Δmc + mc
E. coli
S. aureus
truncated
dead

Figure S5

15674631 Streptococcus pyogenes SF370
24379904 Streptococcus mutans UA159
116628032 Streptococcus thermophilus LMD-9
57651802 Staphylococcus aureus COL
118498035 Francisella novicida U112
218767809 Neisseria meningitidis Z2491
15601926 Pasteurella multocida Pm70
218563224 Campylobacter jejuni NCTC 11168
16130492 Escherichia coli K-12
Motifs
Cutting domain
dsRNA binding domain

Figure S6

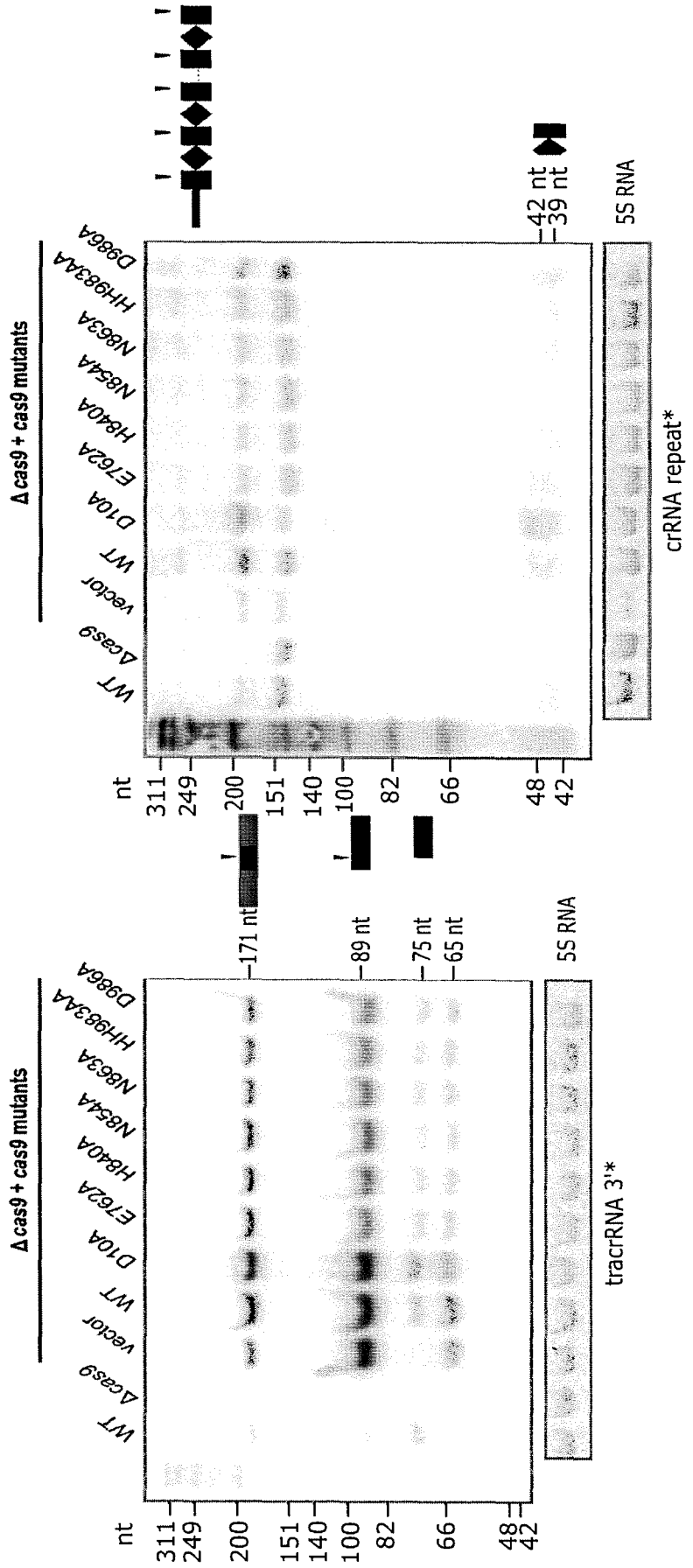

Δ cas9 + cas9 mutants
WT
Δcas9
vector
WT
D10A
E762A
H840A
N854A
N863A
HH983AA
D986A
nt
311
249
200
151
140
100
82
66
48
42
171 nt
89 nt
75 nt
65 nt
5S RNA
tracrRNA 3'*
42 nt
39 nt
5S RNA
crRNA repeat*

