



(51) International Patent Classification:

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

C12N 15/10 (2006.01) C12N 15/63 (2006.01)

C12N 15/90 (2006.01) C07K 14/195 (2006.01)

(21) International Application Number:

PCT/US2018/059090

(22) International Filing Date:

02 November 2018 (02.11.2018)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

62/580,880 02 November 2017 (02.11.2017) US

62/587,381 16 November 2017 (16.11.2017) US

(71) Applicant: ARBOR BIOTECHNOLOGIES, INC.

[US/US]; 790 Memorial Drive, Suite 1B, Cambridge, Massachusetts 02139 (US).

(72) Inventors: SCOTT, David A.; 75 Richdale Avenue, Ste 6, Cambridge, Massachusetts 02140 (US). CHENG, David R.; 286 Beacon Street, Apt. 10, Boston, Massachusetts 02116 (US). YAN, Winston X.; 8 Enterprise St, #2, Boston, Massachusetts 02125 (US).

(74) Agent: WON, Doyon et al.; Fish & Richardson P.C., PO Box 1022, Minneapolis, Minnesota 55440-1022 (US).

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

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

(54) Title: NOVEL CRISPR-ASSOCIATED TRANSPOSON SYSTEMS AND COMPONENTS

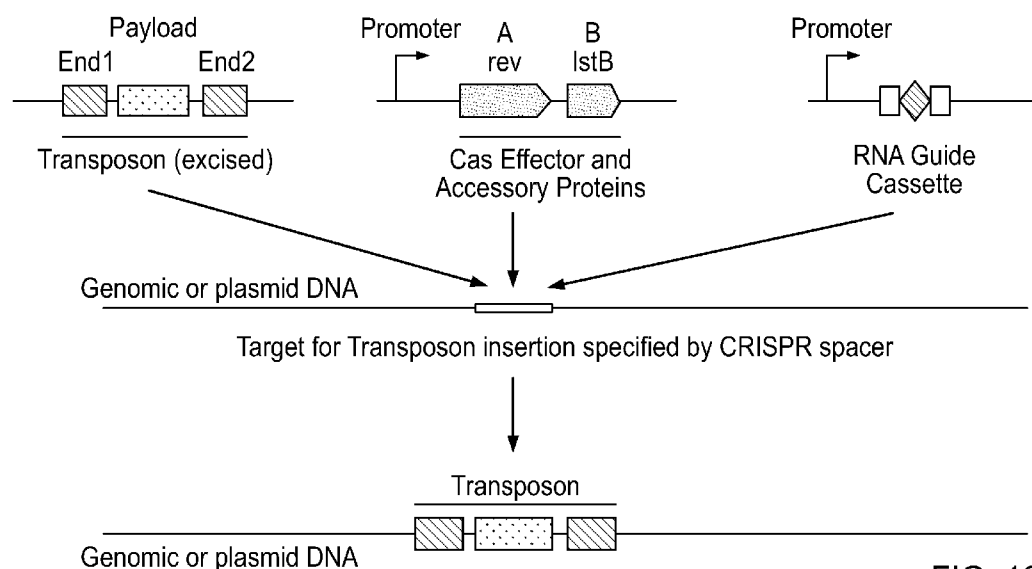


FIG. 10

(57) **Abstract:** The disclosure describes novel systems, methods, and compositions for the manipulation of nucleic acids in a targeted fashion. The disclosure describes non-naturally occurring, engineered CRISPR systems, components, and methods for targeted modification of DNA, RNA, and protein substrates. Each system includes one or more protein components and one or more nucleic acid components that together target DNA, RNA, or protein substrates.



Published:

- *with international search report (Art. 21(3))*
- *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

NOVEL CRISPR-ASSOCIATED TRANSPOSON SYSTEMS AND COMPONENTS**CROSS REFERENCE TO RELATED APPLICATIONS**

This application claims the benefit of priority of U.S. Application No. 62/580,880, filed on November 2, 2017, and U.S. Application No. 62/587,381, filed on November 16, 2017. The content of each of the foregoing applications is hereby incorporated by reference in its entirety.

FIELD OF THE INVENTION

The present disclosure relates to novel CRISPR systems and components, systems for detecting CRISPR systems, and methods and compositions for use of the CRISPR systems in, for example, nucleic acid targeting and manipulation.

BACKGROUND

Recent application of advances in genome sequencing technologies and analysis have yielded significant insights into the genetic underpinning of biological activities in many diverse areas of nature, ranging from prokaryotic biosynthetic pathways to human pathologies. To fully understand and evaluate the vast quantities of information produced by genetic sequencing technologies, equivalent increases in the scale, efficacy, and ease of technologies for genome and epigenome manipulation are needed. These novel genome and epigenome engineering technologies will accelerate the development of novel applications in numerous areas, including biotechnology, agriculture, and human therapeutics.

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) and the CRISPR-associated (Cas) genes, collectively known as the CRISPR-Cas or CRISPR/Cas systems, are currently understood to provide immunity to bacteria and archaea against phage infection. The CRISPR-Cas systems of prokaryotic adaptive immunity are an extremely diverse group of proteins effectors, non-coding elements, as well as loci architectures, some examples of which have been engineered and adapted to produce important biotechnologies.

The components of the system involved in host defense include one or more effector proteins capable of modifying DNA or RNA and an RNA guide element that is responsible to targeting these protein activities to a specific sequence on the phage DNA or RNA. The RNA guide is composed of a CRISPR RNA (crRNA) and may require an additional trans-activating

RNA (tracrRNA) to enable targeted nucleic acid manipulation by the effector protein(s). The crRNA consists of a direct repeat responsible for protein binding to the crRNA and a spacer sequence that is complementary to the desired nucleic acid target sequence. CRISPR systems can be reprogrammed to target alternative DNA or RNA targets by modifying the spacer sequence of the crRNA.

CRISPR-Cas systems can be broadly classified into two classes: Class 1 systems are composed of multiple effector proteins that together form a complex around a crRNA, and Class 2 systems consist of a single effector protein that complexes with the crRNA to target DNA or RNA substrates. The single-subunit effector composition of the Class 2 systems provides a simpler component set for engineering and application translation, and have thus far been an important source of programmable effectors. Thus, the discovery, engineering, and optimization of novel Class 2 systems may lead to widespread and powerful programmable technologies for genome engineering and beyond.

Citation or identification of any document in this application is not an admission that such document is available as prior art to the present invention.

SUMMARY

The present disclosure provides methods for computational identification of new CRISPR-Cas systems from genomic databases, together with the development of the natural loci into an engineered system, and experimental validation and application translation.

In one aspect, provided herein is an engineered, non-naturally occurring Clustered Interspaced Short Palindromic Repeat (CRISPR) - Cas system of CLUST.009467 including a Guide consisting of a direct repeat sequence and a spacer sequence capable of hybridizing to a target nucleic acid, wherein the Guide comprises a CRISPR RNA or DNA, and any one of the following: a CRISPR-associated protein containing both an HTH domain and an rve integrase domain capable of binding to the Guide, either as a monomer or multimer, and of targeting the target nucleic acid sequence complementary to the Guide spacer; an IstB domain-containing protein; and a payload nucleic acid flanked by transposon end sequences.

In some embodiments, the target nucleic acid is a DNA or an RNA. In some embodiments, the target nucleic acid is double-stranded DNA.

In some embodiments, targeting of the target nucleic acid by the protein and the Guide results in a modification in the target nucleic acid. In some embodiments, the modification in the target nucleic acid is a double-stranded cleavage event. In some embodiments, the modification in the target nucleic acid is a single-stranded cleavage event.

5 In some embodiments, the system described herein further comprises a donor template nucleic acid. In some embodiments, the donor template nucleic acid is a DNA. In some embodiments, the target nucleic acid is a double-stranded DNA and the targeting of the double-stranded DNA results in scarless DNA insertion.

10 In some embodiments, the modification results in cell toxicity. In some embodiments, the system described herein is within a cell. In some embodiments, the cell comprises a eukaryotic cell. In some embodiments, the cell comprises a prokaryotic cell.

In another aspect, provided herein are methods of targeting and editing a target nucleic acid, the method comprising contacting the target nucleic acid with a system described herein.

15 In another aspect, the disclosure provides methods of targeting the insertion of a payload nucleic acid at a site of a target nucleic acid, the method including contacting the target nucleic acid with a system described herein.

In another aspect, the disclosure provides methods of targeting the excision of a payload nucleic acid from a site at a target nucleic acid, the method including contacting the target nucleic acid with a system described herein.

20 In some embodiments, the CRISPR-associated protein comprises an amino acid sequence having at least 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% similarity to an amino acid sequence provided in any one of Tables 2 – 3.

25 In some embodiments, the Guide comprises a nucleic acid sequence having at least 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% similarity to a nucleic acid sequence provided in Table 4.

In some embodiments, the payload nucleic acid is flanked by transposon end sequences comprising a nucleic acid sequence having at least 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% similarity to a nucleic acid sequence contained in Table 5.

30 In certain embodiments, the CRISPR-associated protein can include at least one nuclear localization signal. In some embodiments, the CRISPR-associated protein includes at least one nuclear export signal.

In some embodiments, at least one component of the system is encoded by a codon-optimized nucleic acid for expression in a cell. In some embodiments, the codon-optimized nucleic acid is present within at least one vector. In some embodiments, the at least one vector
5 comprise one or more regulatory elements operably linked to a nucleic acid encoding the component of the system.

In certain embodiments, the one or more regulatory elements comprise at least one promoter. In some embodiments, the at least one promoter comprises an inducible promoter or a constitutive promoter.

10 In some embodiments, the at least one vector includes a plurality of vectors. In some embodiments, the at least one vector comprises a viral vector. In some embodiments, the viral vector is selected from the group consisting of a retroviral vector, a lentiviral vector, an adenoviral vector, an adeno-associated vector, and a herpes simplex vector.

In some embodiments, the system is present in a delivery system. In some embodiments,
15 the delivery system comprises a delivery vehicle selected from the group consisting of a liposome, an exosome, a microvesicle, and a gene-gun.

In another aspect, provided herein is a cell including one or more systems described herein. In some embodiments, the cell is a eukaryotic cell, e.g., a mammalian cell or a plant cell. In some embodiments, the mammalian cell is a human cell. In some embodiments, the cell is a
20 prokaryotic cell.

The term “CRISPR-associated transposon” as used herein refers to a mobile genetic element having terminal transposon ends on both sides, which is acted upon by a CRISPR system described herein. In some embodiments, the CRISPR-associated transposon includes a gene encoding a CRISPR-associated transposase that is capable of facilitating the mobility (e.g.,
25 excision or deletion) of the CRISPR-associated transposon from a first site in a nucleic acid to a second site in a nucleic acid.

The term “CRISPR-associated transposase” as used herein refers to a protein including one or more transposase domains that is encoded by a gene that in nature is present in a CRISPR-associated transposon. In some embodiments, the CRISPR-associated transposase is capable of
30 facilitating the mobility of a CRISPR-associated transposon from a first site in a nucleic acid to a second site in a nucleic acid. In some embodiments, the CRISPR-associated transposase has

integration activity. In some embodiments, the CRISPR-associated transposase has excision activity. In some embodiments, the CRISPR-associated transposase specifically targets a CRISPR-associated transposon for mobilization via an RNA guide.

The term “Guide” for a CRISPR-associated transposase system refers to either an RNA or DNA sequence that includes one or more direct repeat and spacer sequences, and that is capable of hybridizing to a target nucleic acid and to the proteins and/or nucleic acid of the CRISPR-associated transposon complex.

The term “cleavage event,” as used herein, refers to a DNA break in a target nucleic acid created by a nuclease of a CRISPR system described herein. In some embodiments, the cleavage event is a double-stranded DNA break. In some embodiments, the cleavage event is a single-stranded DNA break.

The term “CRISPR system” as used herein refers to nucleic acids and/or proteins involved in the expression of, or directing the activity of, CRISPR-effectors, including sequences encoding CRISPR effectors, RNA guides, and other sequences and transcripts from a CRISPR locus.

The term “CRISPR array” as used herein refers to the nucleic acid (e.g., DNA) segment that includes CRISPR repeats and spacers, starting with the first nucleotide of the first CRISPR repeat and ending with the last nucleotide of the last (terminal) CRISPR repeat. Typically, each spacer in a CRISPR array is located between two repeats. The term “CRISPR repeat,” or “CRISPR direct repeat,” or “direct repeat,” as used herein, refers to multiple short direct repeating sequences, which show very little or no sequence variation within a CRISPR array.

The term “CRISPR RNA” or “crRNA” as used herein refers to an RNA molecule comprising a guide sequence used by a CRISPR effector to specifically target a nucleic acid sequence. Typically, crRNAs contains a sequence that mediates target recognition and a sequence that forms a duplex with a tracrRNA. The crRNA: tracrRNA duplex binds to a CRISPR effector. The term “donor template nucleic acid,” as used herein refers to a nucleic acid molecule that can be used by one or more cellular proteins to alter the structure of a target nucleic acid after a CRISPR enzyme described herein has altered a target nucleic acid. In some embodiments, the donor template nucleic acid is a double-stranded nucleic acid. In some embodiments, the donor template nucleic acid is a single-stranded nucleic acid. In some embodiments, the donor template nucleic acid is linear. In some embodiments, the donor

template nucleic acid is circular (e.g., a plasmid). In some embodiments, the donor template nucleic acid is an exogenous nucleic acid molecule. In some embodiments, the donor template nucleic acid is an endogenous nucleic acid molecule (e.g., a chromosome).

The term “CRISPR effector,” “effector,” “CRISPR-associated protein,” or “CRISPR enzyme” as used herein refers to a protein that carries out an enzymatic activity or that binds to a target site on a nucleic acid specified by an RNA guide. In some embodiments, a CRISPR effector has endonuclease activity, nickase activity, exonuclease activity, transposase activity, and/or excision activity.

The term “guide RNA” or “gRNA” as used herein refers to an RNA molecule capable of directing a CRISPR effector having nuclease activity to target and cleave a specified target nucleic acid.

The term “RNA guide” as used herein refers to any RNA molecule that facilitates the targeting of a protein described herein to a target nucleic acid. Exemplary “RNA guides” include, but are not limited to, crRNAs, crRNAs fused to tracrRNAs, or crRNAs hybridized to tracrRNAs.

The term “mobile genetic element” as used herein refers to a nucleic acid capable of being specifically recognized and mobilized from a nucleic acid. In some embodiments, the mobile genetic element comprises nucleic acid sequences at flanking terminal ends that are specifically recognized by a CRISPR-associated transposase.

The term “origin of replication,” as used herein, refers to a nucleic acid sequence in a replicating nucleic acid molecule (e.g., a plasmid or a chromosome) at which replication is initiated.

As used herein, the term “target nucleic acid” refers to a specific nucleic acid sequence that is to be modified by a CRISPR system described herein. In some embodiments, the target nucleic acid comprises a gene. In some embodiments, the target nucleic acid comprises a non-coding region (e.g., a promoter). In some embodiments, the target nucleic acid is single-stranded. In some embodiments, the target nucleic acid is double-stranded.

The terms “trans-activating crRNA” or “tracrRNA” as used herein refer to an RNA including a sequence that forms a structure required for a CRISPR effector to bind to a specified target nucleic acid.

A “transcriptionally-active site” as used herein refers to a site in a nucleic acid sequence

comprising promoter regions at which transcription is initiated and actively occurring.

The term “collateral RNase activity,” as used herein in reference to a CRISPR enzyme, refers to non-specific RNase activity of a CRISPR enzyme after the enzyme has modified a specifically-targeted nucleic acid.

5 Unless otherwise defined, 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. Although methods and materials similar or equivalent to those described herein can be used in the practice or testing of the present invention, suitable methods and materials are described below. All publications, patent applications, patents, and other references mentioned
10 herein are incorporated by reference in their entirety. In case of conflict, the present specification, including definitions, will control. In addition, the materials, methods, and examples are illustrative only and not intended to be limiting.

Other features and advantages of the invention will be apparent from the following detailed description, and from the claims.

15

BRIEF FIGURE DESCRIPTION

FIGs. 1A-1D shows the conserved effectors, CRISPR array, and Tnp end elements by bacterial genome accession and species for representative CLUST.009467 loci.

20 **FIGs. 2A and 2B** shows a phylogenetic tree of CLUST.009467 effector A proteins.

FIGs. 3A and 3B shows an alignment of CLUST.009467 effector A proteins by genome accession and species, highlighting sequence location of conserved residues, with color denoting nucleotide polarity (Yellow: Non-polar side chain, Green: Polar side chain, Blue: Basic, Red: Acidic).

25 **FIG. 4** shows PFAM domains identified within CLUST.009467 effector A proteins.

FIGs. 5A and 5B shows a phylogenetic tree of CLUST.009467 effector B proteins.

FIGs. 6A and 6B shows an alignment of CLUST.009467 effector B proteins by genome accession and species, highlighting sequence location of conserved residues, with color denoting nucleotide polarity (Yellow: Non-polar side chain, Green: Polar side chain, Blue: Basic, Red: Acidic).
30

FIG. 7 shows PFAM domains identified within CLUST.009467 effector B proteins.

FIG. 8 shows a schematic of natural and engineered components for the CRISPR

transposition system of CLUST.009467.

FIG. 9 shows a schematic of transposon excision in the engineered system of CLUST.009467.

FIG. 10 shows a schematic of RNA guided transposon insertion in the engineered system of CLUST.009467.

FIG. 11 shows a schematic CRISPR negative selection screening

FIGs. 12 and 13 show depletion distributions for direct repeats and spacers targeting pACYC and *E. coli* essential genes. To quantify depletion, a fold-depletion ratio was calculated as $R_{treated} / R_{input}$ for each direct repeat and spacer. The normalized input read count is computed as:

$$R_{input} = \# \text{ reads containing DR+spacer} / \text{total reads}$$

without expressing the BDLR01000066 CLUST.009467 system and RNA guide. The treated read count is computed as

$$R_{treated} = (1 + \# \text{ reads containing DR+spacer}) / \text{total} \# \text{ reads}$$

with expression of the BDLR01000066 CLUST.009467 system and RNA guide. A strongly depleted target has a fold depletion greater than 3, which is marked by the red lines.

FIGs. 14A and 14B show the target site mapping of depleted RNA guides targeting pACYC (A) and *E. coli* essential genes (B) for the BDLR01000066 CLUST.009467 CRISPR-Cas system.

FIGs. 15A-C shows a weblogo of the sequences flanking the left (A) and right (B) sides of depleted targets for the BDLR01000066 CLUST.009467 CRISPR-Cas system.

FIGs. 16 and 17 show depletion distributions for direct repeats and spacers targeting pACYC and *E. coli* essential genes for the BDLR01000066 CLUST.009467 system with effector A knockout. A strongly depleted target has a fold depletion greater than 3, which is marked by the red lines.

FIGs. 18 and 19 show depletion distributions for direct repeats and spacers targeting pACYC and *E. coli* essential genes for the CP002736 CLUST.009467 system with effector A knockout. A strongly depleted target has a fold depletion greater than 3, which is marked by the red lines.

FIGs. 20A and 20B show the target site mapping of depleted RNA guides targeting pACYC (A) and *E. coli* essential genes (B) for the CP002736 CLUST.009467 CRISPR-Cas

system.

FIGs. 21A-C shows a weblogo of the sequences flanking the left (A) and right (B) sides of depleted targets for the CP002736 CLUST.009467 CRISPR-Cas system.

FIGs. 22 and 23 show depletion distributions for direct repeats and spacers targeting pACYC and *E. coli* essential genes for the CP002736 CLUST.009467 system with effector A knockout. A strongly depleted target has a fold depletion greater than 3, which is marked by the red lines.

DETAILED DESCRIPTION

The disclosure relates to the use of computational methods and algorithms to predict new CRISPR-Cas systems and identify their components.

In one embodiment, the disclosure includes new computational methods for identifying novel CRISPR loci by:

detecting all potential CRISPR arrays in prokaryotic data sources (contig, scaffold, or complete genome);

identifying all predicted protein coding genes in close proximity to a CRISPR array (e.g. 10 kb);

forming protein clusters (putative protein families) around identified genes, using, for example, mmseqs2;

selecting clusters of proteins of unknown function, and identifying homologs in the wider prokaryotic set of proteins using, e.g., BLAST or UBLAST;

identifying clusters of proteins with a large percentage of homologs co-occurring with CRISPR arrays; and

predicting the functional domains of the proteins in the identified cluster, e.g., by using hmmsearch on each member of the cluster individually, or by, for example, using a profile hidden Markov model (HMM) constructed from the multiple alignment.

External databases of functional domains include, for example, Pfam and Uniprot. Multiple alignment can be done using, e.g., mafft.

In another aspect, the disclosure relates to defining the minimal elements of novel CRISPR systems by:

identifying conserved elements (both coding genes and non-coding) in the loci

surrounding each cluster (in one aspect, this can be done manually by inspection on a case by case analysis);

identifying specific conserved non-coding elements that may be terminal repeats required for transposon activity;

5 identifying the RNA guide associated with each protein (identify conserved direct repeat structures and then attach a non-natural spacer sequence to the direct repeat. The effect of the non-natural spacer is to induce the effector to target a novel DNA or RNA substrate);

identifying the minimal RNA or DNA target (identify RNA and DNA targeting a by testing targeting activity of the effector(s) in combination with crRNAs containing multiple
10 different engineered spacer sequences targeting a high diversity of DNA and RNA substrates);

identifying the minimal system necessary to achieve activity; and

testing all conserved elements together and then systematically removing different proteins while preserving activity.

The broad natural diversity of CRISPR-Cas defense systems contains a wide range of
15 activity mechanisms and functional elements that can be harnessed for programmable biotechnologies. In a natural system, these mechanisms and parameters enable efficient defense against foreign DNA and viruses while providing self vs. non-self discrimination to avoid self-targeting. In an engineered system, the same mechanisms and parameters also provide a diverse toolbox of molecular technologies and define the boundaries of the targeting space. For instance,
20 systems Cas9 and Cas13a have canonical DNA and RNA endonuclease activity and their targeting spaces are defined by the protospacer adjacent motif (PAM) on targeted DNA and protospacer flanking sites (PFS) on targeted RNA, respectively.

The methods described herein can be used to discover additional mechanisms and parameters within single subunit Class 2 effector systems that can be more effectively harnessed
25 for programmable biotechnologies.

POOLED-SCREENING

To efficiently validate the activity of the engineered novel CRISPR-Cas systems and simultaneously evaluate in an unbiased manner different activity mechanisms and functional
30 parameters, we used a new pooled-screening approach in *E. coli*. First, from the computational identification of the conserved protein and noncoding elements of the novel CRISPR-Cas

system, DNA synthesis and molecular cloning was used to assemble the separate components into a single artificial expression vector, which in one embodiment is based on a pET-28a+ backbone. In a second embodiment, the effectors and noncoding elements are transcribed on a single mRNA transcript, and different ribosomal binding sites are used to translate individual effectors.

Second, the natural crRNA and targeting spacers were replaced with a library of unprocessed crRNAs containing non-natural spacers targeting the essential genes of the host *E. coli*, or a second plasmid bearing antibiotic resistance, pACYC184. This crRNA library was cloned into the vector backbone containing the protein effectors and noncoding elements (e.g. pET-28a+), and then subsequently transformed the library into *E. coli* along with the pACYC184 plasmid target. Consequently, each resulting *E. coli* cell contains no more than one targeting spacer.