Figure S7

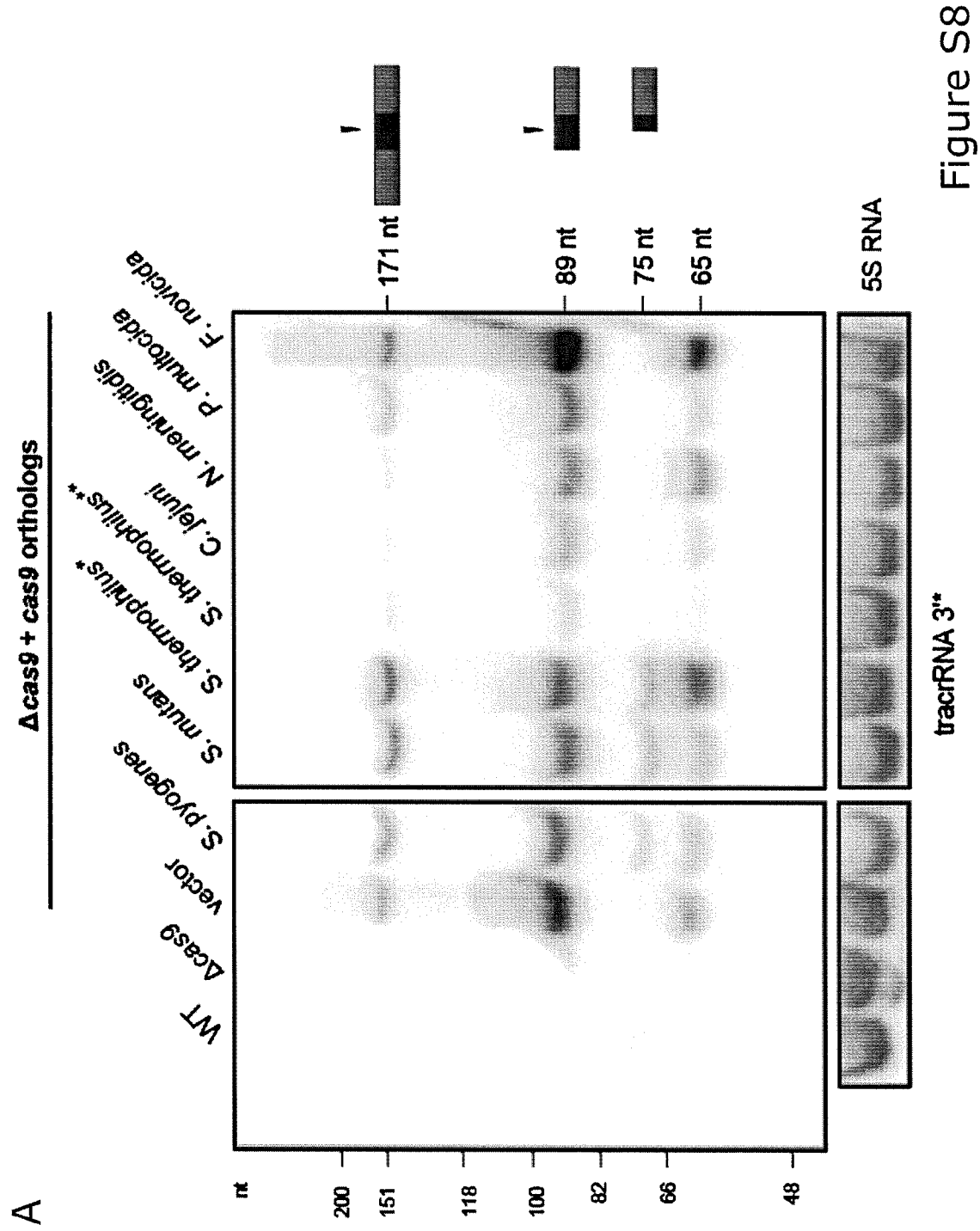

A
Δcas9 + cas9 orthologs
WT
Δcas9
vector
S. pyogenes
S. mutans
S. thermophilus*
S. thermophilus**
C. jejuni
N. meningitidis
P. multocida
F. novicida
nt
200
151
118
100
82
66
48
171 nt
89 nt
75 nt
65 nt
5S RNA
tracrRNA 3'*

Figure S8

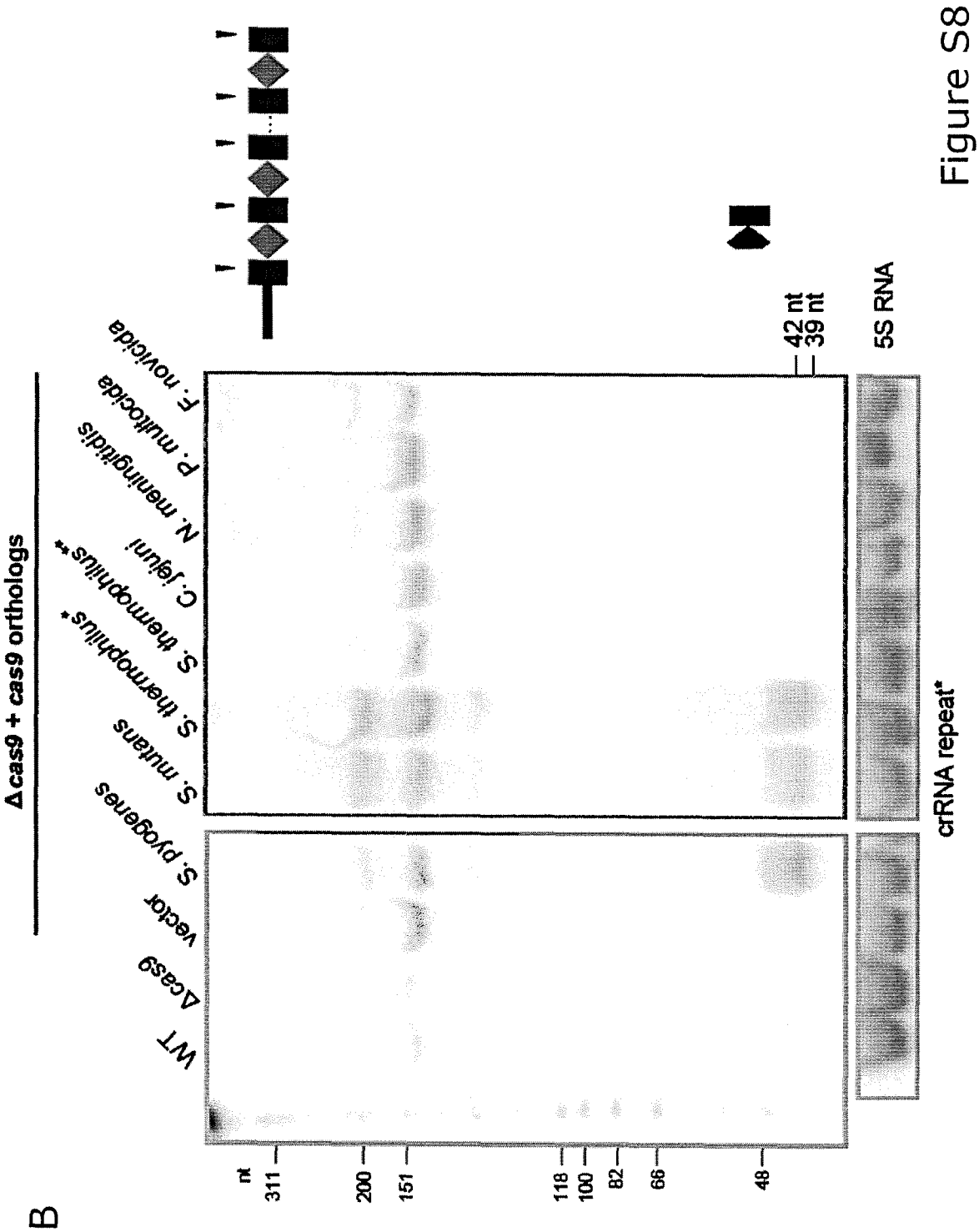

B
Δcas9 + cas9 orthologs
WT
Δcas9
vector
S. pyogenes
S. mutans
S. thermophilus*
S. thermophilus**
C. jejuni
N. meningitidis
P. multocida
F. novicida
nt
311
200
151
118
100
82
66
48
42 nt
39 nt
5S RNA
crRNA repeat*