Third, the *E. coli* were grown under antibiotic selection. In one embodiment, triple antibiotic selection is used: kanamycin for ensuring successful transformation of the pET-28a+ vector containing the engineered CRISPR-Cas effector system, and chloramphenicol and tetracycline for ensuring successful co-transformation of the pACYC184 target vector. Since pACYC184 normally confers resistance to chloramphenicol and tetracycline, under antibiotic selection, positive activity of the novel CRISPR-Cas system targeting the plasmid will eliminate cells that actively express the effectors, noncoding elements, and specific active elements of the crRNA library. Examining the population of surviving cells at a later time point compared to an earlier time point results in a depleted signal compared to the inactive crRNAs.

Since the pACYC184 plasmid contains a diverse set of features and sequences that may affect the activity of a CRISPR-Cas system, mapping the active crRNAs from the pooled screen onto pACYC184 provides patterns of activity that can be suggestive of different activity mechanisms and functional parameters in a broad, hypothesis-agnostic manner. In this way, the features required for reconstituting the novel CRISPR-Cas system in a heterologous prokaryotic species can be more comprehensively tested and studied.

The key advantages of the *in vivo* pooled-screen described herein include:

(1) Versatility - Plasmid design allows multiple effectors and/or noncoding elements to be expressed; library cloning strategy enables both transcriptional directions of the computationally predicted crRNA to be expressed;

(2) Comprehensive tests of activity mechanisms & functional parameters - Evaluates diverse interference mechanisms, including DNA or RNA cleavage; DNA excision and/or insertion via transposition; examines co-occurrence of features such as transcription, plasmid DNA replication; and flanking sequences for crRNA library can be used to reliably determine PAMs with complexity equivalence of 4N's;

(3) Sensitivity -- by using as targets the low copy number of pACYC184 and the single copy of the *E. coli* genome, this screen design enables high sensitivity for CRISPR-Cas activity since even modest interference rates can result in loss of cell viability through loss of antibiotic resistance or essential gene targeting; and

(4) Efficiency - Optimized molecular biology steps to enable greater speed and throughput, as RNA-sequencing and protein expression samples can be directly harvested from the surviving cells in the screen.

The novel CRISPR-Cas families described herein were evaluated using this *in vivo* pooled-screen to evaluate their operational elements, mechanisms and parameters, as well as their ability to be active and reprogrammed in an engineered system outside of their natural cellular environment.

CRISPR Enzymes Associated with Mobile Genetic Elements

This disclosure provides mobile genetic elements (e.g., CRISPR-associated transposons) associated with CRISPR systems described herein. These mobile genetic elements can be genetically altered to delete and/or add one or more components (e.g., a gene encoding a therapeutic product), thereby resulting in mobile genetic elements that can be readily inserted into or removed from a target nucleic acid. In some embodiments, the CRISPR systems encoded by these mobile genetic elements include one or more effector proteins, referred to herein as "CRISPR-associated transposases," which facilitate the movement of the mobile genetic elements from a first site in a nucleic acid to a second site in a nucleic acid. As described in further detail below, in some embodiments, the activity (e.g., excision activity or integration activity) of the CRISPR-associated transposases can be directed to a particular site on a target nucleic acid using a RNA guide that is complementary to a nucleic acid sequence on the target nucleic acid. In some embodiments, the RNA guides are engineered to specifically target the insertion, excision, and/or mobilization of the mobile genetic element to any site in a target

nucleic acid of interest.

In some embodiments, the disclosure provides CRISPR-associated transposons that include one or more genes encoding a CRISPR-associated transposase and an RNA guide. In some embodiments, the CRISPR-associated transposons include a payload nucleic acid. In some
5 embodiments, the payload nucleic acid includes a gene of interest. In some embodiments, the gene of interest is operably linked to a promoter (e.g., a constitutive promoter or an inducible promoter). In some embodiments, the gene of interest encodes a therapeutic protein.

This disclosure further provides a class of CRISPR effectors, referred to herein as CRISPR-associated transposases, capable of facilitating the movement of a mobile genetic
10 element (e.g., a CRISPR-associated transposon), wherein the targeting of the mobile genetic element is facilitated by RNA guides. Mobile genetic elements in CLUST.004377, CLUST.009467, and CLUST.009925 comprise CRISPR-associated transposons, including genes encoding CRISPR-associated transposases having a rve integrase domain (also referred to herein as rve integrase domain-containing effectors).

RNA-guided DNA Insertion

In some embodiments, a CRISPR system described herein mediates the insertion of a nucleotide payload into a target nucleic acid sequence. This activity is facilitated by one or more CRISPR-associated transposases present in the CRISPR system. In some embodiments, the CRISPR-associated transposase comprises one or more of a rve integrase domain a TniQ
20 domain, a TniB domain, or a TnpB domain. In some embodiments, the site of the target nucleic acid where the nucleotide payload is to be inserted is specified by a guide RNA.

RNA-guided DNA Excision

In some embodiments, a CRISPR system described herein mediates the excision of a nucleotide payload from a target nucleic acid sequence. This activity is facilitated by one or
25 more CRISPR-associated transposases present in the CRISPR system. In some embodiments, the CRISPR-associated transposase comprises one or more of a rve domain a TniQ domain, a TniB domain, or a TnpB domain. In some embodiments, the site of the target nucleic acid where the nucleotide payload is to be excised is specified by a guide RNA.

RNA-guided DNA Mobilization

In some embodiments, a CRISPR system described herein mediates the excision of a nucleotide payload from a first site in a target nucleic acid, mobilization of the nucleotide payload, and insertion of the nucleotide payload at a second site in a target nucleic acid. This activity is facilitated by one or more CRISPR-associated transposases present in the CRISPR system. In some embodiments, the CRISPR-associated transposase comprises one or more of a rve domain, a TniQ domain, a TniB domain, or a TnpB domain. In some embodiments, the first site of the target nucleic acid where the nucleotide payload is to be excised is specified by a guide RNA. In some embodiments, the second site in the target nucleic acid where the nucleotide payload is to be inserted is specified by a guide RNA.

Transposase Activity

In some embodiments, a CRISPR effector described herein (e.g., a CRISPR-associated transposase) comprises transposase activity. DNA transposition is one of the mechanisms by which genome rearrangement and horizontal gene transfer occurs in prokaryotic and eukaryotic cells. Transposons are DNA sequences that are capable of being moved within a genome. During DNA transposition a transposase recognizes transposon-specific sequences that flank an intervening DNA sequence. The transposase recognizes the transposon, excises the transposon from one location in a nucleic acid sequence, and inserts it (including the inverted repeat sequences with the intervening sequence) into another location. In some instances, the end sequences of the transposon comprise short inverted terminal repeats comprising duplications of a short segment of the sequence flanking the insertion sites that are characteristic for each transposon.

The mechanisms involved in transposition mediated by several transposon systems including *Tc1*, *Tol2*, *Minos*, *Himar1*, *Hsmar1*, *Mos1*, *Frog Prince*, *Piggyback*, and *Sleeping Beauty* have been characterized (see, e.g., Chitilian *et al.* (2014) *Stem Cells* 32: 204-15; Ivics *et al.* (1997) *Cell* 91: 501-10; Miskey *et al.* (2003) *Nucleic Acids Res.* 31: 6873-81; Urschitz *et al.* (2013) *Mob. Genet. Elements* 3: e25167; Jursch *et al.* (2013) *Mob. DNA* 4: 15; Pflieger *et al.* (2014) *J. Biol. Chem.* 289: 100-11); and Hou *et al.* (2015) *Cancer Biol. Ther.* 16(1): 8-16). For example, the Sleeping Beauty transposase (SBase) functions through a “cut-and-paste” process (see, e.g., Yant *et al.* (2004) *Mol. Cell. Biol.* 4: 9239-47). In some embodiments, a CRISPR-

associated transposase described herein comprise a transposase domain or a transposase-like domain. For example, in some embodiments, the CRISPR-associated transposase comprises a Mu-transposase domain or module. In some embodiments, the CRISPR-associated transposase comprises a TniQ transposase domain. In some embodiments, the CRISPR-associated
5 transposase comprises a TniB transposase domain. In some embodiments, the transposase domain is a OrfB_IS605 domain.

In some embodiments, the transposase domain of a CRISPR-associated transposase described herein is capable of one or more of the following activities: (a) excising a nucleic acid fragment from its situs on in a genome (*i.e.*, excision activity); (b) mediating the integration of a
10 nucleic acid fragment into a situs in a genome (*i.e.*, integration activity); and/or (c) specifically recognizing a transposon element (*e.g.*, a specific inverted repeat sequence). In some embodiments, a transposase domain of a CRISPR-associated transposase described herein is modified to eliminate one or more activities. In some embodiments, a transposase domain of a CRISPR-associated transposase described herein may be mutated such that the transposase
15 domain comprises excision activity, but does not comprise integration activity. In some embodiments, a transposase domain of a CRISPR-associated transposase described herein is mutated such that a transposase domain comprises integration activity, but does not comprise excision activity. In some embodiments, a transposase domain of a CRISPR-associated transposase described herein is mutated such that the transposase domain comprises integration
20 activity and excision activity, but lacks the ability to recognize a specific nucleic acid sequence. In some embodiments, a transposase domain of a CRISPR-associated transposase described herein is mutated such that the transposase domain comprises integration activity, but lacks excision activity and the ability to recognize a specific nucleic acid sequence. In some
25 embodiments, a transposase domain of a CRISPR-associated transposase described herein is mutated such that the transposase domain comprises excision activity, but lacks integration activity and the ability to recognize a specific nucleic acid sequence.

In some embodiments, the CRISPR-associated transposases described herein can be used to facilitate the random insertion of a nucleic acid sequence into the genome of a cell. In some
30 embodiments, the CRISPR-associated transposases described herein may be used to facilitate the targeted insertion or excision of a nucleic acid sequence into or out of the genome of a cell. In some embodiments, the nucleic acid sequence that is inserted into the cell is flanked by inverted

repeat sequences that are specifically recognized by a transposase domain of the CRISPR-associated transposase. In some embodiments, the intervening nucleic acid sequence between these inverted repeat sequences is at least 10 bp, at least 20 bp, at least 30 bp, at least 40 bp, at least 50 bp, at least 60 bp, at least 70 bp, at least 80 bp, at least 90 bp, at least 100 bp, at least 150
5 bp, at least 200 bp, at least 250 bp, at least 300 bp, at least 400 bp, at least 500 bp, at least 600 bp, at least 700 bp, at least 800 bp, at least 900 bp, at least 1000 bp, at least 1,500 bp, at least 2,000 bp, at least 3,000 bp, at least 4,000 bp, at least 5,000 bp in length, at least 6,000 bp in length, at least 7,000 bp in length, at least 8,000 bp in length, at least 9,000 bp in length, or at least 10,000 bp in length,. In some embodiments, the intervening nucleic acid sequence between
10 these inverted repeat sequences is less than 10 bp, less than 20 bp, less than 30 bp, less than 40 bp, less than 50 bp, less than 60 bp, less than 70 bp, less than 80 bp, less than 90 bp, less than 100 bp, less than 150 bp, less than 200 bp, less than 250 bp, less than 300 bp, less than 400 bp, less than 500 bp, less than 600 bp, less than 700 bp, less than 800 bp, less than 900 bp, less than 1000 bp, less than 1,500 bp, less than 2,000 bp, less than 3,000 bp, less than 4,000 bp, less than
15 5,000 bp in length, or less than 10,000 bp in length.

Effectors Comprising Helix-Turn-Helix Domains

In some embodiments, effectors within the systems described herein may contain a helix-turn-helix (HTH) domain. HTH domains typically include or consist of two α -helices forming
20 an internal angle of approximately 120 degrees that are connected by a short strand or turn of amino acid residues, and are present in a multitude of prokaryotic and eukaryotic DNA-binding proteins, such as transcription factors (see, e.g., Aravind et al. (2005) FEMS Microbiol. Rev. 29(2): 231-62). Beyond transcriptional regulation, HTH domains are involved in a wide range of functions including DNA repair and replication, RNA metabolism and protein-protein
25 interactions. In some embodiments, an effector described herein comprises DNA-binding activity. In some embodiments, an effector described herein comprises RNA-binding activity. In some embodiments, an effector described herein comprises a HTH domain that mediates a protein-protein interaction.

CRISPR ENZYME MODIFICATIONS

Modulating insertion/excision/transposition activity of CRISPR Enzymes

The activity of CRISPR-associated transposases may be altered in ways to change the relative efficiencies of insertion and excision. Altering these activities enable greater control in the different modes of transposase directed genome editing; for instance, CRISPR-associated transposases with predominantly insertion activity can be used for locus-specific insertion of transgenes to enable applications including, but not limited to, therapeutics (e.g., fetal hemoglobin expression for treatment of thalassemia), genetic engineering (e.g., trait stacking in plant genome engineering), or engineered cells (e.g., introducing control circuits or custom chimeric antigen receptors for CAR-T cell engineering). Alternatively, CRISPR-associated transposases with predominantly excision activity can also be used in applications, such as restoring the genome of engineered cells through excising inserted transcription factors. These CRISPR-associated transposases can be modified to have diminished excision or insertion activity, e.g., excision or insertion inactivation of at least 40%, at least 50%, at least 60%, at least 70%, at least 80%, at least 90%, at least 95%, at least 97%, or 100% as compared with the wild type CRISPR-associated transposases. Modulation of the insertion or excision activity can be done with several methods known in the art, such as introducing mutations into the catalytic core of the transposase. In some embodiments, catalytic residues for the transposase activities are identified, and these amino acid residues can be substituted by different amino acid residues (e.g., glycine or alanine) to diminish the transposase activity. An example in which modifications to the catalytic core of a transposase yielded differential effects on excision versus insertion is with the generation of an excision competent/integration defective variant of the *piggyBac* transposase (Exc⁺/Int⁻ PB) Li, Xianghong et al. “*piggyBac* transposase tools for genome engineering,” Proc. Nat’l. Acad. Sci., 1073.10 (2013): 2279-2287. In other embodiments, the variants to modulate excision/insertion activity can be combined with other CRISPR-associated transposase modifications, such as mutations that relax or make more stringent the targeting space or PAM constraints. Furthermore, these mutations are not restricted to the protein effectors of the transposon, but may be found on the noncoding elements such as the transposase ends. Together, these may yield a set of mutations that enable the tuning of the enzymatic activities of the CRISPR-associated transposase.

Generation of Fusion Proteins

Additionally, CRISPR-associated transposases, whether in its native functional form or with mutations to modulation its activity, can provide a foundation from which fusion proteins with additional functional proteins can be created. The inactivated CRISPR enzymes can
5 comprise or be associated with one or more functional domains (e.g., via fusion protein, linker peptides, “GS” linkers, etc.). These functional domains can have various activities, e.g., methylase activity, demethylase activity, transcription activation activity, transcription repression activity, transcription release factor activity, histone modification activity, RNA cleavage activity, DNA cleavage activity, nucleic acid binding. In some embodiments, the functional
10 domains are Krüppel associated box (KRAB), VP64, VP16, Fok1, P65, HSF1, MyoD1, and biotin-APEX.

The positioning of the one or more functional domains on the CRISPR transposase is one that allows for correct spatial orientation for the functional domain to affect the target with the attributed functional effect. For example, if the functional domain is a transcription activator
15 (e.g., VP16, VP64, or p65), the transcription activator is placed in a spatial orientation that allows it to affect the transcription of the target. Likewise, a transcription repressor is positioned to affect the transcription of the target, and a nuclease (e.g., Fok1) is positioned to cleave or partially cleave the target. In some embodiments, the functional domain is positioned at the N-terminus of the CRISPR enzyme. In some embodiments, the functional domain is positioned at
20 the C-terminus of the CRISPR enzyme. In some embodiments, the CRISPR enzyme is modified to comprise a first functional domain at the N-terminus and a second functional domain at the C-terminus.

The addition of functional domains to the CRISPR-associated transposase or onto other effector proteins in the complex may provide an ability for the transposase system to modify the
25 the physical DNA (e.g., methylation, etc.) or its regulation (e.g., transcriptional or repression) *in situ*.

Split Enzymes

The present disclosure also provides a split version of the CRISPR enzymes described herein. The split version of the CRISPR enzymes may be advantageous for delivery. In some
30 embodiments, the CRISPR enzymes are split to two parts of the enzymes, which together substantially comprises a functioning CRISPR enzyme.

The split can be done in a way that the catalytic domain(s) are unaffected. The CRISPR enzymes may function as a nuclease or may be inactivated enzymes, which are essentially RNA-binding proteins with very little or no catalytic activity (e.g., due to mutation(s) in its catalytic domains).

5 In some embodiments, the nuclease lobe and α -helical lobe are expressed as separate polypeptides. Although the lobes do not interact on their own, the guide RNA recruits them into a ternary complex that recapitulates the activity of full-length CRISPR enzymes and catalyzes site-specific DNA cleavage. The use of a modified guide RNA abrogates split-enzyme activity by preventing dimerization, allowing for the development of an inducible dimerization system.

10 The split enzyme is described, e.g., in Wright, Addison V., et al. "Rational design of a split-Cas9 enzyme complex," Proc. Nat'l. Acad. Sci., 112.10 (2015): 2984-2989, which is incorporated herein by reference in its entirety.

In some embodiments, the split enzyme can be fused to a dimerization partner, e.g., by employing rapamycin sensitive dimerization domains. This allows the generation of a chemically

15 inducible CRISPR enzyme for temporal control of CRISPR enzyme activity. The CRISPR enzymes can thus be rendered chemically inducible by being split into two fragments and rapamycin-sensitive dimerization domains can be used for controlled reassembly of the CRISPR enzymes.

The split point is typically designed *in silico* and cloned into the constructs. During this

20 process, mutations can be introduced to the split enzyme and non-functional domains can be removed. In some embodiments, the two parts or fragments of the split CRISPR enzyme (i.e., the N-terminal and C-terminal fragments), can form a full CRISPR enzyme, comprising, e.g., at least 70%, at least 80%, at least 90%, at least 95%, or at least 99% of the sequence of the wild-type CRISPR enzyme.

25 ***Self-Activating or Inactivating Enzymes***

The CRISPR enzymes described herein can be designed to be self-activating or self-inactivating. In some embodiments, the CRISPR enzymes are self-inactivating. For example, the target sequence can be introduced into the CRISPR enzyme coding constructs. Thus, the CRISPR enzymes can modify the target sequence, as well as the construct encoding the enzyme

30 thereby self-inactivating their expression. Methods of constructing a self-inactivating CRISPR system is described, e.g., in Epstein, Benjamin E., and David V. Schaffer. "Engineering a Self-

Inactivating CRISPR System for AAV Vectors,” Mol. Ther., 24 (2016): S50, which is incorporated herein by reference in its entirety.

In some other embodiments, an additional guide RNA, expressed under the control of a weak promoter (e.g., 7SK promoter), can target the nucleic acid sequence encoding the CRISPR enzyme to prevent and/or block its expression (e.g., by preventing the transcription and/or translation of the nucleic acid). The transfection of cells with vectors expressing the CRISPR enzyme, guide RNAs, and guide RNAs that target the nucleic acid encoding the CRISPR enzyme can lead to efficient disruption of the nucleic acid encoding the CRISPR enzyme and decrease the levels of CRISPR enzyme, thereby limiting the genome editing activity.

In some embodiments, the genome editing activity of the CRISPR enzymes can be modulated through endogenous RNA signatures (e.g., miRNA) in mammalian cells. The CRISPR enzyme switch can be made by using a miRNA-complementary sequence in the 5'-UTR of mRNA encoding the CRISPR enzyme. The switches selectively and efficiently respond to miRNA in the target cells. Thus, the switches can differentially control the genome editing by sensing endogenous miRNA activities within a heterogeneous cell population. Therefore, the switch systems can provide a framework for cell-type selective genome editing and cell engineering based on intracellular miRNA information (Hirosawa, Moe et al. “Cell-type-specific genome editing with a microRNA-responsive CRISPR–Cas9 switch,” Nucl. Acids Res., 2017 Jul 27; 45(13): e118).

Inducible CRISPR Enzymes

The CRISPR enzymes can be inducible, e.g., light inducible or chemically inducible. This mechanism allows for activation of the functional domain in the CRISPR enzymes. Light inducibility can be achieved by various methods known in the art, e.g., by designing a fusion complex wherein CRY2PHR/CIBN pairing is used in split CRISPR Enzymes (see, e.g., Konermann et al. “Optical control of mammalian endogenous transcription and epigenetic states,” *Nature*, 500.7463 (2013): 472). Chemical inducibility can be achieved, e.g., by designing a fusion complex wherein FKBP/FRB (FK506 binding protein / FKBP rapamycin binding domain) pairing is used in split CRISPR Enzymes. Rapamycin is required for forming the fusion complex, thereby activating the CRISPR enzymes (see, e.g., Zetsche, Volz, and Zhang, “A split-Cas9 architecture for inducible genome editing and transcription modulation,” *Nature Biotech.*, 33.2 (2015): 139-142).

Furthermore, expression of the CRISPR enzymes can be modulated by inducible promoters, e.g., tetracycline or doxycycline controlled transcriptional activation (Tet-On and Tet-Off expression system), hormone inducible gene expression system (e.g., an ecdysone inducible gene expression system), and an arabinose-inducible gene expression system. When delivered as RNA, expression of the RNA targeting effector protein can be modulated via a riboswitch, which can sense a small molecule like tetracycline (see, e.g., Goldfless, Stephen J. et al. "Direct and specific chemical control of eukaryotic translation with a synthetic RNA-protein interaction," *Nucl. Acids Res.*, 40.9 (2012): e64-e64).

Various embodiments of inducible CRISPR enzymes and inducible CRISPR systems are described, e.g., in US8871445, US20160208243, and WO2016205764, each of which is incorporated herein by reference in its entirety.

Functional Mutations

Various mutations or modifications can be introduced into CRISPR enzymes as described herein to improve specificity and/or robustness. In some embodiments, the amino acid residues that recognize the Protospacer Adjacent Motif (PAM) are identified. The CRISPR enzymes described herein can be modified further to recognize different PAMs, e.g., by substituting the amino acid residues that recognize PAM with other amino acid residues.

In some embodiments, at least one Nuclear Localization Signal (NLS) is attached to the nucleic acid sequences encoding the CRISPR enzyme. In some embodiments, at least one Nuclear Export Signal (NES) is attached to the nucleic acid sequences encoding the CRISPR enzyme. In a preferred embodiment a C-terminal and/or N-terminal NLS or NES is attached for optimal expression and nuclear targeting in eukaryotic cells, e.g., human cells.

In some embodiments, the CRISPR enzymes described herein are mutated at one or more amino acid residues to alter one or more functional activities. For example, in some embodiments, the CRISPR enzyme is mutated at one or more amino acid residues to alter its helicase activity. In some embodiments, the CRISPR enzyme is mutated at one or more amino acid residues to alter its nuclease activity (e.g., endonuclease activity or exonuclease activity). In some embodiments, the CRISPR enzyme is mutated at one or more amino acid residues to alter its ability to functionally associate with a guide RNA. In some embodiments, the CRISPR enzyme is mutated at one or more amino acid residues to alter its ability to functionally associate with a target nucleic acid.

In some embodiments, the CRISPR enzymes described herein are capable of modifying a target nucleic acid molecule. In some embodiments, the CRISPR enzyme modifies both strands of the target nucleic acid molecule. However, in some embodiments, the CRISPR enzyme is mutated at one or more amino acid residues to alter its nucleic acid manipulation activity. For example, in some embodiments, the CRISPR enzyme may comprise one or more mutations which render the enzyme incapable of cleaving a target nucleic acid or inserting/excising a target sequence.

In some embodiments, a CRISPR enzyme described herein may be engineered to comprise a deletion in one or more amino acid residues to reduce the size of the enzyme while retaining one or more desired functional activities (e.g., nuclease activity and the ability to functionally interact with a guide RNA). The truncated CRISPR enzyme may be advantageously used in combination with delivery systems having load limitations.