Figure S8

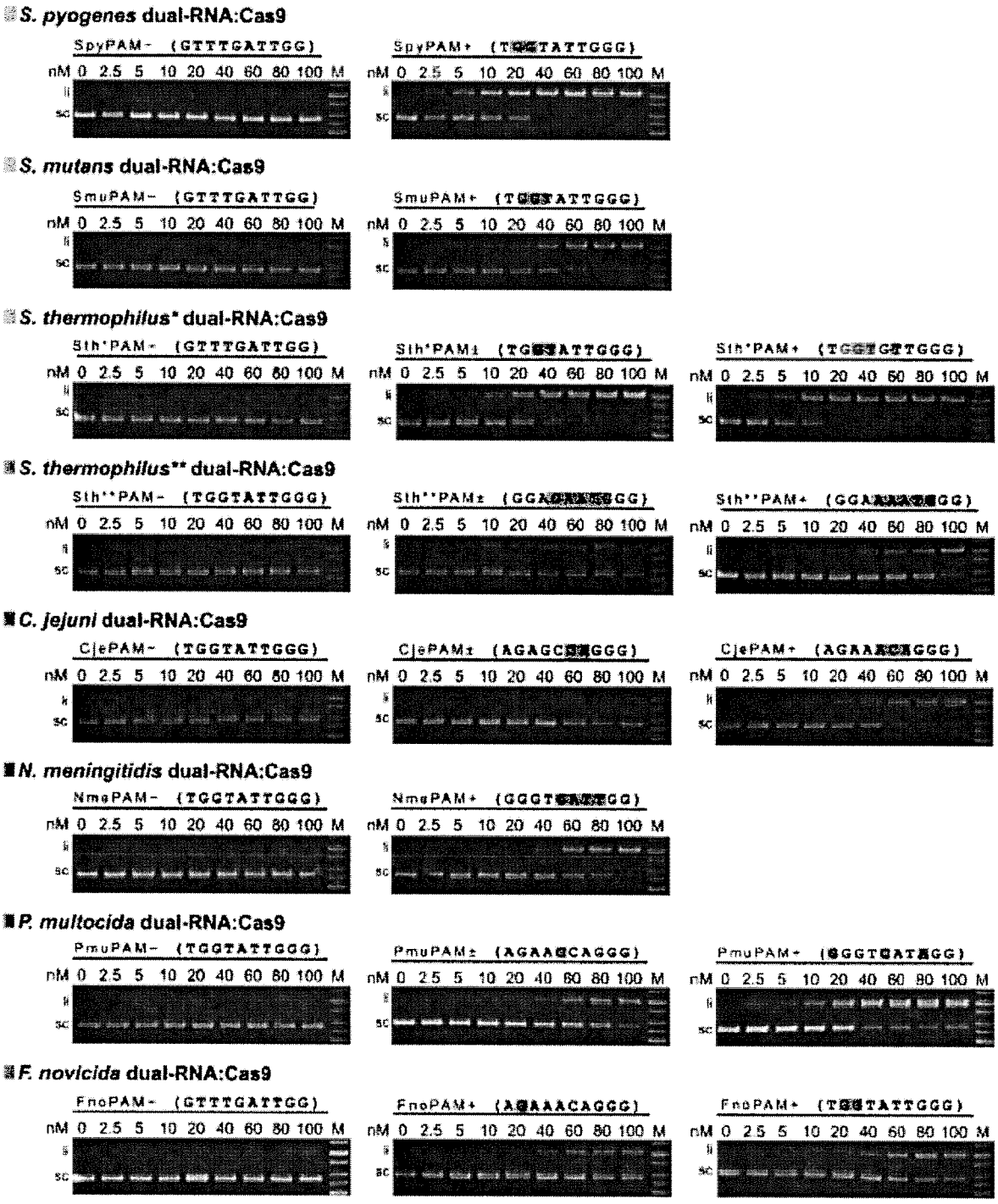

S. pyogenes dual-RNA:Cas9
SpyPAM- (GTTTGATTGG)
SpyPAM+ (TGGTATTGGG)
S. mutans dual-RNA:Cas9
SmuPAM- (GTTTGATTGG)
SmuPAM+ (TGGTATTGGG)
S. thermophilus* dual-RNA:Cas9
Sth*PAM- (GTTTGATTGG)
Sth*PAM± (TGGTATTGGG)
Sth*PAM+ (TGGTGTTGGG)
S. thermophilus** dual-RNA:Cas9
Sth**PAM- (TGGTATTGGG)
Sth**PAM± (GGA GG)
Sth**PAM+ (GGA GG)
C. jejuni dual-RNA:Cas9
CjePAM- (TGGTATTGGG)
CjePAM± (AGAGC GGG)
CjePAM+ (AGAA GGG)
N. meningitidis dual-RNA:Cas9
NmePAM- (TGGTATTGGG)
NmePAM+ (GGGT GG)
P. multocida dual-RNA:Cas9
PmuPAM- (TGGTATTGGG)
PmuPAM± (AGAAGCAGGG)
PmuPAM+ (GGGTGATGGG)
F. novicida dual-RNA:Cas9
FnoPAM- (GTTTGATTGG)
FnoPAM+ (AGAAACAGGG)
FnoPAM+ (TGGTATTGGG)
nM 0 2.5 5 10 20 40 60 80 100 M
sc

Figure S9

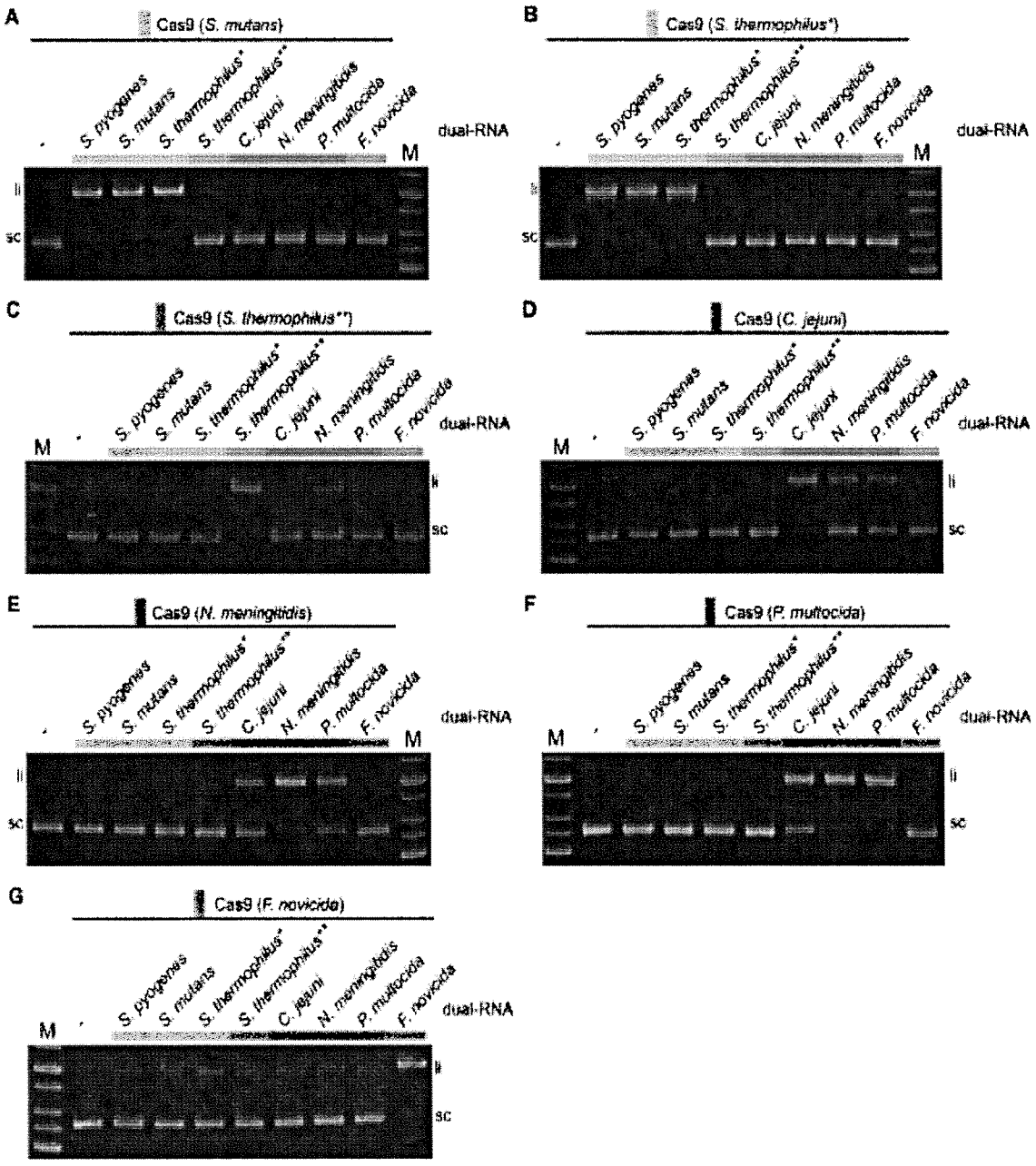

A
Cas9 (S. mutans)
B
Cas9 (S. thermophilus*)
C
Cas9 (S. thermophilus**)
D
Cas9 (C. jejuni)
E
Cas9 (N. meningitidis)
F
Cas9 (P. multocida)
G
Cas9 (F. novicida)
S. pyogenes
S. mutans
S. thermophilus*
S. thermophilus**
C. jejuni
N. meningitidis
P. multocida
F. novicida
dual-RNA
M
li
sc