In one aspect, the present disclosure provides nucleotide sequences that are at least 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% identical to the sequences described herein. In another aspect, the present disclosure also provides amino acid sequences that are at least 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% identical to the sequences described herein.

In some embodiments, the nucleic acid sequences have at least a portion (e.g., at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20, 30, 40, 50, 60, 70, 80, 90, or 100 nucleotides, e.g., contiguous or non-contiguous nucleotides) that are the same as the sequences described herein. In some embodiments, the nucleic acid sequences have at least a portion (e.g., at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20, 30, 40, 50, 60, 70, 80, 90, or 100 nucleotides, e.g., contiguous or non-contiguous nucleotides) that is different from the sequences described herein.

In some embodiments, the amino acid sequences have at least a portion (e.g., at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20, 30, 40, 50, 60, 70, 80, 90, or 100 amino acid residues, e.g., contiguous or non-contiguous amino acid residues) that is the same as the sequences described herein. In some embodiments, the amino acid sequences have at least a portion (e.g., at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20, 30, 40, 50, 60, 70, 80, 90, or 100 amino acid residues, e.g., contiguous or non-contiguous amino acid residues) that is

different from the sequences described herein.

To determine the percent identity of two amino acid sequences, or of two nucleic acid sequences, the sequences are aligned for optimal comparison purposes (e.g., gaps can be introduced in one or both of a first and a second amino acid or nucleic acid sequence for optimal alignment and non-homologous sequences can be disregarded for comparison purposes). In general, the length of a reference sequence aligned for comparison purposes should be at least 80% of the length of the reference sequence, and in some embodiments is at least 90%, 95%, or 100% of the length of the reference sequence. The amino acid residues or nucleotides at corresponding amino acid positions or nucleotide positions are then compared. When a position in the first sequence is occupied by the same amino acid residue or nucleotide as the corresponding position in the second sequence, then the molecules are identical at that position. The percent identity between the two sequences is a function of the number of identical positions shared by the sequences, taking into account the number of gaps, and the length of each gap, which need to be introduced for optimal alignment of the two sequences. For purposes of the present disclosure, the comparison of sequences and determination of percent identity between two sequences can be accomplished using a Blosum 62 scoring matrix with a gap penalty of 12, a gap extend penalty of 4, and a frameshift gap penalty of 5.

GUIDE RNA MODIFICATIONS

Spacer Lengths

The spacer length of guide RNAs can range from about 15 to 50 nucleotides. In some embodiments, the spacer length of a guide RNA is at least 16 nucleotides, at least 17 nucleotides, at least 18 nucleotides, at least 19 nucleotides, at least 20 nucleotides, at least 21 nucleotides, or at least 22 nucleotides. In some embodiments, the spacer length is from 15 to 17 nucleotides, from 17 to 20 nucleotides, from 20 to 24 nucleotides (e.g., 20, 21, 22, 23, or 24 nucleotides), from 23 to 25 nucleotides (e.g., 23, 24, or 25 nucleotides), from 24 to 27 nucleotides, from 27 to 30 nucleotides, from 30 to 45 nucleotides (e.g., 30, 31, 32, 33, 34, 35, 40, or 45 nucleotides), from 30 or 35 to 40 nucleotides, from 41 to 45 nucleotides, from 45 to 50 nucleotides, or longer. In some embodiments, the direct repeat length of the guide RNA is at least 16 nucleotides, or is from 16 to 20 nucleotides (e.g., 16, 17, 18, 19, or 20 nucleotides). In some embodiments, the direct repeat length of the guide RNA is 19 nucleotides.

The guide RNA sequences can be modified in a manner that allows for formation of the CRISPR complex and successful binding to the target, while at the same time not allowing for successful effector activity (i.e., without excision activity / without insertion activity / without nuclease activity). These modified guide sequences are referred to as “dead guides” or “dead guide sequences.” These dead guides or dead guide sequences may be catalytically inactive or conformationally inactive with regard to nuclease activity. Dead guide sequences are typically shorter than respective guide sequences that result in active modification. In some embodiments, dead guides are 5%, 10%, 20%, 30%, 40%, or 50%, shorter than respective guide RNAs that have nuclease activity. Dead guide sequences of guide RNAs can be from 13 to 15 nucleotides in length (e.g., 13, 14, or 15 nucleotides in length), from 15 to 19 nucleotides in length, or from 17 to 18 nucleotides in length (e.g., 17 nucleotides in length).

Thus, in one aspect, the disclosure provides non-naturally occurring or engineered CRISPR systems including a functional CRISPR enzyme as described herein, and a guide RNA (gRNA) wherein the gRNA comprises a dead guide sequence whereby the gRNA is capable of hybridizing to a target sequence such that the CRISPR system is directed to a genomic locus of interest in a cell without detectable nucleic acid modification activity.

A detailed description of dead guides is described, e.g., in WO 2016094872, which is incorporated herein by reference in its entirety.

Inducible Guides

Guide RNAs can be generated as components of inducible systems. The inducible nature of the systems allows for spatiotemporal control of gene editing or gene expression. In some embodiments, the stimuli for the inducible systems include, e.g., electromagnetic radiation, sound energy, chemical energy, and/or thermal energy.

In some embodiments, the transcription of guide RNA can be modulated by inducible promoters, e.g., tetracycline or doxycycline controlled transcriptional activation (Tet-On and Tet-Off expression systems), hormone inducible gene expression systems (e.g., ecdysone inducible gene expression systems), and arabinose-inducible gene expression systems. Other examples of inducible systems include, e.g., small molecule two-hybrid transcription activations systems (FKBP, ABA, etc.), light inducible systems (Phytochrome, LOV domains, or cryptochrome), or Light Inducible Transcriptional Effector (LITE). These inducible systems are described, e.g., in WO 2016205764 and US 8795965, both of which are incorporated herein by reference in the

entirety.

Chemical Modifications

Chemical modifications can be applied to the guide RNA's phosphate backbone, sugar, and/or base. Backbone modifications such as phosphorothioates modify the charge on the phosphate backbone and aid in the delivery and nuclease resistance of the oligonucleotide (see, 5 e.g., Eckstein, "Phosphorothioates, essential components of therapeutic oligonucleotides," *Nucl. Acid Ther.*, 24 (2014), pp. 374-387); modifications of sugars, such as 2'-O-methyl (2'-OMe), 2'-F, and locked nucleic acid (LNA), enhance both base pairing and nuclease resistance (see, e.g., Allerson et al. "Fully 2'-modified oligonucleotide duplexes with improved *in vitro* potency and 10 stability compared to unmodified small interfering RNA," *J. Med. Chem.*, 48.4 (2005): 901-904). Chemically modified bases such as 2-thiouridine or N6-methyladenosine, among others, can allow for either stronger or weaker base pairing (see, e.g., Bramsen et al., "Development of therapeutic-grade small interfering RNAs by chemical engineering," *Front. Genet.*, 2012 Aug 20; 3:154). Additionally, RNA is amenable to both 5' and 3' end conjugations with a variety of 15 functional moieties including fluorescent dyes, polyethylene glycol, or proteins.

A wide variety of modifications can be applied to chemically synthesized guide RNA molecules. For example, modifying an oligonucleotide with a 2'-OMe to improve nuclease resistance can change the binding energy of Watson-Crick base pairing. Furthermore, a 2'-OMe modification can affect how the oligonucleotide interacts with transfection reagents, proteins or 20 any other molecules in the cell. The effects of these modifications can be determined by empirical testing.

In some embodiments, the guide RNA includes one or more phosphorothioate modifications. In some embodiments, the guide RNA includes one or more locked nucleic acids for the purpose of enhancing base pairing and/or increasing nuclease resistance.

25 A summary of these chemical modifications can be found, e.g., in Kelley et al., "Versatility of chemically synthesized guide RNAs for CRISPR-Cas9 genome editing," *J. Biotechnol.* 2016 Sep 10; 233:74-83; WO 2016205764; and US 8795965 B2; each which is incorporated by reference in its entirety.

Sequence Modifications

30 The sequences and the lengths of the guide RNAs, tracrRNAs, and crRNAs described

herein can be optimized. In some embodiments, the optimized length of guide RNA can be determined by identifying the processed form of tracrRNA and/or crRNA, or by empirical length studies for guide RNAs, tracrRNAs, crRNAs, and the tracrRNA tetraloops.

The guide RNAs can also include one or more aptamer sequences. Aptamers are oligonucleotide or peptide molecules that can bind to a specific target molecule. The aptamers can be specific to gene effectors, gene activators, or gene repressors. In some embodiments, the aptamers can be specific to a protein, which in turn is specific to and recruits / binds to specific gene effectors, gene activators, or gene repressors. The effectors, activators, or repressors can be present in the form of fusion proteins. In some embodiments, the guide RNA has two or more aptamer sequences that are specific to the same adaptor proteins. In some embodiments, the two or more aptamer sequences are specific to different adaptor proteins. The adaptor proteins can include, e.g., MS2, PP7, Q β , F2, GA, fr, JP501, M12, R17, BZ13, JP34, JP500, KU1, M11, MX1, TW18, VK, SP, FI, ID2, NL95, TW19, AP205, ϕ Cb5, ϕ Cb8r, ϕ Cb12r, ϕ Cb23r, 7s, and PRR1. Accordingly, in some embodiments, the aptamer is selected from binding proteins specifically binding any one of the adaptor proteins as described herein. In some embodiments, the aptamer sequence is a MS2 loop. A detailed description of aptamers can be found, e.g., in Nowak et al., "Guide RNA engineering for versatile Cas9 functionality," *Nucl. Acid. Res.*, 2016 Nov 16;44(20):9555-9564; and WO 2016205764, which are incorporated herein by reference in their entirety.

Guide: Target Sequence Matching Requirements

In classic CRISPR systems, the degree of complementarity between a guide sequence and its corresponding target sequence can be about 50%, 60%, 75%, 80%, 85%, 90%, 95%, 97.5%, 99%, or 100%. In some embodiments, the degree of complementarity is 100%. The guide RNAs can be about 5, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 35, 40, 45, 50, 75, or more nucleotides in length.

To reduce off-target interactions, e.g., to reduce the guide interacting with a target sequence having low complementarity, mutations can be introduced to the CRISPR systems so that the CRISPR systems can distinguish between target and off-target sequences that have greater than 80%, 85%, 90%, or 95% complementarity. In some embodiments, the degree of complementarity is from 80% to 95%, e.g., about 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, or 95% (for example, distinguishing between a target having 18

nucleotides from an off-target of 18 nucleotides having 1, 2, or 3 mismatches). Accordingly, in some embodiments, the degree of complementarity between a guide sequence and its corresponding target sequence is greater than 94.5%, 95%, 95.5%, 96%, 96.5%, 97%, 97.5%, 98%, 98.5%, 99%, 99.5%, or 99.9%. In some embodiments, the degree of complementarity is 100%.

It is known in the field that complete complementarity is not required, provided there is sufficient complementarity to be functional. For CRISPR nucleases, modulations of cleavage efficiency can be exploited by introduction of mismatches, e.g., one or more mismatches, such as 1 or 2 mismatches between spacer sequence and target sequence, including the position of the mismatch along the spacer/target. The more central (i.e., not at the 3' or 5' ends) a mismatch, e.g., a double mismatch, is located; the more cleavage efficiency is affected. Accordingly, by choosing mismatch positions along the spacer sequence, cleavage efficiency can be modulated. For example, if less than 100% cleavage of targets is desired (e.g., in a cell population), 1 or 2 mismatches between spacer and target sequence can be introduced in the spacer sequences.

METHODS OF USING CRISPR SYSTEMS

The CRISPR-associated transposon systems described herein have a wide variety of utilities including modifying (e.g., deleting, inserting, translocating, inactivating, or activating) a target polynucleotide in a multiplicity of cell types. The CRISPR transposon systems have a broad spectrum of applications in, e.g., tracking and labeling of nucleic acids, drug screening, and treating various genetic disorders.

High-Throughput Screening

The CRISPR systems described herein can be used for preparing next generation sequencing (NGS) libraries. For example, to create a cost-effective NGS library, the CRISPR systems can be used to disrupt the coding sequence of a target gene, and the CRISPR enzyme transfected clones can be screened simultaneously by next-generation sequencing (e.g., on the Illumina system). CRISPR-associated transposases may enable more efficient preparation due to the ability to directly insert barcodes and adaptor sequences on a transposase payload.

Engineered Microorganisms

Microorganisms (e.g., *E. coli*, yeast, and microalgae) are widely used for synthetic

biology. The development of synthetic biology has a wide utility, including various clinical applications. For example, the programmable CRISPR systems can be used to split proteins of toxic domains for targeted cell death, e.g., using cancer-linked RNA as target transcript. Further, pathways involving protein-protein interactions can be influenced in synthetic biological systems with e.g. fusion complexes with the appropriate effectors such as kinases or enzymes.

In some embodiments, guide RNA sequences that target phage sequences can be introduced into the microorganism. Thus, the disclosure also provides methods of vaccinating a microorganism (e.g., a production strain) against phage infection.

In some embodiments, the CRISPR systems provided herein can be used to engineer microorganisms, e.g., to improve yield or improve fermentation efficiency. For example, the CRISPR systems described herein can be used to engineer microorganisms, such as yeast, to generate biofuel or biopolymers from fermentable sugars, or to degrade plant-derived lignocellulose derived from agricultural waste as a source of fermentable sugars. More particularly, the methods described herein can be used to modify the expression of endogenous genes required for biofuel production and/or to modify endogenous genes, which may interfere with the biofuel synthesis. These methods of engineering microorganisms are described e.g., in Verwaal et al., “CRISPR/Cpf1 enables fast and simple genome editing of *Saccharomyces cerevisiae*,” *Yeast*, 2017 Sep 8. doi: 10.1002/yea.3278; and Hlavova et al., “Improving microalgae for biotechnology—from genetics to synthetic biology,” *Biotechnol. Adv.*, 2015 Nov 1; 33:1194-203, both of which are incorporated herein by reference in the entirety.

Application in Plants

The CRISPR systems described herein have a wide variety of utility in plants. In some embodiments, the CRISPR systems can be used to engineer genomes of plants (e.g., improving production, making products with desired post-translational modifications, or introducing genes for producing industrial products). In some embodiments, the CRISPR systems can be used to introduce a desired trait to a plant (e.g., with or without heritable modifications to the genome), or regulate expression of endogenous genes in plant cells or whole plants.

In some embodiments, the CRISPR systems can be used to identify, edit, and/or silence genes encoding specific proteins, e.g., allergenic proteins (e.g., allergenic proteins in peanuts, soybeans, lentils, peas, green beans, and mung beans). A detailed description regarding how to identify, edit, and/or silence genes encoding proteins is described, e.g., in Nicolaou et al.,

“Molecular diagnosis of peanut and legume allergy,” *Curr. Opin. Allergy Clin. Immunol.*, 2011 Jun; 11(3):222-8, and WO 2016205764 A1; both of which are incorporated herein by reference in the entirety.

Gene Drives

5 Gene drive is the phenomenon in which the inheritance of a particular gene or set of genes is favorably biased. The CRISPR systems described herein can be used to build gene drives. For example, the CRISPR systems can be designed to target and disrupt a particular allele of a gene, causing the cell to copy the second allele to fix the sequence. Because of the copying, the first allele will be converted to the second allele, increasing the chance of the second allele
10 being transmitted to the offspring. A detailed method regarding how to use the CRISPR systems described herein to build gene drives is described, e.g., in Hammond et al., “A CRISPR-Cas9 gene drive system targeting female reproduction in the malaria mosquito vector *Anopheles gambiae*,” *Nat. Biotechnol.*, 2016 Jan; 34(1):78-83, which is incorporated herein by reference in its entirety.

Pooled-Screening

15 As described herein, pooled CRISPR screening is a powerful tool for identifying genes involved in biological mechanisms such as cell proliferation, drug resistance, and viral infection. Cells are transduced in bulk with a library of guide RNA (gRNA)-encoding vectors described herein, and the distribution of gRNAs is measured before and after applying a selective
20 challenge. Pooled CRISPR screens work well for mechanisms that affect cell survival and proliferation, and they can be extended to measure the activity of individual genes (e.g., by using engineered reporter cell lines). CRISPR-associated transposases may enable more efficient disruption due to the potential for inserting larger sequences. Arrayed CRISPR screens, in which only one gene is targeted at a time, make it possible to use RNA-seq as the readout. In some
25 embodiments, the CRISPR systems as described herein can be used in single-cell CRISPR screens. A detailed description regarding pooled CRISPR screenings can be found, e.g., in Datlinger et al., “Pooled CRISPR screening with single-cell transcriptome read-out,” *Nat. Methods.*, 2017 Mar; 14(3):297-301, which is incorporated herein by reference in its entirety.

Saturation Mutagenesis (Bashing)

The CRISPR systems described herein can be used for *in situ* saturating mutagenesis. In some embodiments, a pooled guide RNA library can be used to perform *in situ* saturating mutagenesis for particular genes or regulatory elements. In other embodiments, a pooled library of DNA inserts containing a saturating mutagenesis library can be inserted by the CRISPR-associated transposase. Such methods can reveal critical minimal features and discrete vulnerabilities of these genes or regulatory elements (e.g., enhancers). These methods are described, e.g., in Canver et al., “BCL11A enhancer dissection by Cas9-mediated *in situ* saturating mutagenesis,” *Nature*, 2015 Nov 12; 527(7577):192-7, which is incorporated herein by reference in its entirety.

RNA-Related Applications

The CRISPR systems described herein can have various RNA-related applications, e.g., modulating gene expression, inhibiting RNA expression, screening RNA or RNA products, determining functions of lincRNA or non-coding RNA, inducing cell dormancy, inducing cell cycle arrest, reducing cell growth and/or cell proliferation, inducing cell anergy, inducing cell apoptosis, inducing cell necrosis, inducing cell death, and/or inducing programmed cell death. A detailed description of these applications can be found, e.g., in WO 2016205764 A1, which is incorporated herein by reference in its entirety.

Modulating Gene Expression

The CRISPR systems described herein can be used to modulate gene expression. The CRISPR systems can be used, together with suitable guide RNAs, to target gene expression, via control of RNA processing. The control of RNA processing can include, e.g., RNA processing reactions such as RNA splicing (e.g., alternative splicing), viral replication, and tRNA biosynthesis. The RNA targeting proteins in combination with suitable guide RNAs can also be used to control RNA activation (RNAa). RNA activation is a small RNA-guided and Argonaute (Ago)-dependent gene regulation phenomenon in which promoter-targeted short double-stranded RNAs (dsRNAs) induce target gene expression at the transcriptional/epigenetic level. RNAa leads to the promotion of gene expression, so control of gene expression may be achieved that way through disruption or reduction of RNAa. In some embodiments, the methods include the use of the RNA targeting CRISPR as substitutes for e.g., interfering ribonucleic acids (such as

siRNAs, shRNAs, or dsRNAs). The methods of modulating gene expression are described, e.g., in WO 2016205764, which is incorporated herein by reference in its entirety.

Controlling RNA Interference

Control over interfering RNAs or microRNAs (miRNA) can help reduce off-target effects by reducing the longevity of the interfering RNAs or miRNAs *in vivo* or *in vitro*. In some embodiments, the target RNAs can include interfering RNAs, i.e., RNAs involved in the RNA interference pathway, such as small hairpin RNAs (shRNAs), small interfering (siRNAs), etc. In some embodiments, the target RNAs include, e.g., miRNAs or double stranded RNAs (dsRNA).

In some embodiments, if the RNA targeting protein and suitable guide RNAs are selectively expressed (for example spatially or temporally under the control of a regulated promoter, for example a tissue- or cell cycle-specific promoter and/or enhancer), this can be used to protect the cells or systems (*in vivo* or *in vitro*) from RNA interference (RNAi) in those cells. This may be useful in neighboring tissues or cells where RNAi is not required or for the purposes of comparison of the cells or tissues where the effector proteins and suitable guide RNAs are and are not expressed (i.e., where the RNAi is not controlled and where it is, respectively). The RNA targeting proteins can be used to control or bind to molecules comprising or consisting of RNAs, such as ribozymes, ribosomes, or riboswitches. In some embodiments, the guide RNAs can recruit the RNA targeting proteins to these molecules so that the RNA targeting proteins are able to bind to them. These methods are described, e.g., in WO 2016205764 and WO 2017070605, both of which are incorporated herein by reference in the entirety.

Therapeutic Applications

The CRISPR-associated transposon systems described herein can have diverse therapeutic applications. Without wishing to be limiting, one framework to organize the range of therapeutic applications enabled by CRISPR-associated transposon systems is determining whether the therapeutic genetic modification is a correction of a native locus, or locus-agnostic gene augmentation.

For therapeutic correction of a native locus, the new CRISPR systems can be used to correct mutations responsible for monogenic diseases (e.g., Duchenne Muscular Dystrophy, Cystic Fibrosis, etc.), or introduce beneficial mutations (e.g., Pcsk9 for lowered cardiovascular

disease risk, BCL11a for increasing fetal hemoglobin expression in treating hemoglobinopathies, CCR5 for HIV resistance, etc.) Using CRISPR-associated transposons may have key advantages. The first advantage is the potential to use a single therapeutic construct to correct a diverse set of genetic mutations, whether in a single patient or across the patient population. This is due to the fact that transposition enables the replacement of a large gene fragment, rather than the short-range corrections enabled by homology directed repair following DNA cleavage or base editing. Second, using CRISPR-associated transposons may enable therapeutic modifications in a broad range of post-mitotic cells and tissues, given that the enzymatic mechanism of action of the transposon is anticipated to be independent from DNA repair mechanisms such as homologous recombination or homology directed repair.

For locus-agnostic gene augmentation, the new CRISPR systems can be used to introduce gene fragments that provide a therapeutic benefit. This includes gene therapies to replace missing or defective native enzymes, including, but not limited to, *RPE65* in Leber's Congenital Amaurosis, adenosine de-aminase in Severe Combined Immunodeficiency (SCID), and any number of defective enzymes causing diseases of inborn errors of metabolism. In addition to supplementing defective enzymes, the CRISPR-associated transposons may provide augment existing cellular properties, such as the introduction of custom chimeric antigen T-cell receptors for the production of cell therapies. The CRISPR-associated transposons may have the advantages of greater therapeutic durability when the transgene is incorporated genomically (vs. episomal expression in some recombinant viral vectors), and of greater control in transgene insertion, to ensure that the transposon is directed towards chromosomal locations that are at once "safe harbors" for genome editing but still transcriptionally active. This is differentiated from the possible deleterious effects of pseudo-random insertion of integrating viruses, as well as the promiscuous insertion using transposases such as Tn5, *piggyBac*, and *Sleeping Beauty*.

Altogether, a programmable transposon would enable a range of genome modifications that may prove highly valuable to therapeutic development, whether directly via therapeutic gene corrections, or indirectly by enabling the engineering of cells and disease models.

Delivery

Through this disclosure and the knowledge in the art, the CRISPR systems described

herein, or components thereof, nucleic acid molecules thereof, or nucleic acid molecules encoding or providing components thereof, can be delivered by various delivery systems such as vectors, e.g., plasmids, viral delivery vectors. The new CRISPR enzymes and/or any of the RNAs (e.g., guide RNAs) can be delivered using suitable vectors, e.g., plasmids or viral vectors, such as adeno-associated viruses (AAV), lentiviruses, adenoviruses, and other viral vectors, or combinations thereof. The proteins and one or more guide RNAs can be packaged into one or more vectors, e.g., plasmids or viral vectors.