Figure S10

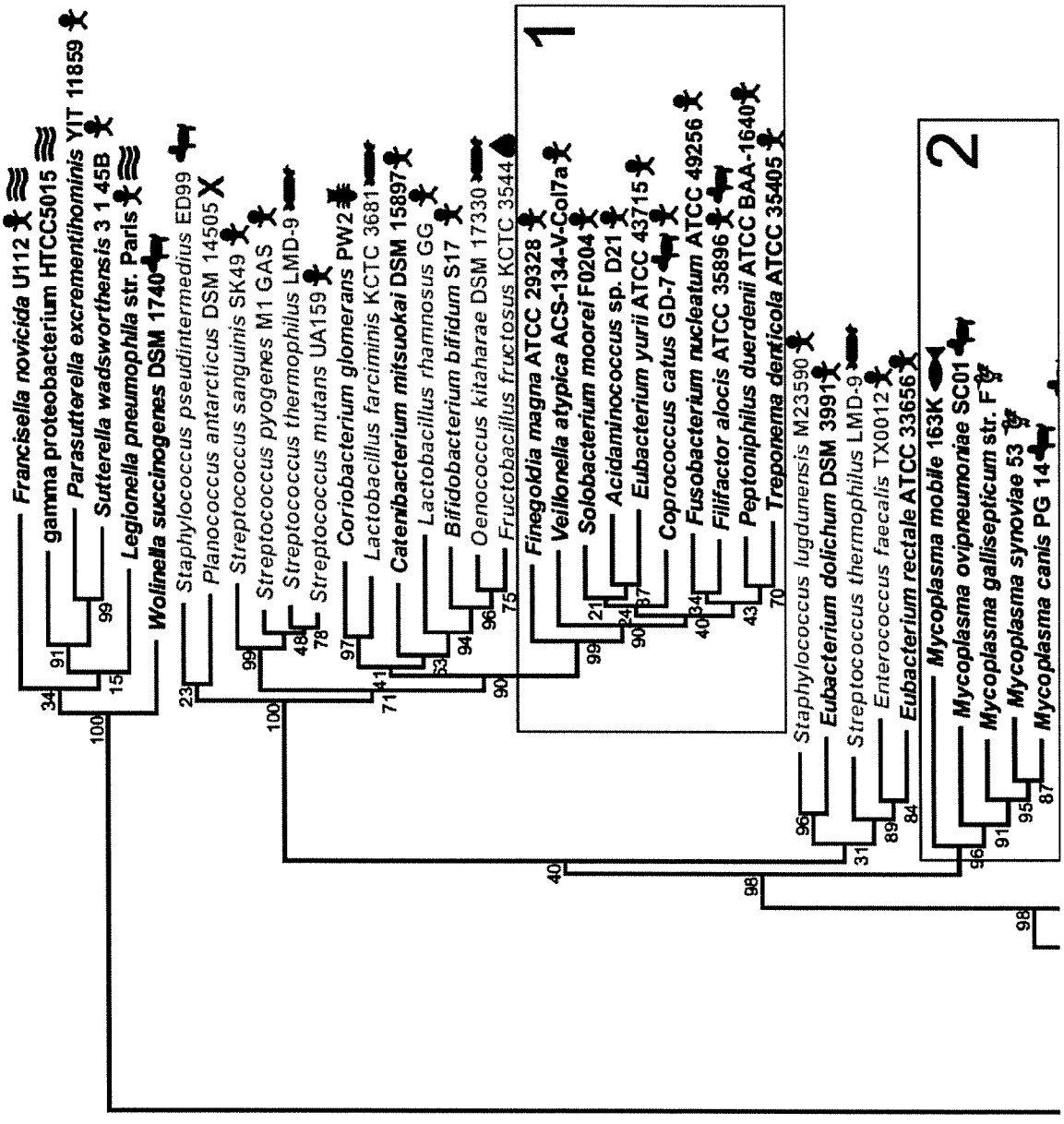

Francisella novicida U112
gamma proteobacterium HTCC5015
Parasutterella excrementihominis YIT 11859
Sutterella wadsworthensis 3 1 45B
Legionella pneumophila str. Paris
Wolinella succinogenes DSM 1740
Staphylococcus pseudintermedius ED99
Planococcus antarcticus DSM 14505
Streptococcus sanguinis SK49
Streptococcus pyogenes M1 GAS
Streptococcus thermophilus LMD-9
Streptococcus mutans UA159
Coriobacterium glomerans PW2
Lactobacillus farciminis KCTC 3681
Catenibacterium mitsuokai DSM 15897
Lactobacillus rhamnosus GG
Bifidobacterium bifidum S17
Oenococcus kitaharae DSM 17330
Fructobacillus fructosus KCTC 3544
Finegoldia magna ATCC 29328
Veillonella atypica ACS-134-V-Col7a
Solobacterium moorei F0204
Acidaminococcus sp. D21
Eubacterium yurii ATCC 43715
Coprococcus catus GD-7
Fusobacterium nucleatum ATCC 49256
Filifactor alocis ATCC 35896
Peptoniphilus duerdenii ATCC BAA-1640
Treponema denticola ATCC 35405
Staphylococcus lugdunensis M23590
Eubacterium dolichum DSM 3991
Streptococcus thermophilus LMD-9
Enterococcus faecalis TX0012
Eubacterium rectale ATCC 33656
Mycoplasma mobile 163K
Mycoplasma ovipneumoniae SC01
Mycoplasma gallisepticum str. F
Mycoplasma synoviae 53
Mycoplasma canis PG 14
1
2

Figure S11

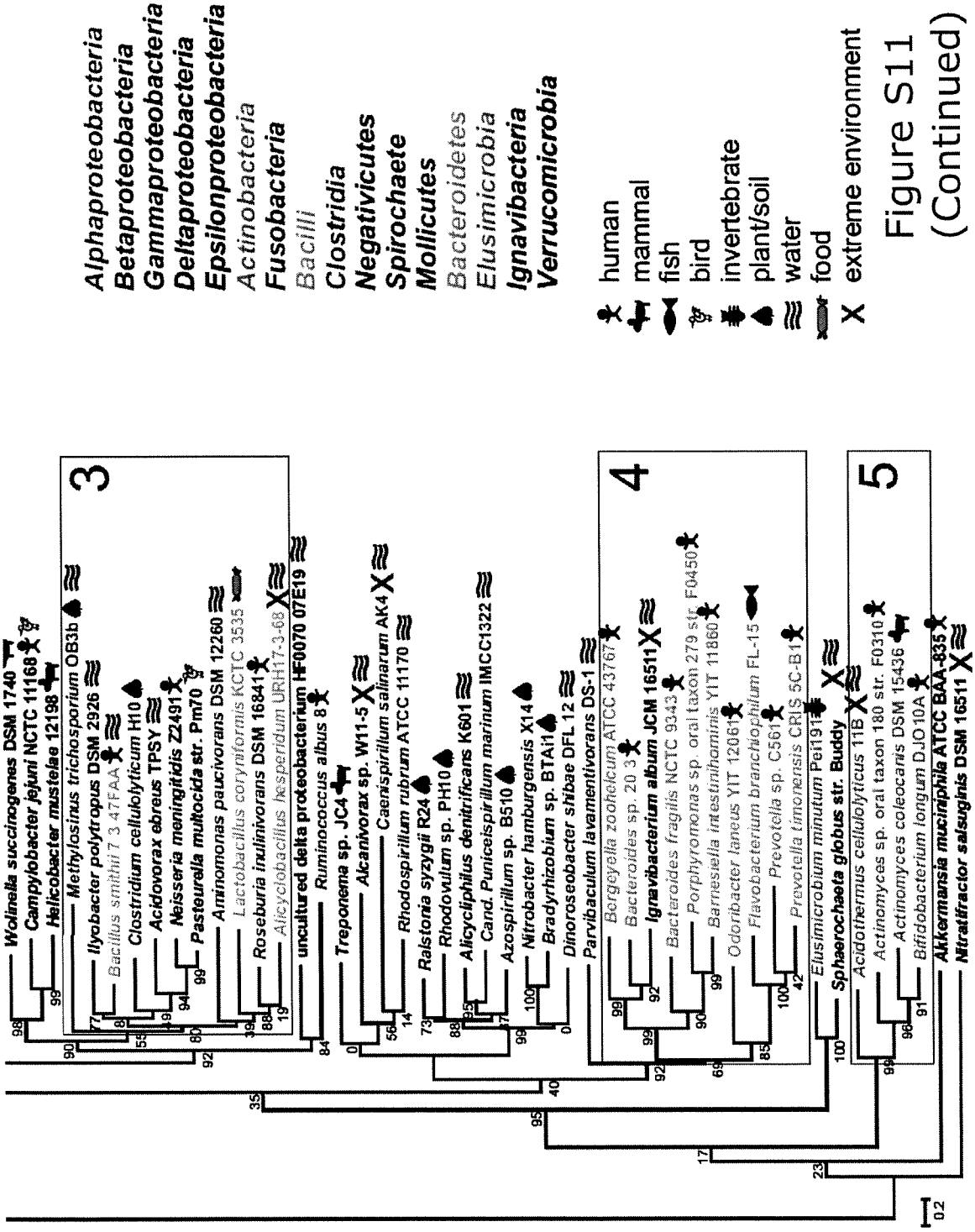


Figure S11
(Continued)