In some embodiments, the vectors, e.g., plasmids or viral vectors, are delivered to the tissue of interest by, e.g., intramuscular injection, intravenous administration, transdermal administration, intranasal administration, oral administration, or mucosal administration. Such delivery may be either via a single dose, or multiple doses. One skilled in the art understands that the actual dosage to be delivered herein may vary greatly depending upon a variety of factors, such as the vector choices, the target cells, organisms, tissues, the general conditions of the subject to be treated, the degrees of transformation/modification sought, the administration routes, the administration modes, the types of transformation/modification sought, etc.

In certain embodiments, the delivery is via adenoviruses, which can be at a single dose containing at least 1×10^5 particles (also referred to as particle units, pu) of adenoviruses. In some embodiments, the dose preferably is at least about 1×10^6 particles, at least about 1×10^7 particles, at least about 1×10^8 particles, and at least about 1×10^9 particles of the adenoviruses. The delivery methods and the doses are described, e.g., in WO 2016205764 A1 and U.S. Patent No. 8,454,972 B2, both of which are incorporated herein by reference in the entirety.

In some embodiments, the delivery is via plasmids. The dosage can be a sufficient number of plasmids to elicit a response. In some cases, suitable quantities of plasmid DNA in plasmid compositions can be from about 0.1 to about 2 mg. Plasmids will generally include (i) a promoter; (ii) a sequence encoding a nucleic acid-targeting CRISPR enzymes, operably linked to the promoter; (iii) a selectable marker; (iv) an origin of replication; and (v) a transcription terminator downstream of and operably linked to (ii). The plasmids can also encode the RNA components of a CRISPR complex, but one or more of these may instead be encoded on different vectors. The frequency of administration is within the ambit of the medical or veterinary practitioner (e.g., physician, veterinarian), or a person skilled in the art.

In another embodiment, the delivery is via liposomes or lipofectin formulations and the

like, and can be prepared by methods known to those skilled in the art. Such methods are described, for example, in WO 2016205764 and U.S. Pat. Nos. 5,593,972; 5,589,466; and 5,580,859; each of which is incorporated herein by reference in its entirety.

In some embodiments, the delivery is via nanoparticles or exosomes. For example,
5 exosomes have been shown to be particularly useful in delivery RNA.

Further means of introducing one or more components of the new CRISPR systems to the cell is by using cell penetrating peptides (CPP). In some embodiments, a cell penetrating peptide is linked to the CRISPR enzymes. In some embodiments, the CRISPR enzymes and/or guide RNAs are coupled to one or more CPPs to effectively transport them inside cells (e.g., plant
10 protoplasts). In some embodiments, the CRISPR enzymes and/or guide RNA(s) are encoded by one or more circular or non-circular DNA molecules that are coupled to one or more CPPs for cell delivery.

CPPs are short peptides of fewer than 35 amino acids either derived from proteins or from chimeric sequences capable of transporting biomolecules across cell membrane in a
15 receptor independent manner. CPPs can be cationic peptides, peptides having hydrophobic sequences, amphipathic peptides, peptides having proline- rich and anti-microbial sequences, and chimeric or bipartite peptides. Examples of CPPs include, e.g., Tat (which is a nuclear transcriptional activator protein required for viral replication by HIV type I), penetratin, Kaposi fibroblast growth factor (FGF) signal peptide sequence, integrin $\beta 3$ signal peptide sequence,
20 polyarginine peptide Args sequence, Guanine rich-molecular transporters, and sweet arrow peptide. CPPs and methods of using them are described, e.g., in Hällbrink et al., "Prediction of cell-penetrating peptides," *Methods Mol. Biol.*, 2015;1324:39-58; Ramakrishna et al., "Gene disruption by cell-penetrating peptide-mediated delivery of Cas9 protein and guide RNA," *Genome Res.*, 2014 Jun;24(6):1020-7; and WO 2016205764 A1; each of which is incorporated
25 herein by reference in its entirety.

Various delivery methods for the CRISPR systems described herein are also described, e.g., in US 8795965, EP 3009511, WO 2016205764, and WO 2017070605; each of which is incorporated herein by reference in its entirety.

EXAMPLES

The invention is further described in the following examples, which do not limit the scope of the invention described in the claims.

5 Example 1 – Identification of Minimal Components for a CLUST.009467 CRISPR System (FIGs. 1 – 7)

This protein family describes a mobile genetic element associated with CRISPR systems found in organisms including but not limited to *Bifidobacterium* and *Clostridium* (FIG. 1). The naturally occurring loci containing this effector are depicted in FIG. 1, indicative that Effector A (~510 amino acids) has a high co-occurrence with the effector protein Effector B (~250 aa).
10 CLUST.009467 effectors include the exemplary effectors detailed in TABLES 1 – 5.

- An HMM profile constructed from the multiple sequence alignment of Effector A, and Pfam and Uniprot database searches revealed an rve domain indicative of integrase activity, as well HTH and DDE domains from the Tc3 family of transposons, indicative of nucleic acid binding and transposition activity (FIGs. 2-4).
15
- A domain search as described above for Effector B revealed an IstB domain originating from the IS21 family of transposons, as well as a Bac_DnaA domain indicative of DNA unwinding activity. Together these domains are indicative of accessory functions involved in the transposition process (FIGs. 5-7).
20

Table 1. Representative CLUST.009467 Effector Proteins

Species	Effector A Accession	Effector B Accession	# Spacers	Cas1	Cas2	Effector Size
<i>Bifidobacterium longum</i> subsp. <i>infantis</i> EK3 (NZ_JNWB01000026)	WP_032744684.1	WP_032744686.1	68	Y	N	492
<i>Geobacillus</i> sp. Y4.1MC1 (NC_014650)	WP_013401401.1	WP_013399917.1	2	N	N	505
<i>Geobacillus thermoglucosidasius</i> C56-YS93 (CP002835)	AEH46354.1	AEH46355.1	3	N	N	505
<i>Bifidobacterium longum</i> subsp. <i>longum</i> BBMN68 (CP002286)	ADQ03535.1	ADQ01872.1	17	N	N	500
<i>Aminomonas paucivorans</i> DSM 12260 (CM001022)	EFQ22705.1	EFQ22706.1	79	Y	Y	489
<i>Syntrophomonas wolfei</i> subsp. <i>wolfei</i> str. Goettingen G311 (NC_008346)	WP_011639494.1	WP_011639493.1	57	Y	Y	501
<i>Bifidobacterium breve</i> UCC2003 (CP000303)	WP_015439061.1	WP_015439060.1	46	Y	Y	499
<i>Desulfotomaculum nigrificans</i> CO-1-SRB (CP002736)	WP_013809817.1	WP_013809816.1	62	Y	Y	496
<i>Bifidobacterium longum</i> subsp. <i>longum</i> (NZ_MRBG01000001)	WP_007056033.1	WP_007056032.1	9	N	N	493
<i>Mesotoga infera</i> (CARH01000041)	CCU84826.1	CCU84827.1	9	N	N	486

Clostridium stercorarium subsp. thermolacticum DSM 2910 (CP014672)	ANW98740.1	ANW98739.1	153	Y	Y	507
Carboxydocella sp. JDF658 (NZ_BDLR01000066)	WP_079907503.1	N/A	25	Y	Y	503
Bifidobacterium breve NCFB 2258 (CP006714)	WP_025263135.1	WP_015439062.1	37	Y	Y	499
Geobacillus sp. Y4.1MC1 (CP002293)	WP_013400884.1	WP_013399917.1	29	N	N	505
Bifidobacterium adolescentis (LNKL01000001)	WP_011742685.1	WP_011742684.1	50	Y	Y	487
Polymorphum gilvum SL003B-26A1 (CP002568)	ADZ71261.1	ADZ71260.1	30	Y	Y	498
Thermovenabulum gondwanense (LOHZ01000022)	KYO67332.1	WP_068747786.1	227	N	N	511
Actinobacteria bacterium CG2_30_50_142 (MNXI01000142)	OIO83550.1	OIO83549.1	6	N	N	494
Desulfotomaculum kuznetsovii DSM 6115 (CP002770)	AEG16465.1	WP_013823974.1	58	Y	Y	512
Anoxybacillus geothermalis (NZ_JYCG01000314)	WP_044746322.1	WP_044746317.1	6	N	N	501
Anoxybacillus sp. B2M1 (NZ_CP015435)	WP_044746322.1	WP_044746317.1	21	N	N	501
Clostridium josui JCM 17888 (NZ_JAGE01000001)	WP_024833244.1	WP_024833243.1	28	Y	Y	503
Numidum massiliense (NZ_CTDZ01000009)	WP_074011080.1	N/A	71	Y	Y	500
Bifidobacterium aesculapii (NZ_BCFK01000046)	WP_082440433.1	N/A	4	N	N	496

Table 21. Amino acid sequences of Representative CLUST.009467 Effector A Proteins

>WP_032744684.1 [Bifidobacterium longum subsp. infantis EK3] MTKPVSIQQRIRILDAQGVSWREIARRLGVSRTVRYATMEDCSFKPAVRKGRSLIDAYS GTVDSWLFADRLMPRKQRHTAR RVYARLVEEEGFEFSYSSVQRYVVRWREHRSDGDGYLELDWSAGVMQVDFGEAVATIGGGDVKVHCLVATFPHSNMRYVAAMP GENAECVCEGLAQIFDHIGMVPVRLVLDNATGAGHRVAWNKVTVVRVFAMFCDHYRLETRFCNPFYSGNEKGSVENAVGFLRRNL MVPKPNASHRQLTRHLLSRCDATADVHDYRSGRPFVRELDFEDRGEMQPLPRARFDAVEWVERKADKEGNIQIGSVRYLAGPSW RGWTLLAGLRAFEVEIRTDGRHVNTLPSYGDGGRTVRNPATLLPALARKPNAWGESPLRGELPDGLVLRDLALDPGSRREL TLLHRAGVASGFGNAVLAMAAIVEGGRAPDEASLLALARRIAQGDPPDDGAGRLAVYDAFNPNPDGKGTES (SEQ ID NO: 1)
>WP_013401401.1 [Geobacillus sp. Y4.1MC1] MLAVAEIHYIRYEAQKGCSSYDIARMDRPTVKKYAEMEDFNPSKVKQTRKAKVMDPVKPILDQWIKEDLTKKKKYRRTAK RMYEILKEEYGFTGSDRSVRLVSKRKQELLEQSEPAALPLESKPATAQVDFGEAPFLYQGGKYVDFPYLVVSPYNSAAYVQVM PAQNQECFLEGLKRIHYMGRVPRVIRFDNLSPAVKTI LPNGERELTETQRFVLYHYGFECEFCNPASGNEKGNVESKVKYIRN NFFLPEQTIYQLESFNESLWEKCEKDWNRPHYEKERLIAELFEEEQALFLQPAKEFECEVRYEQVTADKYGFHLENNLYSTSP RFAKQKVLAKISYHEIAILTEHELI IKHERLYGTQKQSMKWQPYLTLMAKRPNALKYTD FYEKMPEEWKNYFSNCTVQEKKEA LQLLAVLLKEHDFEVSTQALRIASQYGHPKVESIKQVYQQLINGRGIREPIQPKKHVPDMPEAIRGVRHYDRFLFESQGDVASWN K (SEQ ID NO: 2)
>AEH46354.1 [Geobacillus thermoglucosidasius C56-YS93] MLAVAEIHYIRYEAQKGCSSYDIARMDRPTVKKYAEMEDFNPSKVKQTRKAKVMDPVKPILDQWIKEDLTKKKKYRRTAK RMYEILKEEYGFTGSDRSVRLVSKRKQELLEQSEPAALPLESKPATAQVDFGEAPFLYQGGKYVDFPYLVVSPYNSAAYVQVM PAQNQECFLEGLKRIHYMGRVPRVIRFDNLSPAVKTI LPNGERELTETQRFVLYHYGFECEFCNPASGNEKGNVESKVKYIRN NFFLPEQTIYQLESFNESLWEKCEKDWNRPHYEKERLIAELFEEEQALFLQPAKEFECEVRYEQVTADKYGFHLENNLYSTSP RFAKQKVLAKISYHEIAILTEHELI IKHERLYGTQKQSMKWQPYLTLMAKRPNALKYTD FYEKMPEEWKNYFSNCTVQEKKEA LQLLAVLLKEHDFEVSTQALRIASQYGHPKVESIKQVYQQLINGRGIREPIQPKKHVPDMPEAIRGVRHYDRFLFESQGDVASWN K (SEQ ID NO: 2)
>ADQ03535.1 [Bifidobacterium longum subsp. longum BBN68] MAIPMPIVQDIRRLDRQGMRAQIARRLHVDRGTVAKYADMEDCSFKPKADRRYGSKIDPYAHLVDEWLEADRLLPKQRHTIR RVHDLRLAETDYDGEYSTTMRVYVHRWREANRGVPDREGYVRLEWAAGSMQADFGVARARIAGEMADVHCLVSLFYSNMRLCVA LPGENAECCLCHGLMLVFEHIGGVPPVIVMDNATGAGRRNAKGEVALTGVSFAFVAHYRLEVRFNPFYSGNEKGSVENAVGFLRR NLMVPPMRAESYQGLSRLLLERCGLDGLSYCPRLLDVPAEVDFEERAALMPLPSTAFDPVRWESRTADKYGLVDIDSRYLA GPDSARSRVLAARIWDTVTLTSPATGELFAEYPRQYGRSRNVEDPALVLPRLAVKPRAWRESSIRPDVDDIRAWLDSMDEKTL RESLKAIGDACRAAGFDPAMQACGEILRSNRDMGLHADSLTPIALRMRDGEWEYPPGGIEEPDLSGYDRFITGTDDGGEER

(SEQ ID NO: 3)

>EFQ22705.1

[Aminomonas paucivorans DSM 12260]

MYGVVDREYIRKQHFREGWSIRKIARQLGMCRRKTVRRLLLEDSDQVPTTYTLKEPRPRPVTPGYLEVIRTWLTEDLQAPRKQRHTAR
RVYDRLVTERGFGGSESIIRKVVAELKREIAPKKGFLPLEADPGEQAQVDWGEAMVRLEGESTRVHLFCMRLRRSGTFFVYAFP
DEGLEAFLAGHRLAFAFEFFGGVPRECVYDNLKSAVTKVLQGGPHREENRQFSALRGHYLFESVFCNPRSGHEKGAVEHLVGVFVRN
VLVPVFDLPSLEELNLCLARWCEKQRRSRGTSFDEEATCLLPLPALPHPCALQTVAVVSPTSLSRVFEGNVYSVPVHGEGEAVNL
STTWDRIRISRNGTLLAEHPRLSGKGKASMELVHVLPLQFKPGAVRNAAVLRRLEPQKARTLLCAQPEGYREFCAILLHR
DHSLEVLTEALEEALALKRVTAETVRQLVWNRTPSVVPEASVPDPLAELPAGCAPDPSRYDTLLGEAVA (SEQ ID NO: 4)

>WP_011639494.1

[Syntrophomonas wolfei subsp. wolfei str. Goettingen G311]

MLTMIQQYHIKHLNFHKGKSLRSIAEETGHDFTVKKYAEQTDNEIRKPKRGRPSKLDPVKPIIDAWLKEDLNRPVKQRHTAR
RIYDRLCDEHQDIFNACERTVRTYVSARKKELYGEEGYLPLEHPAGEAQVDFGEIVMIEQGGQVKGYELILSLPYSNAGYPQI
FRGQNQECCLLTGLKDI FEYLEHVPRI IWFDMNSAAVAGIGEKGERKLVQDFYRFALHYGFKPQFCNPGKGHEKGHVENKVGYSR
RNYFVPEPSFSDIEEFNQGLFAVAEKDHCRCRHYRKGVRINELLQDDLTAMLPLPAKPFITIGRMQKLSANKYGVKVCVDTNVYSAS
PQVANKELYMRGLAHQVELLNEQYQPIVKKHRLYGQNLMDWLPYLTTLAKRPNALKYTGFIYRELDPWQDYFNNLDYAEKKK
SLNLLVRMITETDATICLLETLDGKADADSILLSYRRLTEPTFNDFLPLLSAGINSPAIYTPDLTCYDTFLKAGECR
(SEQ ID NO: 5)

>WP_015439061.1

[Bifidobacterium breve UCC2003]

MTIPVSVQQIRIRLLDQGGTPWRRAIARDLNVSRDTRVRYALMEDCSFKPKPSTTRPKLDGFTFVIDEWLQADRFPMPKQRHTAK
RVYDRLVTEKGFTGSYSYSAVQRYVKQWRQEHPRACDGFMELEWRPGVMQVDFGEAVVVLAVGKTKVHCLVASFPYSNKRIVVALP
GENAQVCCEGLVEI FEHIGLVPMVLVMDNATGAAHRTAWDKITIVGLFQLFLEHYRIEARFCNPRSGWEKGSVENAVGFLRRNL
MVPPLLNVESYAQLSKHLLSRCDELGNDKHYRADESNDLFEEDRAGMRPLPRVRFDVWDQERRADKEGVIQVGSYRYLAGPAW
RSWPLLVGLRAFDVEIRTRDGRKVASLPRAYGEQAGTVRSPLSLPALARKPRAWGESPVRGDFPEGLRTLIDAQDSARRKATF
RLLERVGAACGFDAACKAAEHIIQGRPIDEAALSIMARRIASGEAPADVAPDLSVYDDFMKTGDGDTLKEREGGKDR (SEQ
ID NO: 6)

>WP_013809817.1

[Desulfotomaculum nigrificans CO-1-SRB]

MLTVPQQEYIRLMREIEGCKIAEIRRTGVSWITAKKYADRDNWSQKTGKKHHRPIIDPFKEIIDTWLTEDQCIIPAKQRHTAA
AIHRRLLVREFNFKGSDRTIRDYVSKRRKELKLEEAQAYQRLHHPGGEAQVDFCTIQVSKNAKLMEYKLLVASFPYSNASFVYPV
PKENQECFLEGLKRLFDQMGVVKRIWFDNLAAAVVHI EKDGQRKCTDAFLRFVAHYRFEPVFCNPASGNEKGHVENKVGYSRR
NWCVPVPIFTDHTLASELVSCALADRKRKHAKGVTIEELWQESAKLLVLPEEPFQVFRLEQAKVNNYGEVRFDGVSFPLFE
VHPEEKVILKIYWDHIEVLNQNYKLIGSFPRPYTHKTADVWDKVIHNLKRKPRSAFYSQFVRMMPISIVQEFILLDDLEIRNR
LYWLSNWLKQYSLDEVAMVLKQEDWNQAMFSDVVTHRLYALRHPIVATDEISIPGTPTYTPDLTIYDRLHKGGRSE (SEQ ID
NO: 7)

>WP_007056033.1

[Bifidobacterium longum subsp. longum]

MTVPVSVQQIRIRVLDAHGMSWRSIAKEVGVARDTVAKYAGREDSPGFPERSVSRPSKLDFFRATVDGWLQADRFPMPKQRHTAK
RVYDRLVSEKGYEGSYSPVQRYVKRWREHRLPSDGYLELKWHPGEAQVDFGMAQAVVGGDRVDVHCLVVTFFPYSNMRYCVALP
GENAECVCAGLRTVFEHIGAVPLTLVLVDNATGAGHRIAWDKVTVVHVFELFVEHYRLETRFCNPNNGNEKGSVENAVGFLRRNI
MVPMLNAESHGQLTRYMLERCDAKETHYRKGAIPIGGLFAGEKMDMQPLPAKPYDAIRWEVRKADKGRVQIDGNYYLAGPSW
RGWTLDVGLRAFDVTIRTQDGRTCARLPRVYGDSPATVRNPATLLPALSARKTHAWTDSTIRDDFPDKLRIADRMDAKTRRTTF
RVIAKASAASGFEEAAVRAGEHLVEQGHAIDEASVTMTARRIAAGEKPYEQSVPDLTGYDVFMKPRQPWERKEA (SEQ ID
NO: 8)

>CCU84826.1

[Mesotoga infera]

MTQAHHIKYLYEVENKSLNSISKETGDFDRTVKKYVEKQDWNERTRNKPSKIESVKEIIDGILIQDMKHPKQRHTAVRIYQR
LLEEYPDKLQVKERTVRRYVSQKKKELYGQMGYLPKHPGGEAQADFGEVLLIEKGKEKKFYTFNLSPFSSNSSYVQLFKSTNT
ESLLTGLKVVFEYMDRVKVIWFDNLSSAVISILKNGERVLTDRFQRFALHYGFKASFNCNPSGHEKGNIENKVGYNRRNFFVP
IPSFDSIEDFNRELLKVSDEDMRRNHYKKDVLISKLFEEDRGMNELPVTTFEVSLLSVKANKYGMVTFEKNRYSTAPELSLR
ELWLRADAFTVTILDERYERVVTHARLYGEKQEAFNWYLYLPTIVKRPNALKYVAFYDQLPVQWKEFFERSDRDGKAGLKLSS
KMLSESDMATATRALIESDGEIDSILVTYHRLTQPGIGEVPLSENIPQLNYEADLSVYDGLLGSAG (SEQ ID NO: 9)

>ANW98740.1

[Clostridium stercorarium subsp. thermolacticum DSM 2910]

MLSMTQIKDIRKMYFEEGKNISQIARETGHDRKTVRAYLDKVDWNQKPKPVKKETAFFKLNPKDDIDTWLNEDKKARRKQRHT

AKRIYNRLVEKYGERFNC SYRTVAGYVAVKKKEI FNAREGFLPLEHVPGEAQADFGDADFYENGRHYRGKSLTSLFPHSNKGYT
QLFKGENQECLFEGLKAI FEHIGGVPPRIWFDNASTIVAKVI KGGGRNLTDDFMRFMEHYRFKAVFCNVDAGHEKGNVENKVG
HRRNMLVPVPRFEDI SEFNKELLIRCEEDAKRQHYRKNGTIEELYRDDKAALLELPKTTFTDTSKYITVKTNGYKGFLLNKG
LHE YSSAPKFANKYVLVRLTAFHVTVLDESHREIVRHERLYGDYKQSQMWLPYLTQLARRPGALKYTGIIYQMLPQPVKEYMEEL
SK QDRGKVLRLVIADLTQKSSFEKAIAKTVSTALSYGAADVDSLINLHRYLYEKVLQLEPIHLPEHIPLHNRYPDFMAYDRSLKAGE
EKC (SEQ ID NO: 10)

>WP_079907503.1

[Carboxydocella sp. JDF658]

MLAMDQIYNIRFQAIRKGKSLRSIARDTGHDFATVKKYVNKEDFNEVKQRRQRESKLEPYKYLIDEWLQEDLKAKPKQRHTAQR
VYDRLKEKFGEFEIISRTVRSYVAQKKSELFQKEGYLPLEHPPGEAQVDFGQAEFTENGVKITGHYLNVSFPYSNGGYTQVF
KSENQESFLTGLKNIFEHIGCVDPRIWLDNLSAAVVTIKKHGERDLTEMFKRFAHYGFECNFCNPNSGHEKGSVESKVGYHRR
NFFVPIPEFKDMAEYNKNLLELADIKMNRQHYKKQTIAELFLEDKAMKLKPEKEFEVYKLEKAKADNYGKVKFDNRISTSP
EYAGKQVWIKASAFDIVILDQDYRYIEKHKRLYGEQKESMKWLPLYLKLMAKRPTALKYSGFFKELPQTIQDYFERCEYQKKAA
LQVLSKMVALTDIHTATKAFETIKHGLTDLDSIWATFYTMTNQSPAASEIKLGDKIPDITPYSVDNSSYDCLLKGGDAGWKN
(SEQ ID NO: 11)

>WP_025263135.1

[Bifidobacterium breve NCFB 2258]

MTIPVSVQQIRILLDQGGTPWRAIARDLNVSRDTPVRKYALMEDCSPKPKPSTTRPRKLDGFTPVIDEWLQADRFMSRKQRHTAK
RVYDRLVTEKGTGTSYSAVQRYVKQWRQEHPRACDGFMELEWRPQVDFGEAVVLAGVKTKVHCLVASFPYSNKRYVVALP
GENAQCVCEGLVEIFEHIGLVMVLVMDNATGAHRTAWDKITIVGLFQLFLEHYRIEARFCNPRSGWEKGSVENAVGFLRRNL
MVP LLNVESYAQLSKHLLSRCDELGN DKHYRADES VNDLFEEDRAGRPLPRVRFDVWDQERRADKEGVIQVGSYRYLAGPAW
RSWFLLVGLRAFDEIRTRDGRKVASLPRAYGEQAGTVRSPSSLLPALARKPRAWGESFVRGDFPEGLRTLIDAQDSARRKATF
RLLERVGAACGFDAACKAAEHIIQGRPIDEAALSIMARRIASGEAPADVPAPDLSVYDDFMKTGDGDTLKEREGGKDR (SEQ
ID NO: 12)

>WP_013400884.1

[Geobacillus sp. Y4.1MC1]

MLAVAEIHYIRYEAQKGCSSYDI AKRMNRDPRTVKKYAEMEDFNPSKVKQTRKAKVMDPVKPIILDQWIKEDLTKKKKYRRTAK
RMYEILKEEYGTGSDRSVRLYVSKRKQELLEQSEPAALPLESKPATAQVDFGEAPFLYQKGKYVDFPYLVVSFPYSNAAVQVM
PAQNQECFLEGLKRI FHYMGRVPRVIRFDNLSPAVKTI LPNGERELTETTFQRFVLHYGFCEFCNPASGNEKGNVESKVKYIRN
HFFLPEQTIYQLESFNESLWEKCEKDWNRPHYEKERLIAELFEEEQALFLQLPAKEFECEVRYEQVTADKYGFHLENNLYSTSP
RFAKQKVLAKISYHEIAILTEEHELIIKHERLYGTKQKSMKWQPYLTLMAKRPNALKYTD FYEKMPEEWKNYFSNCTVQEKKEA
LQLLAVLLKEHDFEVSQTALRIASQYGHPKVESIKQVFYQLINGRGIREPIQPKKHVPDMPEAIRGVRHYDRLFESQGDVASWN
K (SEQ ID NO: 13)

>WP_011742685.1

[Bifidobacterium adolescentis]

MTTPVRIQQSIRQLETGKLTHTQIARELGVSRTTVVKYATRDYSVPPTGGTASRSLVAGEYARKADEWLEADRRMPRKQRHTA
RRVHERLVEECGFEGSYSSVQRVVKRWQRHRGEAEGFSELEWAPGSAQVDFGQALAVVAGVERTVHFLVVSFPYSNMRYVAAL
PGETAECVCHGLNKVFSHAGMVPRVLVFDNATGVGHRRADGTVTQTRLFSLFCAHYGFETRFCNPYSGWEKGSTENAVGFLRRN
LMVPTPSAEGWDRDLTDAWLERCDEIARGVHYREDVPIRDLFETDLDHMLPLPSPMFDACDWRGVKADRTGTVTIDSNRYLAGPK
WHSMRLMAGVRALRVELRSMDGEPITLERRWGRSPDTAMDPLSLAI IARKPRSWGESP IRSDFPEETRVLLDRMEPRERGLL
VDDIRHAAVVS GFRAAVMAVVEIVAAGRRPDRAAIDQTARRIAQGDGPESRARLDYTSRFMRREDGDE (SEQ ID NO: 14)

>ADZ71261.1

[Polymorphum gilvum SL003B-26A1]

MDLYLKVRLAVSEGMSRRQA AKHFNISRDSVSKMLSYSSTPPGYRRQAPVRRPKLDAFVSTIDHWLDEDLKVPRKQRHTAKRVFD
RLRAECGFTGGYTI IKDYMREREQRQEVFVPLSHPPGHAQADFGAEAVVIGGVEQKAHFFVLDLPHSDACYVRAYPAVSEAW
IDGHVHAFAGGAVPQSIYVDNDRCLVSKILADGTRKRTVLFSGFLSHYLIRDRYGRPGKGNKGSVEGLVGYARRNFMVPIPR
FATWDEFNAWLEEQCRKRQRDRLRGESETIAERLQDLAAMRPLPASPFACDQASGRVSSQALVRYRTNDYSVPVAFGHQDVW
IRGYVDEVVIGCRGEI IARHPRSWEREDVVFDP LHYLPLIEQKINALDQAAPLQGWDLPEAFATLRLMEARMAKHGREYVQV
LRLLESFELADLHVAVKQALQIGAIGFDAVKHLLCQVERRPRLDLSIYPYLPRATVEKTS AKAYMRLSTHAGDAA (SEQ
ID NO: 15)

>KYO67332.1

[Thermovenabulum gondwanense]

MLTMTQINDIRHSFFFKGENISEIAREFFKKDRKTVRKYIYQEDWNMMLKAKEIENAFPKLNPFKADIDQWLEEDKKARKKQRHT
AKRIYDRLCEKYKDNFNCSYRTVAAYVASKKKEIYQDNSSCLPLEHIPGEAQVDFGEADFYENGILYHGFYLNVSFPYSNAGFT
QLFKGENRECLFQGLKNI FEYIGGVPPILWFDDPSTLVAAILKDGGERKLTDAFLRFKEHYGFEAVFCNPAAGHEKGNVENKVG
YHRRNFLVPVPKFEKLEDNFNELLQCEQDMYREHYRKDGTHAEI FNEDEKALLKLPDI PYEVADYITVKTNAYAKFSLNGGKHI
YSTAPKYANSRI LVKITANEVILPDESHREIVRHRRLYGD SKQESMDWLPYLTQLSRCPGALKYSGIYNMLPDP LREYLEGCSK

SER GKALKILASLCIESSFEKALSAVTQAILYGVKDLDSLIAHSRITGITPQLELVRLPEEIPELKAFS FNAEEYDKAFLKGG
 ANIYAKG (SEQ ID NO: 16)

>OI083550.1

[Actinobacteria bacterium CG2_30_50_142]

MLKMDQIQEIKNLSKEDKSLREIARQTGYAFETVKKYIEKQDFSPVLKPKQKRSGLNPNYKHHIDKWLENDKKAKRKQRHTAQR
 VYNRLKEIYGDEFKVSRSVRKYVADKKADSRQQADGSLPLNHPGGEAQCDFGDAQFIENGQIFNGHYLSLSYPYSNAGYSOLF
 KSENLECLQEGLVAFIHAFTGRVPTKIWFNDASTIVKKIREEGKRDVTESFRFELHFGFTSNFCNPDSGNEKGSVENKVGYHRR
 NLLVPIPEFTCIQAYNRELLTRCDADMLREHYKKERPIEELFMEEKLLMSNLPGSVFEVCRLEKAKANNYGKVRFETNTYSASP
 DFAAQQVWVKAGAFITISILSDDFRLIQSHERLYGKNKESMKWEPYLRLLSKRPRALKYSSFFSELPQSVQDYFDRLDYQGGKRAA
 LKVMAMKVEDTDISTAQIAFELAEGRVEDIDSIWATYSSLTKSDIQVKDIAAPAVATYVPDINSIYDVCLRGV (SEQ ID
 NO: 17)

>AEG16465.1

[Desulfotomaculum kuznetsovii DSM 6115]

MAQQQYIKHLYEQECSIAEISRRVGVNWRATAAKYAKKDDWNQPMQOTRRRQPVMGSEFVDIVDTWLMEDMLKPRKERRTAATY
 RQLREQYGFKGSGRTVRAVYVSRKKELRTGAQEKYLRLEHPPGQAQVDFGTSHVVWDGVLREIKYLTFSFPYSNAGFCVPVPSE
 NTECLLYAMILVFEWIGGVPEIWFNDLSAAVVGVGKGEQRELTETFRFMLHYRFEAKFCNSGKNEKGHVENKVGYTRNWL
 IPYPTVSSYEELTAELYRRALEDQMQRSHYKGVITAELEWEEKKALLPLPATPFEPVRFETARVNKYGVQCCGEVYVPSAQV
 GETVLLKLWDCVEVFSSTQERLAHLPRHYTLTKTPIDWKGYFRIFIRKPRGARHATMYRFLPEPVRQYLEACDPEAYRERLKF
 IHTLLEEGFTIEDIARVLEEVAKYPAADSALIRHKLYRITFPDAPLEALNETYTPDSVRRYTPEINIYDRLVPQAGGKGELRD
 QVVGEGAV (SEQ ID NO: 18)

>WP_044746322.1

[Anoxybacillus geothermalis]

MLAVAEIDYIRYANQKGCSSYSDIAKRTNHDPRITIKKYAEMEDFSPTKRKQTRKARVMDLVKPILDKWIQEDLTKKKKFRRTAK
 RMYEMLKEQDFKGSRSRVLVYAKRKKELQESESAALEAKPATAQVDFGEAPLVYKGERVILPFLALSFPYSNGFYFQVM
 PSQNRECFLEGLKRIFHYMGRVPKVIREFDNLSPAVKKILPNGERELTEEFQRFVLHYGFECEFCNPASGNEKGHVEAKVKYIRN
 NFLLPELTIYDLESFNDLSLEKCEERDQWRPHYQKGTIAGLFEEKECFLLQPAKEYECVRYEQVRADKYGYIRVENKLYSTSP
 RFAKQMVLAKEYEIEILTEDYELIVKHERLYGTQKQSMKWQPYLILMAKRPNALKYTDFYEKMPSEWQNYFSHCTIQEKQKT
 LQLLAVLLKEHDFEISTQALRLASQYGHPSVDSIKQVIFYQLINGRGIREPIQPKISVPTMPEASRGLSHYDRFFESTGGER
 (SEQ ID NO: 19)

>WP_044746322.1

[Anoxybacillus sp. B2M1]

MLAVAEIDYIRYANQKGCSSYSDIAKRTNHDPRITIKKYAEMEDFSPTKRKQTRKARVMDLVKPILDKWIQEDLTKKKKFRRTAK
 RMYEMLKEQDFKGSRSRVLVYAKRKKELQESESAALEAKPATAQVDFGEAPLVYKGERVILPFLALSFPYSNGFYFQVM
 PSQNRECFLEGLKRIFHYMGRVPKVIREFDNLSPAVKKILPNGERELTEEFQRFVLHYGFECEFCNPASGNEKGHVEAKVKYIRN
 NFLLPELTIYDLESFNDLSLEKCEERDQWRPHYQKGTIAGLFEEKECFLLQPAKEYECVRYEQVRADKYGYIRVENKLYSTSP
 RFAKQMVLAKEYEIEILTEDYELIVKHERLYGTQKQSMKWQPYLILMAKRPNALKYTDFYEKMPSEWQNYFSHCTIQEKQKT
 LQLLAVLLKEHDFEISTQALRLASQYGHPSVDSIKQVIFYQLINGRGIREPIQPKISVPTMPEASRGLSHYDRFFESTGGER
 (SEQ ID NO: 19)

>WP_024833244.1

[Clostridium josui JCM 17888]

MLSMDDIKYIRRMHDVEGCSIREIMRRSGYHYETVKKYLDMDNFNPTLLPKEVPSLLDPLKPIIDEWLENDLKAPRKQRHTAK
 RYERLLNEYPGQLEVKLRVTQYYVSKKKKELYAEKSKGYIPLEHPAGEAQVDFCQFFYYDNNTLKEGRKLTVSFPQSNAGAYC
 QAFRGENQECLLQGLKNIMKHMKNKVPFRMVFDNLAAVAHIGSGKDRIITEGFKRFEHYGIEPVFCNASAGWEKGNVENKVG
 ERRNMFVPVPTILDFDQFNRLKCCERDMQREHYRKGHLIADLFEADRQAMLPMPIDFKVSRFQAAKADKYGKVMFETNLYS
 SSPKLAQEHVYLEITSDSVTMDIKYNPVVTHCRLYEKDGESMDWLPYISLMAKRPNAIKYTGIFYQELPEIWQNYLAELSPEKK
 REALLTNTMLQKHDIAAAANALEVALDSGVKSDSILASYRLTNKVQQLQPMQFTNPYIKVPSFKTDNSRYDSLFGKEVSQ
 (SEQ ID NO: 20)

>WP_074011080.1

[Numidum massiliense]

MLTMVHVHYIKSLYEREGLSLREIQRRTTHDFKTIKKYIDMEDWNETPKRKPRRSKLDPYKETIDDWLEGDLKAPRKQKHTARR
 VYERLKEMHGEAFNVSLRTVQYYVAKKKTALKEKTRGYIPLEHTPGEAQVDFGEATFEEKQAQFGYYLNVSPYSNAAAYVQVF
 RAQNTCECLLEGLKRIFAHVGGVPTRLWFDNLAAVATICNGERQLVDSFARFAYHHGFTEPTFCNPNSGNEKGHVENKVGYHRR
 NFFVPIPKFEDLDAYNADLLQRCERDMQRTYKHQEQIAERFQADAAQFKPLPRTAFDVHRMVAKTKDYGKVCQKKNRYSTSP
 GLARQEVWLKVTCRIVIVQDESRTIVTHSRITYEGEGKEAMKWTPYLSLMAKRPNAMKYTAFYRQLPDPWREYLTVKERRPKGLR
 LLAEFLSQGDMQIATTALAEETLAQGVDDADSLATWYRLTGRANECSTLAVSHDVP SHAIYKTDLGSYDHLIRIRKGDGR
 (SEQ ID NO: 21)

```
>WP_082440433.1
[Bifidobacterium aesculapii]
MPIVQDIRRLDRQGLSRAEIARRLHVGRGTVAKYADMEDLSPRRPDDRRYGSKIDEFAVVVDGWLEADRRMPRKQRHTARRVYD
RLVEEHGYGGSYSVVQRYVKRWRETNRESSAAFLDLEWAPGVMQVDFGVALATVAGRGTETHCLVVSFPWSNMRYCVAMPGENA
ECLCAGLMLVFAHVGGVPPVIVMDNATGAGHRNAKGEITLTHVFESFVAHHRIEVRFCNPYSGHEKGSVENAVGFLRRNIMVPP
MRAESYEQLSRLMLERCDVLADSSYCPRLMDVPVAVQVFAEDRAALAPLPSTVFDVVRWESRRADKRGRIIDISRSYLAGGAWAR
RRVQAARWDTVTLVDPDTGVVLAIEYPRPYGDGPRVLQDPALVMPMLAARPGAWGESPIHPDVPDDIRSWLDSMDDKTLRESLK
AIAHACQAAGFAPVMQACGEILRCNKDMGLHEDTPAPIALMRDGDWEYPGAIDGPDLTGYDEFITGTRHDEKETR (SEQ ID
NO: 22)
```

Table 32. Amino acid sequences of Representative CLUST.009467 Effector B Proteins

```
>WP_032744686.1
[Bifidobacterium longum subsp. infantis EK3]
MTRPSRTMADTRRRRASTGRMMDEIMELARSLPLTRQVLADSLETATPAQMEFMLSWMNEELASRERSKRARLLKQAGLPGVKE
LDGYDWTVPVRFVVDYGREALESDFVANSEDVVLFGPPGTGKTHLAVALARKACVEGQPTRFFTAAGLVMRLLRASAEGRDLRE
LALISKARLLVIDELGYVPVDEEGSRLLFQVVTNAYERQSVIYTTNIEFSGWGRVFGDPNMATAIVDRTVHHGRMIRFEGESYR
RTHALMG (SEQ ID NO: 23)

>WP_013399917.1
[Geobacillus sp. Y4.1MC1]
MEQMIKEYAKRLKLSWIPANYHTIHAETNEEFLLKLFEREVQHRDERRINLLLKQATLPKIPNKPYPDWREIQ LAPGITKDYILE
GEFTKNQENLI FYGGVGTGKTFLSTLIALNLMKKQGKRVKIFYTAASIVNALLEANEKGD LGKFLNQIEKLDLLI LDELGYIPLH
KQGAELLFQVISMICYETKSIIIVTTNLPFSQWNNVFGDPILTEAVIDRLIHSHSLIIFNGESHRYKDSISQR (SEQ ID NO:
24)

>AEH46355.1
[Geobacillus thermoglucosidasius C56-YS93]
MEQMMKEYAKRLKLSWIPANYHTIHAETNEEFLLKLFEREVQHRDERRINLLLKQATLPKIPNKPYPDWREIQ LAPGITKDYILE
GEFTKNQENLI FYGGVGTGKTFLSTLIALNLMKKQGKRVKIFYTAASIVNALLEANEKGD LGKFLNQIEKLDLLI LDELGYIPLH
KQGAELLFQVISMICYETKSIIIVTTNLPFSQWNNVFGDPILTEAVIDRLIHSHSLIIFNGESHRYKDSISQR (SEQ ID NO:
25)

>ADQ01872.1
[Bifidobacterium longum subsp. longum BBM68]
MSVRPDPVIPETRRRRRASTTEKSERILKMSRSLTLTRSVLAGTLAEATPNQLDFIERWFTAELDSRERSKRRLRLKQAGFPADK
TLDGYDWTNLMKPADWGQAQLENLDFVAGCEDLVLYGPVGTGKSHLAIAIGRLACERGVVPRFFTATGLLMRLRRAQQENRLDR
ELASIGKARLLIIDEFGYLPIDEEGSRLLFQIISDSYETRISIIYTTNIEFSGWGRVLGDKNMAAALIDRTVHHGRLLIRFEGRSY
RSEHALMTK (SEQ ID NO: 26)

>EFQ22706.1
[Aminomonas paucivorans DSM 12260]
MSGPTHDRDTAVQTLCDRLKLPGLRHYRNGIVPTDPKDYLQESLEAERTSRREKRLKNLLKGARLPYARTLSEYDFLRLPSLP
KEQILSLAQGDFLRRRENVLVLLNSGTGKTHVAIGVAASCIAGYAVRFTTALT LAQELLQAQDEHRLPRVLKSYDRYDLVVVD
ELGYLGLGPGGAPLFQFFAERYERKSVCIITNLEFGRWAEVFGDATLTEALLDRLTHHAHIFVFRGESYRFAQRSARTMTA
(SEQ ID NO: 27)

>WP_011639493.1
[Syntrophomonas wolfei subsp. wolfei str. Goettingen G311]
MKELIKDYCKQLRWGNSIVQNYADIKADTHEEFLAKLLEIEVKGRELNRKNRCLRQAGFDVIKTFNGYSFDHIEIPASIPDDL
KTAALLKNRENLI LYGPVGTGKTHMATAIGVEACNQGKKVGFYRTSTLVNELNDAQNAGNLGKMLKQAKLDLLICDEWGYIPL
DRQGAQLLFQVADCYEKRSIIITNLEFSKWNGIFYDEKLTS AIDRLIHSHSLLVFGGQSYRLTHSTIRK (SEQ ID NO:
28)

>WP_015439060.1
[Bifidobacterium breve UCC2003]
MNQKNGRTGPGPLRERARSLFISQATIDETLEWATPRQLDAIDRLMATEPANREASKRARLMRQARFPVPKSLDGYDFANVRLP
DGYTKEQLTGLDFAAKAQDLVSYGKTGRGKTHLATALGMLAIEQGRSVRFRQTAEVLPQPGKAKRDGALDSLRLDLARADLIIL
DEFGYVPFDIDGARLLYQIIAGSYERRSIIFTNIEFSKWGTIFADDKLAAAIIDRIVHHGRLLIETGFSRRVSQALMFGKTDN
Q (SEQ ID NO: 29)

>WP_013809816.1
```

[Desulfotomaculum nigrificans CO-1-SRB]
 MSQLLSDLCRQLRLAYVSEAVEEEKASPELCTLITEVLKAEIEGRKRMKLNRLLRHAGFPQIKTLEGYSFEPIAFPASCNQGKLV
 SLDWIAQKENVLMGLGAVGTGKTHLAIALGVEACRKGKIVRFFRVSDLVSVLQQKHAEGLTKFRKELLKADLLILDELGYVPFH
 QTGSELLFHVIADCYERQSVIVTSNLEFGQWSSIFGDTKLTAALVDRLVHHAHILAFSGDSYRLRHALGNIQN (SEQ ID
 NO: 30)

>WP_007056032.1
 [Bifidobacterium longum subsp. longum]
 MGTHRPPVMPDMKRRRASTGERIATIMTLARRLPLTRQVLADQLADATPSQMEFMESWMNAEIESRERSKRSRLLKAAGFPAD
 KELGGYDWTPIRFPVDYGRGQITTLDFIAGHDDLVLFGPPGTGKTHLAIALGRNACRKGIPARYFTAAGLVMRLRLRAQSDNRDL
 RELAAIGRTPLLVIDELGYVPIDEDGSRLLFQVVTNAYETQSIIYTTNIEFSGWGRIFGDPNMAAAIIDRTVHHGRMIRFEGES
 WRKTHALMQ (SEQ ID NO: 31)

>CCU84827.1
 [Mesotoga infera]
 MKELISDYCKKRLGKSIATNYLEIAKSHEEFLVKLLQLEVENRNTARKNRYLKQAKFEIVKSFEGYDFENVQIPESITIEAL
 KNGEFIEREENLILYGPVGTGKTHLATAIGVQACSKEKKVRFRTVTLVNELIQAKEEGRLLGRLKQNLKDLLICDEWGYIPL
 DREGSQLLFQLIADCYERKSVIITNLEFSKWNIGIFYDDKLTSAIIDRLVHHCHLIVFTGRSYRLEHSNIRG (SEQ ID NO:
 32)

>ANW98739.1
 [Clostridium stercorearium subsp. thermolacticum DSM 2910]
 MLTSDIAACRRRLRLSRNIVEMSGKIQAVSHQEYLLKLLQSEIRHREELKKDKLLKKAGFYTIKTFESFRFDEVKLPSTPEY
 LKECEFIENKHNIVMYGNVGTGKTHLSIALGVEACKKGLEVRFFRTSALVNRLAEQKKAGTSLGFLKDLNKADLLICDEWGYVP
 LDRIGAQLLFEVISECYERKSVIINTNIEFSRWNVVFYDEQMTGAIIDRLHHCHLLLFPGQSNRMREAVLNT (SEQ ID
 NO: 33)

>WP_015439062.1
 [Bifidobacterium breve NCFB 2258]
 MSRIQNPAPDATTRRASTNRKMETIMRLARRLPLTRQVLADQLETATPSQMEFMDQWMNAEIESRERSKRSRLLKQAGFPAVKT
 LDGYDWENIRFPVDWGRSLESLEFASRPEDVVMFGPPGTGKTHLAMALGRKACLEGMTVRFFTAAGLVMRLRLHASTEGKLDRE
 IASIGKARLLIIDELEGYLPIDIDGARLLFQVVADSYEKRSVVFTTNLEFGRWGEVFGDGDMAAAVIDRIVHHGRIVRFRGESY
 RTHALMQ (SEQ ID NO: 34)

>WP_011742684.1
 [Bifidobacterium adolescentis]
 MSESTDAAGRRRGRTVSKSTPDEVMEACGLPLTRSVLRSVVAGATPGQLGVLADLFRAENASRTESRRARLIRNAGFPVCVK
 GFDGYDWGMASFPADWGREQLMDLGFVDRAEDLVLYGDVCGKTHMAIATGMLACERGMFVRFFTAAGLVMRLRLRARDENRLDA
 ELRAIGRARLLIIDELEGYLPIDIDGARLLFQVVADSYEKRSVVFTTNLEFGRWGEVFGDGDMAAAVIDRIVHHGRIVRFRGESY
 RNSHSLMK (SEQ ID NO: 35)