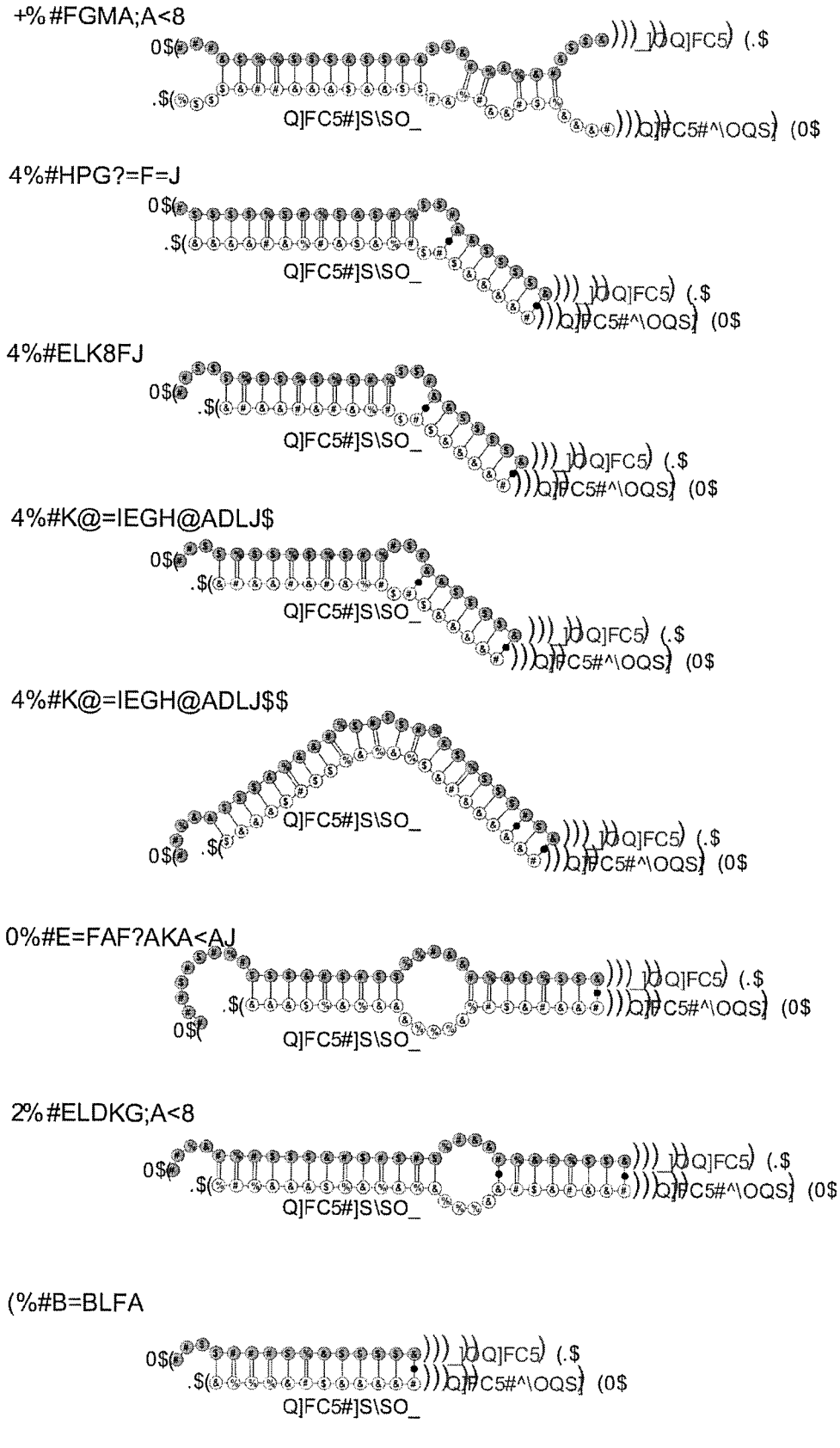

+%#FGMA;A<8
0$(
.$(
Q]FC5#]S\SO_
Q]FC5#^\OQS]
(.$
(0$
4%#HPG?=F=J
4%#ELK8FJ
4%#K@=IEGH@ADLJ$
4%#K@=IEGH@ADLJ$$
0%#E=FAF?AKA<AJ
2%#ELDKG;A<8
(%#B=BLFA

Figure S12