>ADZ71260.1
 [Polymorphum gilvum SL003B-26A1]
 MQVLELARCEWVERRENVIALGPSGTGKTHVALGLAACQKGLSVGFTTAAALVSEMMEARDERRLLRFQKQMAAYKLLIIDE
 LGFVPLSKTGAEELLFELIAQRYERGATLITSNLPFDEWTETLGSRERLTGALLDRLTHHVNIEMNGDSYRLAQSRARRAG
 (SEQ ID NO: 36)

>WP_068747786.1
 [Thermovenabulum gondwanense]
 MLRDEIADCCALKLSQNFANFDKIEAKSHGEYLLKLLKLEIEHRETRKRERLLKKAGFYTIKTFADYIFDEIKLPTGLTQKD
 LKECKFIEEKRNLILYGNVGTGKTHLATAIGVEACKKGLNVKFFRTAALVNRLGEARKRGELSGLLKQFSKLDLLICDEWGYVP
 LEREGAQLLFQVISDCYERRSVVITNLEFSRWASIFYDEQMTTAMIDRLIHHSYLLIFDGQSYRMRQSLMRQLS (SEQ ID
 NO: 37)

>OI083549.1
 [Actinobacteria bacterium CG2_30_50_142]
 MKELIKKHCRKLGLGTRIYENYASIKADTHEEFLEFLLAMEVEARTVNRKSKYVKTAGFDVIKTFENYTFEDVQIPASISIEDI
 KQAAFVERKENLILYGPVGTGKTHLATAIGISACNNDQKVFFRTAALINRLLDKQNKELKKFLRQIEQADLLICDEWGYVPF
 EKEGSQLLFQVISDCYERKSLIITNLEFSKWNIGIFYDEKLTSAIIDRVIHSHLLMFTGGSYRLKHSTINC (SEQ ID NO:
 38)

>WP_044746317.1
 [Anoxybacillus geothermalis]

METQMKEYAKKLKLSWIPANYQTIQAETHEEYLLKLFHEHEVQHRGERRINLLLKQATLPKIPNKPFDWKDIQLAPGITREYVLQ
 GDFIENQENLI FYGGVGTGKTF LSTLISINTIKKKGKKVKFYTAALVNALLEANDKGS LGKFLKQIEKLDLLVIDELGYIPLH
 KQGAELLFQIIISM CYEAKSII VTTNLQFGQWNNIFGDPILTEAVIDRLIHSHLIIFKGD SHRYKDSISLNQ (SEQ ID NO:
 39)

>WP_024833243.1
 [Clostridium josui JCM 17888]
 MKEQLLECKQLRLSARFAENAINANGATNQEYMLEVLKAEI IYRNTKRRNLYLKKAGFDNIKT FNQYDFEDITLPSGVTIDLL
 KQAEFLSRQENLILYGRNGAGKSHMATAIGVEACMQGKHVRFYKTAALVNELLA AKANGSLVQMLKKLSKLDLLICDEWGYIPF
 DAEGSQLLFQVIADCYEKRS LIITNIEFSKWNIGFYDDQLTAALIDRLVHSHLIVFGRDSWR LKHSIMKQSN (SEQ ID
 NO: 40)

Table 43. Nucleotide sequences of Representative CLUST.009467 Direct Repeats

CLUST.009467 Effector Protein Accession	Direct Repeat Nucleotide Sequence
WP_032744684.1 (SEQ ID NO: 1)	CGGATCATCCCCGCGCGTGCGGGGCAAAC (SEQ ID NO: 41)
WP_013401401.1 (SEQ ID NO: 2)	GAGCTTGACGTGAAATCTGGGGAACCTGACGCGGGGAAGAAAGAGGGAATTC AAGG ATTGTCTTGAAT (SEQ ID NO: 42)
AEH46354.1 (SEQ ID NO: 2)	GAGCTTGACGTGAAATCTGGGGAACCTGACGCGGGGAAGAAAGAGCTTGTC AAGGG AATTC AAGGATT (SEQ ID NO: 43)
ADQ03535.1 (SEQ ID NO: 3)	GTTTGCCCCGCATGCGCGGGGATGATCCG (SEQ ID NO: 44)
EFQ22705.1 (SEQ ID NO: 4)	GTTTCCATACCTCATAGGAAGGCTTGAAGC (SEQ ID NO: 45)
WP_011639494.1 (SEQ ID NO: 5)	GTTTCAATCCACGCACCCGCATGGGGTGCGAC (SEQ ID NO: 46)
WP_015439061.1 (SEQ ID NO: 6)	ATTTCAATCCACGCTCCCGGATGGGGAGCGAC (SEQ ID NO: 47)
WP_013809817.1 (SEQ ID NO: 7)	GTTTCAATCCCCTTGATAACCGGGAGGTAATTGTAAC (SEQ ID NO: 48)
WP_007056033.1 (SEQ ID NO: 8)	GTCGCACCCCACTGGGGTGCGTGGATTGAAAT (SEQ ID NO: 49)
CCU84826.1 (SEQ ID NO: 9)	GTTGCGTAAGAACCATAGTGGTATGGAAC (SEQ ID NO: 50)
ANW98740.1 (SEQ ID NO: 10)	CTTTGAATCTAACCATATTGGAATGTAAT (SEQ ID NO: 51)
WP_079907503.1 (SEQ ID NO: 11)	CTTTAAACATCTCCTAATTCCTATGGGAATTGAAA (SEQ ID NO: 52)
WP_025263135.1 (SEQ ID NO: 12)	ATTTCAATCCACGCTCCCGGATGGGGAGCGAC (SEQ ID NO: 47)
WP_013400884.1 (SEQ ID NO: 13)	GTTTTTATCTTACCTATAAGGAATTGAAAC (SEQ ID NO: 53)
WP_011742685.1 (SEQ ID NO: 14)	ATTCCTAAGCTTTTACGCTTAGGACTTCATTGAGGT (SEQ ID NO: 54)
ADZ71261.1 (SEQ ID NO: 15)	GGGTTCATCCCCGCGTAGGCGGGGGAGAC (SEQ ID NO: 55)
KYO67332.1 (SEQ ID NO: 16)	GTTTTTTCTTAACTATGTGGGATTTAAAT (SEQ ID NO: 56)
OIO83550.1 (SEQ ID NO: 17)	GTCTCAATCCTCTATCTACCGAGGTTATGATTTCTAC (SEQ ID NO: 57)
AEG16465.1 (SEQ ID NO: 18)	GTTGCACCCCACTGACCCGACGAAAAGGGGACTGAAAG (SEQ ID NO: 58)
WP_044746322.1 (SEQ ID NO: 19)	GTTTTTATCTTAACTATGAGGAATTGAAAC (SEQ ID NO: 59)
WP_044746322.1 (SEQ ID NO: 19)	GTTTTTATCTTAACTATGAGGAATTGAAAC (SEQ ID NO: 59)
WP_024833244.1 (SEQ ID NO: 20)	GTTTCAATTCCTTATAGGTATTCTATAAAC (SEQ ID NO: 60)
WP_074011080.1 (SEQ ID NO: 21)	GTTTCAATTCCTCATAGGTAAGATCGTAAG (SEQ ID NO: 61)
WP_082440433.1 (SEQ ID NO: 22)	CCCTCAATGAAGTCTCCGGAGAAATACTCCGGAGAT (SEQ ID NO: 62)

Table 54. Non-coding Transposon End Sequences of Representative CLUST.009467 Systems

>Left Transposon End for AEG16465.1
 GGCGGTCACTTCCAGTAAATATTCTACCTGTTCCACAGCCGGAAGAAAAGGGGTCACTCACTGGTTGTCATATTCTGGTGAATA
 GACGGACTGACTGACTGACCTGAAATACCAGAACAATATGCGAAGCAGTTGCAGATTATTCTGGTATCGGAAATAGTTGCAGAT
 TGTTTAGTCTGCGAAAATGGGCAGGGGTTGGGCTAAATTAAAGTGAGAATTGACAGCAGTGCGTCCCAGGGCTGGGCCGTTGCA
 CCCACTGACCCGACGAAAAGGGGACTGAAAGAATCGCAGTACGACCTGTGCGACGAACTGCGCCAGGTTGTTTTTGGCATAATA
 GAATTGCATAAAACGGCAGCCAAAATGCATAGCATGGCGCAGGTAAAGAGGGCATTGCCATCACGCTCTAGAATAACCCCTTC
 TGACTTGGGCAAAGGCAGGAGGGGATACAAGAGGTGATAGCA (SEQ ID NO: 63)

>Right Transposon End for AEG16465.1
 CCAGTCGGAAGATAAAGATAATAGGATTACGACCACAGCGAAGTAAATCTCCGATTGCGAAGGAGAATAATTACGAGCGTGA
 TGCGAAACTCTGCATTTTTTGTATGCCGAATTCATTTTTTTACTTGCCAAACACACCAGGTTG (SEQ ID NO: 64)

>Left Transposon End for EFQ22705.1
 CCTGCGAAAGGAACGGGCATTCCGGTTCGAGGGAGGACGAGAGGAG (SEQ ID NO: 65)

>Right Transposon End for EFQ22705.1
 GCGGCATCCATCAGGCAAGGGGATTGGGGTGGGTCCCTTTTCGAGGATCACACTGGGCGCATTTTCGGTTGACGAAAACACTC

CATTGACAGGCCGAACCAGGTATGGAGCACGGACATCACCTACCTTCCCATGAACAGCGGCTTCCTGTACCTTGTGGCGGTGAT
CGACTGGGCACAGCCGGTACGTCCTGTCTGGGGACTGTCCAACACCTTGGACGCGACCTTCTGCATCGAAGCTCTGGAAGAGGC
CCTGTCCCAGGGGACACCGGAGATATTCAATACCGACCAGGGCAGCCAGTTCACCAGCAAGGCCCTTCTGGGGCCACTGAAGGA
CCGGGAGATCCGCATCAGCATGGATGGCCGAGGGCGAGCGTTGGACAACATCTTCGTGGAACGTTTCTGGAGGACCTTGAAGTA
CGAATGGCTCTACCTGAACGACTACCAGGAGGTTGAGAGCTCAAGTTGGGACTTAGAGAGTATATGAAGTTCTACAACGAGCA
ACGGCTTCATTTCATCGCTGCGCTACCGGACTCCACAGGAAGTGCACCTTACCACCCCGGGTTCATGGCGGTCTGACCTTAACA
AACCTTGCCGCTCCCGTGGTCTTGACATTGGGGTCCACCATACCTCCGGTCACGTCCATCGACTCGTCACGGC (SEQ ID
NO: 66)

>Left Transposon End for WP_032744684.1

GGCGAGCAAAAGAGTTGTTAACTATAACAGTAGATCGCTATCCGGCGATATCTTCAATACTAGAAGATGCCCAAGAGGATATCT
TAGCCTGTTTTTCCCTGCCGGAGAGTCATAGGAAACGGATGAGAACTACCAATATGCTTGAGAGGTTTAATCAGGAGATAAAAC
GCAGGACTAAGGTGGTCAAGATCTTTCCCAACGAGGAAGCTGCGCTAAGACTCATCACCGCCTTGGCCGCCGAGCAGCCGAGG
AATGGATATCAGGCAAGAAGTATCTTGATATGTCGGAAGGGTCAAGTATCAGTGCAAGCGGCGAAACCAGAGCCGGCTCTGG
TTTCAATCTAGGGGTTGATCTTGGGAACGGATTTACAGAAAAATTGGGACTTGATCATAGTGCAACGGAACCTTAATTATTCT
AGCGGTAACTGGAATTTTCTTAGCACCCCAACAGTACTCACCCCTTTTAGTCGAAAGACTTTTCAGGGGGTTGAAAG
(SEQ ID NO: 67)

>Right Transposon End for WP_032744684.1

AAACGGAACCTTGGGGTGCTAGAGAAATTAAGTTCCGAACGCAAGAGTTTTTAGTTGCCAAACACATGTCAACAACCGTTCCCTTA
AATCGGGTAGGAGGTGGGTGATTAATCCACCCACCGTCCTCCACACCACCGTACAAGCGGGTCCGCATACGGCGGTTTCATCAG
AAA (SEQ ID NO: 68)

>Left Transposon End for WP_044746322.1

TTCGGCTATTGATGAACAACCTGATTTCTGAAAATGGGGAAAAATTCACCTCTGAAGAAGCGCAATACGCTATCGATAATTTAAA
ATAATAAAGTCACTATTTAGGAGGGATAGGTGTTTGTGGCAATCTATTAATGTAGTGTATGGCAATTAATTTATGTAGGCTCC
AGCGGCCATACATAAAATAAAAAACCACTTGCCAACATTCCTAAATACCATACCCTCCTAGTTGGGCAACTTGTTCACGGGAG
GTAATGAAGGAG (SEQ ID NO: 69)

>Right Transposon End for WP_044746322.1

CGGAATTTGGCAAGTGGCATTTTTAATTGCCATTTCGTGCATTTTTTACTTGCCAAAAACAGGATAGGCCTACTCGTTTTGAGT
AGGCTTTTATTTGCATAACTATACTCCTTTTTGCCTATCCTTTTAGATGAAGGAGTGAGATC (SEQ ID NO: 70)

>Left Transposon End for WP_013400884.1

GTAACCCCTCCCGCTGATAAGACGTTTATGTATGTAAGGTAACAGGAAATAAATAAAATTCATCTATTTTCTATATTTATG
ATGATTATTGCATTCAGTCGCTCTATAAAAACTTATCGTATACCTAACTAAATAAAAAATTGAAGATTTATGGGAATCGTGTTT
ATGGCAAACTAAAAATGTAACGTTTGGCAATTAATTTATGCAGGTCTATAAAACAAAAACCACTTGCCAACATTCCTCAAT
ACACTACCCTCCTAGTTGGACAAAACAGTTCAACGGGAGGTATTCAAGGAG (SEQ ID NO: 71)

>Right Transposon End for WP_013400884.1

CCATGGAATTTGGCAAGTGTATGTATGCATTTTTAATTGCCATTTTCATACATTTTCTACTTGCCAAAAACAGGGAATCGGAAGA
ACATGAATAGGCGGTGAGTACCTTTTTATAGAAATCATAAATTTCTCTCTCTATCTATATTTTCTACTAGCTGCATATATAT
ATAGTGAAAAAATAACTGGAGATTAGAAGCTATGACGAGCGTACGCCAAGATGCATGGAAGCAAGATGAAGATTTGCTGTTGTC
AGAAGTTGTATTGCGCTATATTCGAGAAGGGGGAGCTCAGCTCGTGGCATTGAGGACCGCTATCGCAACACCGGCATCCTG
CGGATTTCACTGGAATTCATATGTGTGCAAGTAATACACAGAAATGATTGAACTAACGAAAAGGCAGCAGAAGGAAACGTAACA
TTGGTAAGTGAAGGAAAGAAACGAGAAAGACGCTGAAGACAAATTAACAGGAGCGCACGTAATTTATTTTGAAGCCTAGGAA
TCACATACATTCAATTATCAGCGGGTAGCGGATGAAAATCGTGCATTAAGAAAGAGATGTGGGGCAGTTACAGCAAACTGTGACA
AAGTTACAGTTAGAAAAGGAGCTTTGCAAGAGAGTTTGTGCTATTCAAGAAAAATTATAAGACTCTTATCGGCATTATGGGA
TGGTCAAAAAAATGGTTATGCTTAAGAT (SEQ ID NO: 72)

>Left Transposon End for WP_079907503.1

CTTACAATATGGTGGAACCTATAGCTAACTTAATCTTTGTTTTCGGTAAAATAAAAAATCAAGCAAATGGAAATTTAAAAATCA
AGCACAAAACCTGGAATAAAACATTTGCCGGCACCCCTCAAAAACCTGTACCCTTTAGTTGACTAGGCTAAAGGGGGTAAAGAAAG
G (SEQ ID NO: 73)

>Right Transposon End for WP_079907503.1

CAATTAAGACGGCCGCTTCATAAACAGAAGGAAAACCTTGATATTATATGGGCCGGTCCGGACTGGCAAAACACACTTGGCAA
CAGCCATAGGTGTAGAGGCATGTAGTAGAGCTAAGAAAGTTAAATTTTACAGAAGTGCAGCTCTGGTTAATCAGCTGCTGGAAA
CCAAAAGAACCGGGGAGCTGAAGAAATTTCTTAAACAACCTTGAGAATTTAGATCTGCTCATTTGCGATGAATGGGGCTATATTC
CTTTCGAAAAAGAGGTTCACAATTATTATTTCAAGTAATCTCGGAATGTTACGAGAAGAAAAGCCTGATTATTACTACCAACC
TTGAGTTTCAGTAAATGGAATAGCATATTTTATGATGAAAAGCTTACTAGTGCATTAATTGACAGAGTTATTTCATCATAGTCATC
TATTGATGTTTTACCGGAAGTAGTTATAGGCTTAGACACTCAACCATAAATTGCTAAATATGTAAGTTGGGGTGCTGGAAAAATA
AATTTCCAAATGCATGATTTTTTACTTTGCCAAACACACATCACCGGATCTGGGACTGCCATCATCCCTTTAAACATCTCCTAAT
TCCCATTGGGAATTTGAAACGGAAGTAAGAAAACCCGGCTTATGCCGGGTCTTTTTCTGTCGCCCGGCATGGGCGTTGTCTATAG
GGTGAAAGTCCCGAACGGCGAAGGTAGC (SEQ ID NO: 74)

>Left Transposon End for WP_013809817.1

CTTTTAATGTGCTAAAAAACGATTGAGCGCAGGCATTAATCATAACAATTACCTTTCCGGCTCATGCTGGTAATAATCCCATAT
TTCTTTAAGGCATCTTGGTAAGCATAGCTTGCATATTGACTCCCAGCATCGGAATGATGGATTAAACCTGGTTTTGGCCTTTGC
CGACTAACAGCGTGCTCCAGTGCGTCTATTACAAGCTGTTTTGTGTCATGGTACCATCCATAGACCAGCCAACAACCTTTTAAAGG
CATACATCAATTATGACTGCAAGGTAAAGCCAACCTTCAACGGTGGGGATATACGTAATATCTGCCGCCCATAGTTTATTAGGT
TCGGCAACATTGAAATTCTGGTCAACAATATTTGGGACAACATGGTAATTATGTTTTGAATTGGTGGTTGCCTTTAACTTTACA
GCTTCAAGTTTTAAATTCCTTATTGTACCGTTTTGGATTATCTGCCAAAAAAGACACCTCCGATGTTTGATATTTTATCTCAA
TTCGGTGTGTCCATCAAATCGGGTACATTACACAATCCCCTTGATTGTATTTGGCATAATAAAAAATATAGAGTTTGGCAACCAA
AACTTTAGAGTTTGATTGCCATCGGCAGTAATAAAAAAGACACTTGCCAACAGCATACTCTTACCCTACCGTTAAAGTTGCCA
CACAATGACGGTTAGGGGGAAAGAAAGG (SEQ ID NO: 75)
>Right Transposon End for WP_013809817.1
CATTATGTTGGCAAGTGATCTATTAAGAAATTGGCTGCCATTTTATTAATTTTTATCTTGCCAAACACA (SEQ ID NO:
76)
>Left Transposon End for WP_007056033.1
ACCACGCCCCGTATGTGTCTGTTCAATAGAAAGTGGGCCGTGTGAGCGTGTGGATGTCTCCATGTGAGCGGCAGGTTCTGGTCC
ACTTCAGCGTGTGGTGAGAAACGAACACCGGCCCTCCTGCCTGGACTAGTTTCCCGAGTTGAAAGTCAAACCGTTTCGGTTTTTC
TCGGGAAAGGAAAGGA (SEQ ID NO: 77)
>Right Transposon End for WP_007056033.1
CCAATGCAATAACCAACAACCAGCAGGACCGGCCAACGCCGAAACGGCCTGAAACCGCCGCTCACATGGACTAAAAGAAACGC
CGAAACGGCCTACTCCAATTGACAAAACACACCCGTACCCGAGTACTGGCGCATGGCC (SEQ ID NO: 78)
>Left Transposon End for CCU84826.1
AAAGGACAAATAAAAAATACATAAGAGTCATTGCCAGCGATCCGGGGAAAGAATACCCTTTAGTTGTGGCTAAACAAATAGAGGG
GAGATAAGGAGTGGATGCTAACA (SEQ ID NO: 79)
>Right Transposon End for CCU84826.1
ACAGTTAATCCCGGTTCTGGATTTTTATCTGGCCAAGTACTGTATTTCTAGTTGCCAAAAACAATC (SEQ ID NO: 80)
>Left Transposon End for WP_011639494.1
GGCCGTTCTTTTCATTGACGGTTTTACCGCTGGCGGTTCCGCCTGTTTATGGCACAATAAAAAATACAGCGAATGGCAAGTTAAAA
ATACAGCAGTAAATTGGAAAAAAATCATCGCCAGCACCCCCAAAATATACTACCCTTTTTGTTGAAGAGACAAAAGGAGGTAG
ATAAGAAAGG (SEQ ID NO: 81)
>Right Transposon End for WP_011639494.1
TGGAAACTGGGGTGCTGGATTTATTTTTTGCCGTTTGATGTATTTTATACTTGCCAAACACATCCGCCGAATAAAAAGCTGTAT
GTGCCCTCCATACAGACTATTTTTAGGACTGGCTCGGGGTTTTAACAAACCAACAAAATTGGTATTTTCATCTATCTCCCCCCC
TAAATGGATATAAAAAAACACCTGCCCATTTGAAAACAGATGCTTTTCTATTTCGCTTAGCTTTATG (SEQ ID NO: 82)
>Left Transposon End for ADQ03535.1
GAGTCCTTTTCGCGTCTCCGAGCTGTTTGTCTTCAATGTTTGTTCGACAGGTTCCGCGAGTGTTTTTCCGTAGTGTTGGGCTGATGG
TTTTCCCGGTTGTCGTCATCGTTTTCCCGGTGATGTGATAGGCATCGGTGCCGCCGCTGTATATTCCCGAGTTGCCAATCGGTT
TTGATTCCGCCGGTTGCCGCCGGCCGCGCTCGGGAAGGAAAGGAAC (SEQ ID NO: 83)
>Right Transposon End for ADQ03535.1
CCAACACAACGCAGACAGGCAGCGCCGATACCGCACACCCTGCGGAAAACCCGCTGCCTACACTGCGGAAAAACACCGCTCAC
AAAGCGGACGCCAACTTG (SEQ ID NO: 84)
>Left Transposon End for WP_011742685.1
CGGTCGCTTACATCGACGAATTCAAGGACCGTTTTAGGGTCGGGCCGATCCGCAGGGTGCTGGCCGCTTCGCTGGACTGCGGGT
TCATCACGCCGCGCGGCTACCGCATGTTTCAGATCCAGGTGTTTATGGTCAAATAAACGTGTACGGCGTCACCACCAATCTGAA
CGGCATAAGCGTTCATATTGGACACGGACGGCGGACGGAATGGCATCGGCGGAAGAAAAGCGCCGGTGCCGTACGGTTCCTGA
TAGATTCCCGAGTGCTAACAACCAACACATGGTCCTCGGGAGGAAAGGA (SEQ ID NO: 85)
>Right Transposon End for WP_011742685.1
ACGAAAAGAGCGGAACCAACCCGCCGCGGTGACGGTGTCCATTTGAACGCCGATGGTGTTTCAGATTAACTGGTGACGACAC
GCAGATTTATTTGACGAAAGACACAGGCCCGTCAGCCGC (SEQ ID NO: 86)
>Left Transposon End for WP_025263135.1
GAGGTCCTGCGCCTTGGCCGCGAAATCTGTTTATGTTCAATGGTTTAAAGACCGGTTGAGCGTTTGGTTTTTAGTCCGAAAGAG
CGGTAGGAATTAGTCCATTTTCATCGTTGATGGAAACAAGTGACACCGGTCTGTGTTATCGATTAGTTTCCCGAGTTGAAAAA
ACAACCATTGACGGTTCTCGGAAAGGAAAGGA (SEQ ID NO: 87)
>Right Transposon End for WP_025263135.1
TGCCCTTCCGGACCTTTCCCGAATGGAAGTGAAGACCATCAGGCTAAGGCCACGGGTGGTGGCATGCCTTTCCGCCCTAAA
CCCGAAAAACCTTCGTGACTAAACCTCAAAATCCTATTGAACAAAAACAAAACCAACACTGCCGATTACACTCACAATG
(SEQ ID NO: 88)
>Left Transposon End for WP_015439061.1

TCGTCCTCCTTGGTGCAGGCGATGTCGTATTCGCTCAGGTCCACGGGGTCGTCTAGGCCACGGGTCCCAGGCCCGAGGCGTGG
CGTGCCCGCGGCCAGACATACCCCGGCCCGGTCCGTGCCCTCCGGTGGATTTCGAGGATCTCGAGCATGCCGGCCACGGCCCGCCGC
CAGCCGCTCTCGCGCTCGGCGTGCAGCAGGGCACGCAGGCCGTGCGCCGCGTGGCCTCGTCCTGCGCGTCGATCCATTCGCGA
AGCGGGTCGGGCATCGCGTCGCGCACCCGTGTTTATGTTCAATGGTTTTAAGACCGGTTGAGCGTTTGGTTTTAGTCCGAAAG
AGCGGTAGGAATTAGTCCATTTATCGTTGATGGAAAACAAGTGACACCGGTCTGTCGTTATCGATTAGTTTCCCAGTTGAAA
AAACAACCATTTGACGGTTCTCGGAAAAGGAAAGGA (SEQ ID NO: 89)

>Right Transposon End for WP_015439061.1
AACAAATAAACACGAGCAATGACAGGCCAGTTCCGAACCTGGA (SEQ ID NO: 90)

>Left Transposon End for WP_082440433.1
CCGAGCACGTTGGTTCATTGCGCTTCAGCACCAGTGCCGATTGGGCCGTTATTGACACCCGGCACCAGAATCATTGCAATTCGAC
CCAGGTGTTTCGACACCGGTTTATTCAGGGGAACACTGCTCAACACCGGACAAACATTGAAGCAATACACTCCGGAGATGCGC
CCGTTTTCGGCTGCCATGTGCGCCAAATGGCGAT (SEQ ID NO: 91)

>Right Transposon End for WP_082440433.1
AACTCGCCGGCATAGGCAAGGCCAGGCTCCTGATCATCGACGAGTTCGGCTACCTGCCCATCGACGAGGAAGGCAGCCGCTCC
TCTTCAGATCATCAGCGACAGCTACGAGACAAGGAGCATCATCTACACCACCAACATCGAATTCAGCGGATGGGGACGCATCC
TCGGCGACAAGAATGCGCCGACGCCCTCATCGACCGCACCCTCCACCACGGACGACTCATCAGATTCGAGGGCGCTCCTACCG
CAGCCAACACGCCCTCATGACCAAATAACCAACAACAACAACACAGGCAGGCAGCGGCCGATACCGTACACCCCGCGGGA
AACCCGCTGCCACCCCTGCGGAAAACACCGCTCTCTTGTCTGGAGCTCCATATTGGGAACGATAAGAGTCGCGCAGCTTCTTCA
ATACATCGGTGCGGGTTTTACCGTATTAACGATGCACCCGACAACACCTCGCAGCTCCTTCTTCTGGCTTCTCCTCACT
(SEQ ID NO: 92)

>Left Transposon End for ADZ71261.1
AGCCGCGACATCCCGGCTGTGTTCAACAAGAGGCTGACGATTGCCCTCCTCGTCCCACTTGCCAGGGGGAGGCCCTGATTGGGA
AAATATAGGGTTGTCGACAAGCATTCCCCATGGTCTTCAATTTCGGCGGCAGGGCTCTTCTGATCAGAATAATGTAAAGCGAC
TCAAGAAAGCACATTCAGTGAGTTTGGAAATGCTTTCTTTCGAGGAATCTCGGGCTCTGCGTTCAAACTCGAACATAGAAGCCT
GGTGAGAGCGGCGGTGCAAAATTCGGCCACGGTAGCGGCGGGATAGTCTGCGCGGGCGGCGTAAAGTCGTCCACCTATTTT
CCTTCTGCAACATGTGCAGGAGGCTCAGGGGATCTACACC (SEQ ID NO: 93)

>Right Transposon End for ADZ71261.1
AGCCCCCTCAAAAGTCGCCAACGCCGACGCGGAAACTCTGGTCGGGCTACGCCCTCCCGACGTTCCCCGCGTGCGGCGTCAG
GTGGCCTACTTTTACGCCGCCAATGGACGACTTTTGCTCCGCCGTTGACACCTG (SEQ ID NO: 94)

>Left Transposon End for KYO67332.1
AAATTCTATTATAAGGTTGTTCCCTTTGGAATGTTTATTGCATAATTAAAATGGACAGAGTTGCAAAATAAAATCCCCACTTTTG
CAACTCAAAATCCCCAATATGCAAAGGTCATTGCAGGCACCCATAAAAAATATTACCCTTGATAGTTAAGGACAAAACTGCA
AGGGAGGATTGGAGG (SEQ ID NO: 95)

>Right Transposon End for KYO67332.1
CATAAAATTTCCCCACTGGTGCCTGGGGAAAAATCTTGCATATTGGGGAATTTACTCTTGCAAAAAACACTTTGGAACCTGA
AA (SEQ ID NO: 96)

>Left Transposon End for AEH46354.1
TGTTGGTGAGCATAAGTTAACAATTAGCGGTCTTAAAGACTATGCTGGTTTTCAAAATCGCTGACAAAGAATTTAGCTTTACTAT
TACAGCTGATACAACAGCACCAACTGTAACAGCAGCAAAAGTAGTAGATGCTAACACTGTTGAGGTAACATTCTCTGAGCCAGT
TGTC AACATTGGTACAATTACAATTGCTGGTACTCAATTTACTCAATCTGATGTTGGTGCTAATGCTGCTAAAACTGTATCAGC
AAACAAAGACAAAACCTGTTTATACATTTAAACACTGGATTAAATTAGCGGTCTTGGTTCACTAGAAATTGAGTTGAAATA
TCAAAACGTTACTGACGTTGAAGGTAACAAAAATACAACCGAACAAAAATTCAGTTTAAAGCTGAAGATGACACAACAGTACC
AACTGTAACATTAACTTAAAAACAGATAATACATTAGAAGTATTGTAATCTGTAATCAGTTTCTAATGCTGGTACTATTACTGT
AAAAGATTCAAAAGGTACAGTATTAGTTAATAAACTAGCATTATCAACTTACTATAATGTTTATGGCAAACTAAAAATGTAACG
TTTGGCAATTAATTTATGAGGTCTATAAAACAAAAACCACTTGCCAACATTCCCCAAATACACTACCCTCCTAGTTGGACA
AAACAGTTCAACGGGAGGTATTCAAGGA (SEQ ID NO: 97)

>Right Transposon End for AEH46354.1
CCATGGAATTTGGCAAGTGTATGTATGCATTTTAAATGTCATATTTTCTACTTGCCAAAAACATACTAAGTCAATG
ATATCAGGGGGATGATTCTCTCTGATTTATACATAATAATTGGTCGTAGAAATACGATTCTAGGAACAACAAGTTTAAATGGT
TTTATTTTGTTCGCATAATTATCAAGAACGACGGATTAGATTGTTTCGCATAATTTTTTTGAAGGAGAGCTTGTT (SEQ ID
NO: 98)

>Left Transposon End for WP_024833244.1
TGAACCATCCTTTTCATAAGCGATTAAACAGTCTCTTACTGTTGAGTAGTGCACTCCAATTTCTTTTGCCATTAAAGCTTGCTGT
TGTTTCACCTGCCAATATTTTTTCAACGGTCTGTTTCTTGAATGCTTTATCAAACCTTTTTTCTCTCAGCCATTTGAATCCCTCC
ATGTTTTTATTATATAGGGATTCTTGTGTCCATCAAACCTGGATTAAAGGTCACTCTTTTGAGTACATGGGAATTAGAAGGGTT
TCAATTCCTTATAGGTATTCTATAAACATGCATGACAATACTTATTGTTTGATAAAAAATCTATGTTTCAATTCCTTATAGGTA
TTCTATAAACTTGATAGTTTGTACATGTTTGTAGTGAATCTGAAAGTGTAACAAAGTGACGCCAAAAAGTGTAACGGTCTGCG
ATGTAAGTCCATTGTAACACCTTCAATTTGAGTTATCCTTTAGTTTGTCTGACGAACTGAAGGAGGTACAGAAAGGA

```

(SEQ ID NO: 99)
>Right Transposon End for WP_024833244.1
ATACAATTTGGGGTGTACACTTTTGCATTGCATTCGTTACAGAAAGCTATTGCAATAAACATGTTACATACTCTTTGTTATA
TTGA (SEQ ID NO: 100)

>Left Transposon End for WP_074011080.1
AGGCGTTGTATCTGTTTAGTGCAAAATAGAAATGGAGCAAAACGCAACTTAAAAATCCCGCATTGAGCTGGAAAACGAAAAGAC
CATTGCCAGCACCCCAATTAATTACTACCCTCTAATTGACGGCAATCAGTTAGAGGAGGAAAACACGAG (SEQ ID NO:
101)
>Right Transposon End for WP_074011080.1
ATTTGGAAACCGGGGTGCTGGAAATTTTAAATTGCACTTTGCTGGATTTT (SEQ ID NO: 102)

>Left Transposon End for ANW98740.1
AGGCACCACACCAAAATAATATCCTTGTGATTGATCGACCAATCAGTTACAAGGAGGATTGGAGG (SEQ ID NO: 103)
>Right Transposon End for ANW98740.1
AAACTTACGGGAAAGTCACTGGGTTTACATAAAATTTCCCCACCGGTGCCTGGGGAAAAAATTTTGCAAAAGTGGGGAGATCT
CACTTGCAAAAACATACAAATTATTACATGATTATTGATAAATATCATGCGCTCATATCACATGATTTTACAGAAGATAGCTC
TCAGTATAAAGAGTTTCTTAATTTTTTGAGCAGTAAATTAAAAAATTTAAGGGTTATGAAATTATTAAGAGAGAAATGAAAA
AACAGACATATTAAACACATTAGATTTGCTTAAAGAAAAAGATATAGATTTTGATACAAAAAATTTATCTTAAGGAAATTGT
GGAGTGCCTTGAAAAGCACAAAGACATTTTCTTATTAAGATATAGTTAATACAAAAGTAAACTATTTTGGGATAATATTGC
TTTTCTTAATAATAACAATAAAATAAAGAAATGTCGATTGTTATAAAAATACATTTATTGACCCCGTAATACAATATCTTAA
TAAAAACAGAGGAAGAGTGCCCAATCAATGCATTGAATGTGGCAGGAATGTTTCTTCCCAAAAGAAAGTTTCGATGAGATGGTT
AAATGACGTAGGAGTAGATATAAAAGGAAAATCTAACTACTGGAATTTTAAATCCGGATATTGTGGTGTGCTATATGTAA
AACAGTTTATTCGTGCGCTCCATTAGGT (SEQ ID NO: 104)

```

Example 2 – Design of Engineered System for a CLUST.009467 CRISPR System (FIG. 8)

Having identified the minimal CLUST.009467 CRISPR-Cas system components, we

5 composed an engineered system for transposon excision, mobilization, and programmable insertion (**FIG. 8**). Minimally, the natural locus consists of a transposon comprising left and right Tnp ends closely flanking effectors A – B, and a CRISPR array located immediately adjacent to the transposon. In some embodiments, the Cas Effector Cassette of an engineered CLUST.009467 system is composed of effectors A and B under the control of a single artificial

10 promoter. In some embodiments, the Transposon Payload Cassette of an engineered CLUST.009467 system is composed of Tnp ends flanking a nucleic acid cargo ranging from approximately less than 100nt to greater than 25kb (**FIG. 8**). In some embodiments, the RNA Guide Cassette of an engineered CLUST.009467 system is expressed under the control of a single artificial promoter. In some embodiments, the RNA guide may consist of a minimal

15 CRISPR array containing two or more direct repeats and one or more spacers. In some embodiments, the RNA guide may consist of one or more crRNAs containing processed direct repeat and spacer components, and optionally a fused tracrRNA component. In some embodiments, the RNA Guide Cassette contains a tracrRNA expressed under a second artificial promoter.

We selected the BDLR01000066 and CP002736 loci for functional validation of the engineered CLUST.009467 system.

5 *DNA synthesis & effector library cloning*

To test the activity of BDLR01000066 and CP002736 CRISPR-Cas systems, we designed and synthesized a minimal engineered system as described above into the pET28a(+) vector. The synthesized system consisting of an Effector Cassette and acceptor site for an RNA Guide Cassette were included as the cargo region of a Transposon Payload Cassette. The
10 synthesized combined IPTG inducible T7 and lac promoters were used to drive expression of the Effector Cassette, and each CDS sequence was codon optimized for *E. coli* expression and preceded by an *E. coli* ribosome binding sequence. A J23119 (Registry of Standard Biological Parts: http://parts.igem.org/Part:BBa_J23119) promoter was used for expression of the RNA Guide Cassette.

15 In tandem with the effector gene synthesis, we first computationally designed an oligonucleotide library synthesis (OLS) pool containing “repeat-spacer-repeat” sequences, where “repeat” represents the consensus direct repeat sequence found in the CRISPR array associated with the effector, and “spacer” represents sequences tiling the pACYC184 plasmid and *E. coli* essential genes. The spacer length was determined by the mode of the spacer lengths found in the
20 endogenous CRISPR array. The repeat-spacer-repeat sequence was appended with restriction sites enabling the bi-directional cloning of the fragment into the acceptor site for the RNA Guide Cassette, as well as unique PCR priming sites to enable specific amplification of a specific repeat-spacer-repeat library from a larger pool. The library synthesis was performed by Agilent Genomics.

25 We next cloned the CRISPR array library into the plasmid containing the minimal engineered BDLR01000066 or CP002736 CLUST.009467 system using the Golden Gate assembly method. In brief, we first amplified each repeat-spacer-repeat from the CRISPR array library using unique PCR primers, and pre-linearized the plasmid backbone using BsaI to reduce potential background. Both DNA fragments were purified with Ampure XP (Beckman Coulter)
30 prior to addition to Golden Gate Assembly Master Mix (New England Biolabs) and incubated as per manufacturer’s instructions. We further purified and concentrated the Golden Gate reaction

to enable maximum transformation efficiency in the subsequent steps of the bacterial screen.

Example 3 – Functional Screening of a CLUST.009467 CRISPR System (FIGs. 9 – 23)

Functional screening for CLUST.009467

5 To accelerate functional screening of the BDLR01000066 and CP002736 CLUST.009467 systems, we developed a strategy to derive the following functional information in a single screen: 1) crRNA expression direction and processing, 2) nucleic acid substrate type, and 3) targeting requirements such as protospacer adjacent motif (PAM), protospacer flanking sequence (PFS), or target secondary structure. We designed minimal CRISPR array libraries
10 consisting of two consensus direct repeats, each flanking a unique natural-length spacer sequence targeting either the pACYC184 vector, *E. coli* essential genes, or an absent GFP sequence as a negative control. We also designed a bidirectional array library cloning strategy to test both possible CRISPR array expression directions in parallel.

The respective CRISPR array libraries were cloned into the RNA Guide Cassette
15 acceptor sites of the BDLR01000066 or CP002736 CLUST.009467 CRISPR-Cas system expression plasmids such that each element in the resulting plasmid library contained a single RNA guide element expressed in either the forward or reverse orientation. The resulting plasmid libraries were transformed with pACYC184 into *E. coli* using electroporation, yielding a maximum of one plasmid library element per cell. Transformed *E. coli* cells were plated on
20 bioassay plates containing Kanamycin (selecting for the library plasmid), Chloramphenicol (CAM; selecting for intact pACYC184 CAM expression), and Tetracycline (TET; selecting for intact pACYC184 TET expression).

Programmable mobilization and insertion of the BDLR01000066 or CP002736 CLUST.009467 Transposon Payload Cassette was assessed as follows. Following
25 transformation, the CLUST.009467 system mobilizes the Transposon Payload Cassette by excising it from the donor plasmid followed by programmable insertion of the mobilized cassette at a target site specified by the spacer sequence of the expressed RNA guide (**FIGs. 9-10**). Interruption of pACYC184 antibiotic resistance genes or *E. coli* essential genes by targeted insertion of the Transposon Payload Cassette by the CLUST.009467 CRISPR-Cas system results
30 in bacterial cell death and depletion of the targeting RNA guide (**FIG. 11**). To investigate negative selection of RNA guides resulting from programmable transposon insertion, bacteria

were harvested 12h after plating, and plasmid DNA was extracted. We PCR amplified the CRISPR array region of the input plasmid library prior to transformation and the output plasmid library after bacterial selection on antibiotic plates and compared the frequency of RNA guide elements in these samples to identify depleted elements (**FIG. 11**).

5

Bacterial screen transformation

The plasmid library containing the distinct repeat-spacer-repeat elements and Cas proteins was electroporated into Endura or E. cloni electrocompetent *E. coli* (Lucigen) using a Gene Pulser Xcell (Bio-rad) following the protocol recommended by Lucigen. The library was either co-transformed with purified pACYC184 plasmid, or directly transformed into
10 pACYC184-containing Endura or E. cloni electrocompetent *E. coli* (Lucigen), plated onto agar containing Chloramphenicol (Fisher), Tetracycline (Alfa Aesar), and Kanamycin (Alfa Aesar) in BioAssay dishes (Thermo Fisher), and incubated for 10-12h. After estimation of approximate colony count to ensure sufficient library representation on the bacterial plate, the bacteria were
15 harvested and DNA plasmid extracted using a QIAprep Spin Miniprep Kit (Qiagen) to create the 'output library'. By performing a PCR using custom primers containing barcodes and sites compatible with Illumina sequencing chemistry, we generated a barcoded next generation sequencing library from both the pre-transformation 'input library' and the post-harvest 'output library', which were then pooled and loaded onto a Nextseq 550 (Illumina) to evaluate the
20 effectors. At least two independent biological replicates were performed for each screen to ensure consistency.

Bacterial screen sequencing analysis

Next generation sequencing data for screen input and output libraries were demultiplexed
25 using Illumina bcl2fastq. Reads in resulting fastq files for each sample contained the CRISPR array elements for the screening plasmid library. The direct repeat sequence of the CRISPR array was used to determine the array orientation, and the spacer sequence was mapped to the source plasmid pACYC184 or negative control sequence (GFP) to determine the corresponding target. For each sample, the total number of reads for each unique array element (r_a) in a given plasmid
30 library was counted and normalized as follows: $(r_a+1) / \text{total reads for all library array elements}$. The depletion score was calculated by dividing normalized output reads for a given array

element by normalized input reads.

To identify specific parameters resulting in enzymatic activity and bacterial cell death, we used next generation sequencing (NGS) to quantify and compare the representation of individual CRISPR arrays (i.e., repeat-spacer-repeat) in the PCR of the input and output plasmid libraries. We define the array depletion ratio as the normalized output read count divided by the normalized input read count. An array was considered to be strongly depleted if the fold depletion was at least than 3. When calculating the array depletion ratio across biological replicates, we took the maximum depletion ratio value for a given CRISPR array across all experiments (i.e. a strongly depleted array must be strongly depleted in all biological replicates). We generated a matrix including array depletion ratios and the following features for each spacer target: target strand, transcript targeting, ORI targeting, target sequence motifs, flanking sequence motifs, and target secondary structure. We investigated the degree to which different features in this matrix explained target depletion for CLUST.009467 systems, thereby yielding a broad survey of functional parameters within a single screen.

Bacterial screening indicates programmable DNA interference by the BDLR01000066 CLUST.009467 system

Comparison of two bioreplicate screens for the BDLR01000066 CLUST.009467 CRISPR-Cas system showed replication of depleted RNA guide elements for a single direct repeat expression direction (5' CTTT...GAAA - [spacer]...3'). This suggests a specific interaction of the BDLR01000066 CLUST.009467 effector proteins with a specific direct repeat orientation, and accompanying modification of targeted pACYC and *E. coli* essential gene target sequences (**FIGs. 12-13**). Targeting of both strands of pACYC and *E. coli* essential genes indicates DNA targeting is not confined to a specific strand orientation (**FIGs. 14A-B**). Additionally, localization of depleted targets to sequences either within or in close proximity to gene coding sequences is indicative of coding DNA interruption associated interference consistent with the interference arising from mobilization and programmable insertion of the Transposon Payload Cassette at or in close proximity to target sites (**FIGs. 14A-B**). In an attempt to understand targeting requirements for the CLUST.009467 system, we compiled all target and target-flanking sequences associated with depleted RNA guide elements (**FIGs. 15A-C**). Flanking sequences showed no evidence of PAM or PFS elements required for targeting, and

target sequences showed no evidence of sequence bias (**FIGs. 15A-C**). To confirm that BDLR01000066 CLUST.009467 effector A is involved in the depletion phenotype, we deleted effector A from the CLUST.009467 expression plasmid and observed that programmable DNA interference was ablated (**FIGs. 16-17**) and there was absent a preferred direct repeat orientation as observed in FIG. 12, when the BDLR01000066 CLUST.009467 effector A was present. These results indicate that the synthetic BDLR01000066 CLUST.009467 system is capable of interference and that effector A is a necessary component for such activity.

Bacterial screening indicates programmable DNA interference by the CP002736 CLUST.009467 system

As another member of the CLUST.009467 family that displays putative target-dependent interference activity, the CP002736 CLUST.009467 CRISPR-Cas system showed depleted RNA guide elements for a single direct repeat expression direction (5' GTTT...TAAC - [spacer]...3') across two separate bioreplicates. The preference of a particular direct repeat orientation suggests an interaction of the CP002736 CLUST.009467 effector proteins with a specific direct repeat orientation, and accompanying modification of targeted pACYC and *E. coli* essential gene target sequences (**FIGs. 18-19**). Targeting of both strands of pACYC and *E. coli* essential genes indicates DNA targeting is not confined to a specific strand orientation (**FIGs. 20A-B**). Additionally, localization of depleted targets to sequences either within or in close proximity to gene coding sequences is indicative of coding DNA interruption associated interference consistent with the interference arising from mobilization and programmable insertion of the Transposon Payload Cassette at or in close proximity to target sites (**FIGs. 20A-B**). In an attempt to understand targeting requirements for the CLUST.009467 system, we compiled all target and target-flanking sequences associated with depleted RNA guide elements (**FIGs. 21A-C**). Flanking sequences showed no evidence of PAM or PFS elements required for targeting, and target sequences showed no evidence of sequence bias (**FIGs. 21A-C**). To confirm the necessity of the CP002736 CLUST.009467 effector A for observation of the depletion phenotype, we deleted effector A from the CLUST.009467 expression plasmid and performed the screening again, observing that programmable DNA interference was significantly lowered relative to the full synthetic locus (**FIGs. 22-23**).

OTHER EMBODIMENTS

It is to be understood that while the invention has been described in conjunction with the detailed description thereof, the foregoing description is intended to illustrate and not limit the scope of the invention, which is defined by the scope of the appended claims. Other aspects, advantages, and
5 modifications are within the scope of the following claims.

What is claimed is:

1. An engineered, non-naturally occurring Clustered Interspaced Short Palindromic Repeat (CRISPR) - Cas system of CLUST.009467 comprising:
 - a Guide consisting of a direct repeat sequence and a spacer sequence capable of hybridizing to a target nucleic acid, wherein the Guide comprises a CRISPR RNA or DNA, and any one of the following:
 - a. a CRISPR-associated protein containing both an HTH domain and an rve integrase domain capable of binding to the Guide, either as a monomer or multimer, and of targeting the target nucleic acid sequence complementary to the Guide spacer;
 - b. an IstB domain-containing protein; and
 - c. a payload nucleic acid flanked by transposon end sequences.
2. The system of claim 1, wherein the target nucleic acid is a DNA or an RNA.
3. The system of claim 2, wherein the target nucleic acid is double-stranded DNA.
4. The system of any one of claims 1-3, wherein targeting of the target nucleic acid by the protein and the Guide results in a modification in the target nucleic acid.
5. The system of claim 4, wherein the modification in the target nucleic acid is a double-stranded cleavage event.
6. The system of claim 4, wherein the modification in the target nucleic acid is a single-stranded cleavage event.
7. The system of any one of claims 1-6, further comprising a donor template nucleic acid.
8. The system of claim 7, wherein the donor template nucleic acid is a DNA.
9. The system of any one of claims 1-8, wherein the target nucleic acid is a double-stranded

DNA and the targeting of the double-stranded DNA results in scarless DNA insertion.

10. The system of any one of claims 4-8, wherein the modification results in cell toxicity.
11. The system of any one of claims 1 to 10, within a cell.
12. The system of any one of claims 1-11, wherein the cell comprises a eukaryotic cell.
13. The system any one of claims 1-11, wherein the cell comprises a prokaryotic cell.
14. A method of targeting and editing a target nucleic acid, the method comprising contacting the target nucleic acid with a system of any one of claims 1 to 13.
15. A method of targeting the insertion of a payload nucleic acid at a site of a target nucleic acid, the method comprising contacting the target nucleic acid with a system of any one of claims 1 to 13.
16. A method of targeting the excision of a payload nucleic acid from a site at a target nucleic acid, the method comprising contacting the target nucleic acid with a system of any one of claims 1 to 13.
17. The system of any one of claims 1-13, wherein the CRISPR-associated protein comprises an amino acid sequence having at least 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% similarity to an amino acid sequence provided in any one of Tables 2 – 3.
18. The system of any one of claims 1-13, wherein the crRNA or Guide comprises a nucleic acid sequence having at least 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% similarity to a nucleic acid sequence provided in Table 4.
19. The system of claims 1-13, wherein the payload nucleic acid is flanked by transposon end sequences comprising a nucleic acid sequence having at least 85%, 90%, 95%, 96%, 97%, 98%,

99%, or 100% similarity to a nucleic acid sequence contained in Table 5.

20. The system of any one of claims 1-13, wherein the CRISPR-associated protein comprises at least one nuclear localization signal.

21. The system of any one of claims 1-13, wherein the CRISPR-associated protein comprises at least one nuclear export signal.

22. The system of any one of claims 1-13, wherein at least one component of the system is encoded by a codon-optimized nucleic acid for expression in a cell.

23. The system of claim 22, wherein the codon-optimized nucleic acid is present within at least one vector.

24. The system of claim 23, wherein the at least one vector comprise one or more regulatory elements operably-linked to a nucleic acid encoding the component of the system.

25. The system of claim 24, wherein the one or more regulatory elements comprise at least one promoter.

26. The system of claim 25, wherein the at least one promoter comprises an inducible promoter or a constitutive promoter.

27. The system of claim 23, wherein the at least one vector comprises a plurality of vectors

28. The system of claim 23, wherein the at least one vector comprises a viral vector.

29. The system of claim 28, wherein the viral vector is selected from the group consisting of a retroviral vector, a lentiviral vector, an adenoviral vector, an adeno-associated vector, and a herpes simplex vector.

30. The system of any one of claims 1-13, wherein the system is present in a delivery system.
31. The system of claim 30, wherein the delivery system comprises a delivery vehicle selected from the group consisting of a liposome, an exosome, a microvesicle, and a gene-gun.
32. A cell comprising the system of any one of claims 1-13.
33. The cell of claim 32, wherein the cell is a eukaryotic cell.
34. The cell of claim 33, wherein the eukaryotic cell is a mammalian cell or a plant cell.
35. The cell of claim 34, wherein the mammalian cell is a human cell.
36. The cell of claim 33 wherein the cell is a prokaryotic cell.
37. The method of claim 14, wherein the method results in an insertion or substitution of DNA to correct a native locus.
38. A method of claim 15, wherein the method results in a targeted insertion of a DNA payload into a specific genomic target site.
39. A method of claim 16, wherein the method results in a targeted deletion of DNA to correct a native locus.

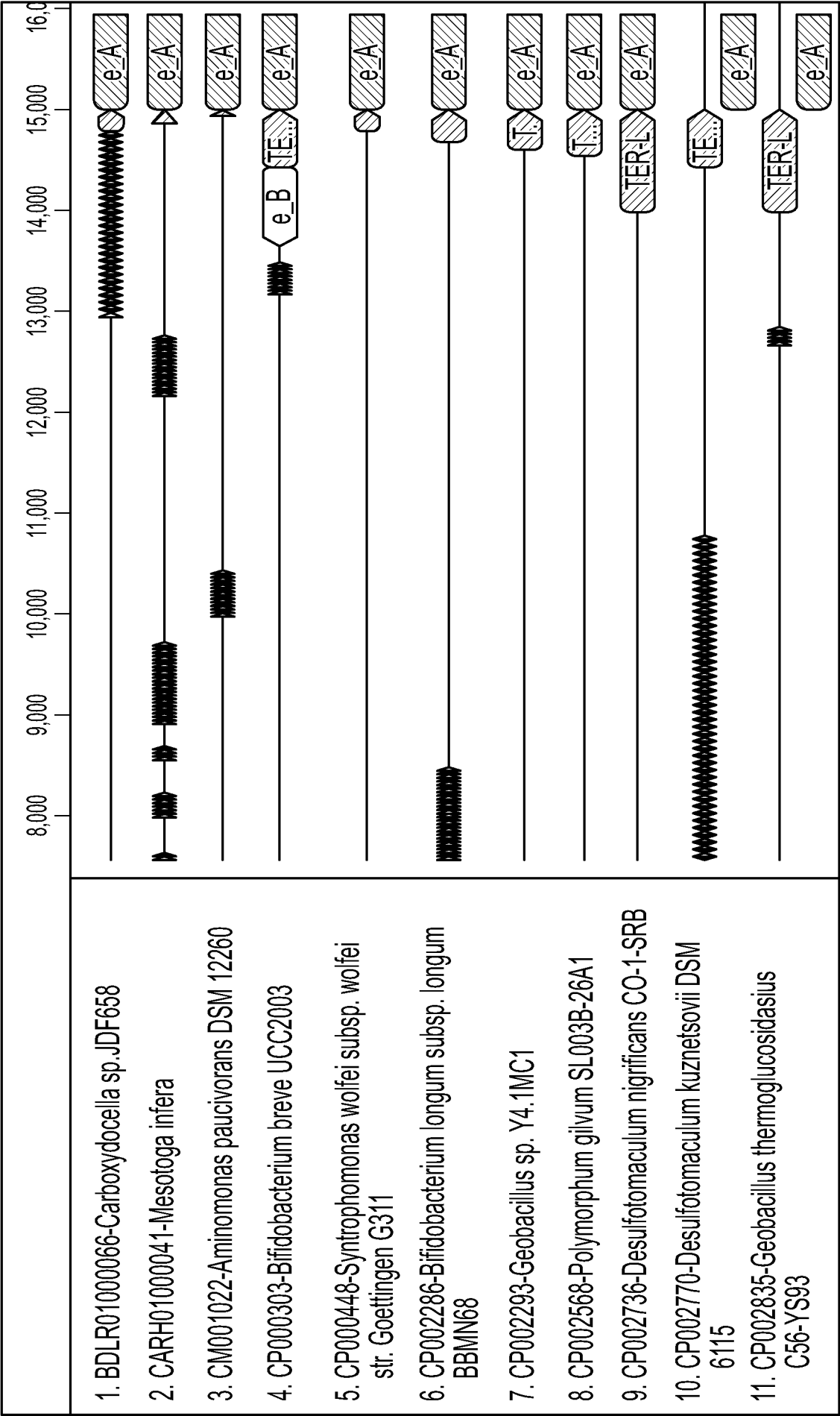


FIG. 1A

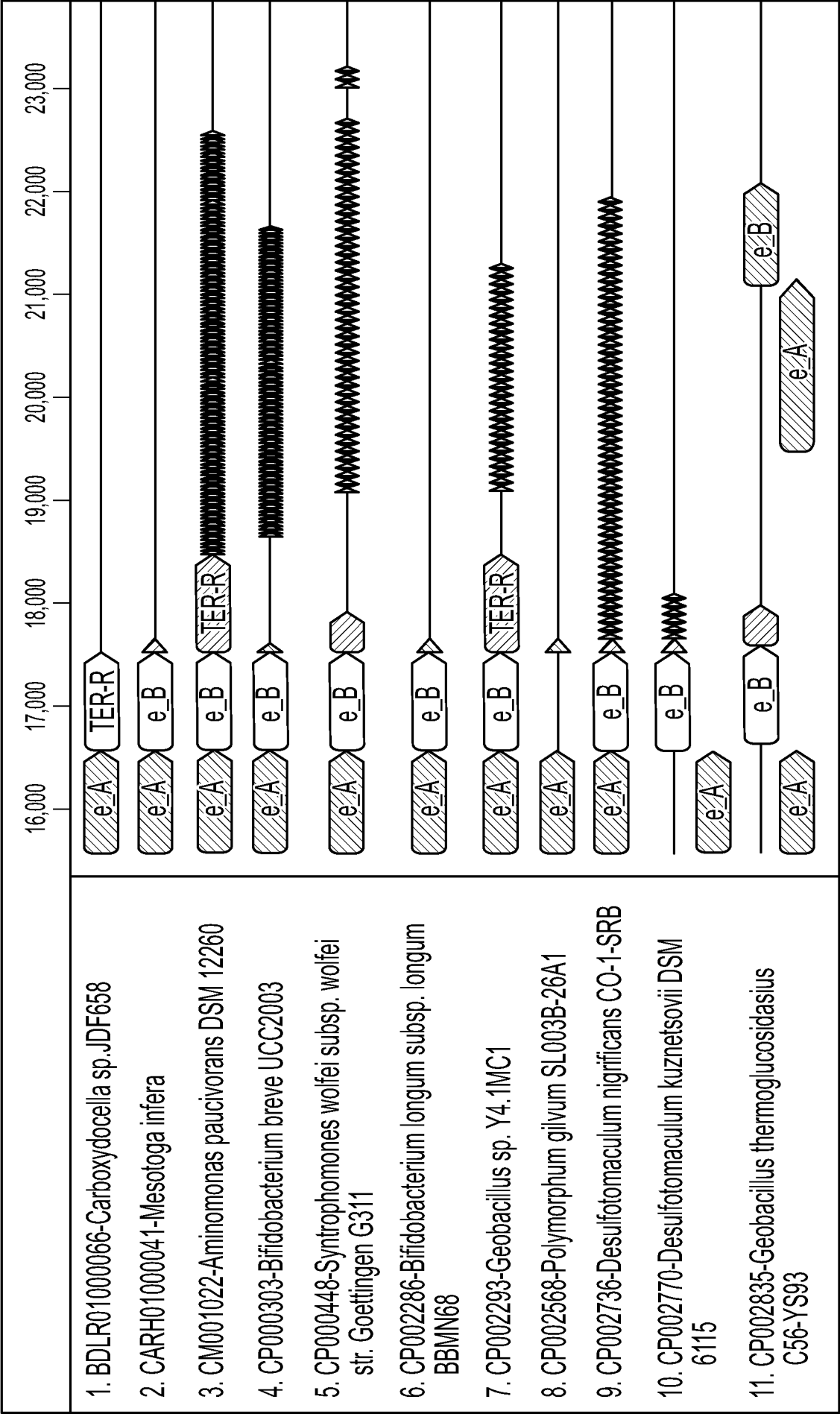


FIG. 1B

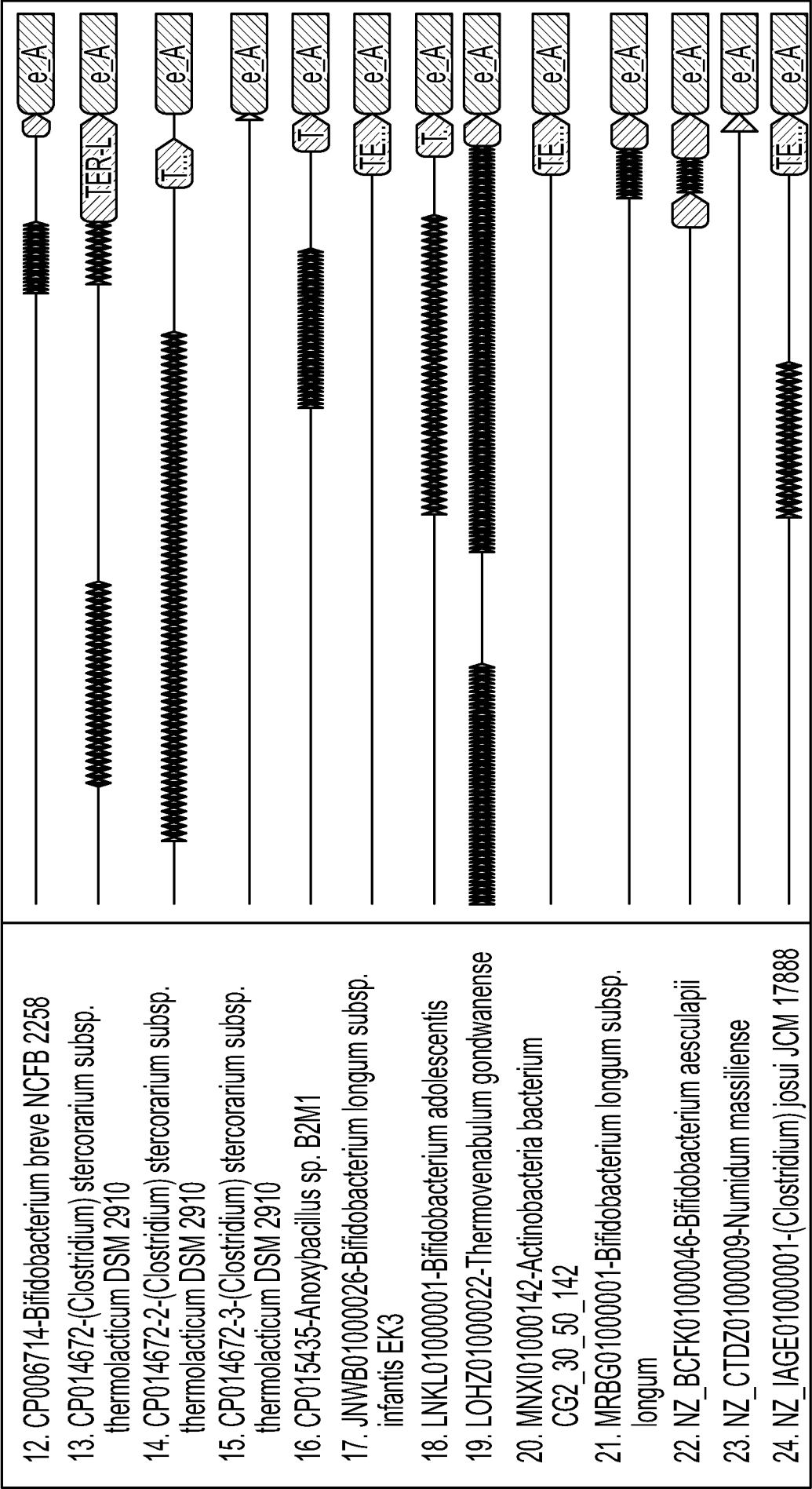


FIG. 1C

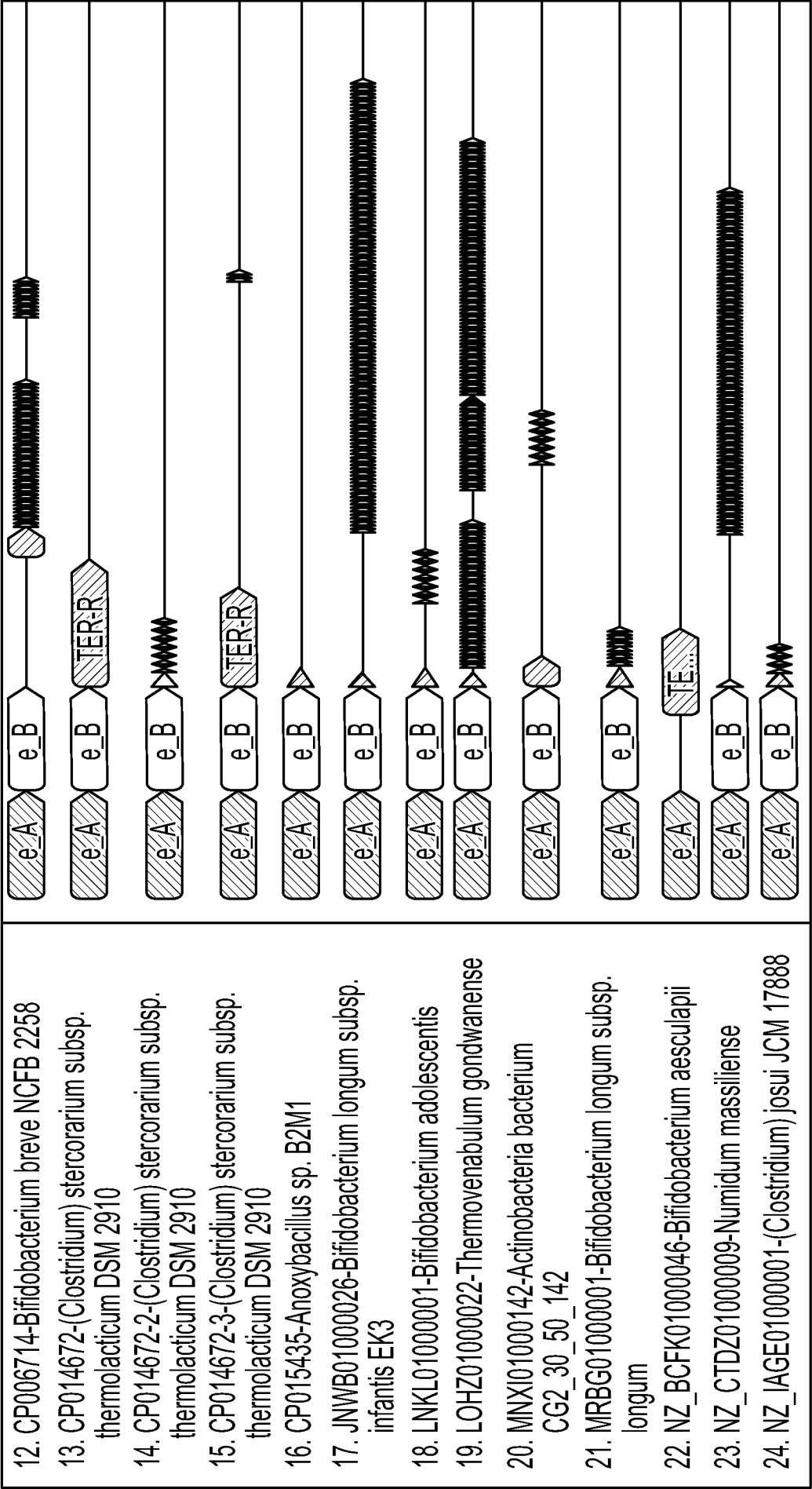


FIG. 1D

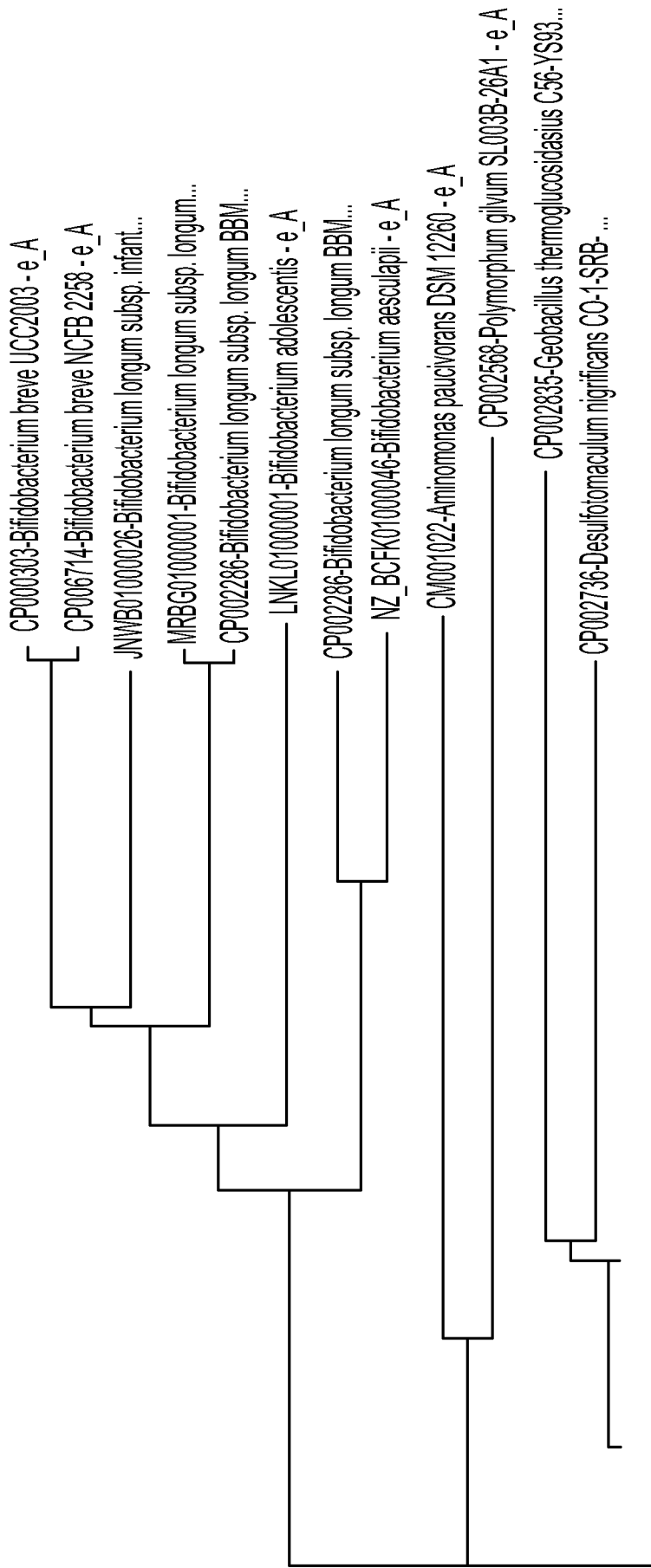


FIG. 2A

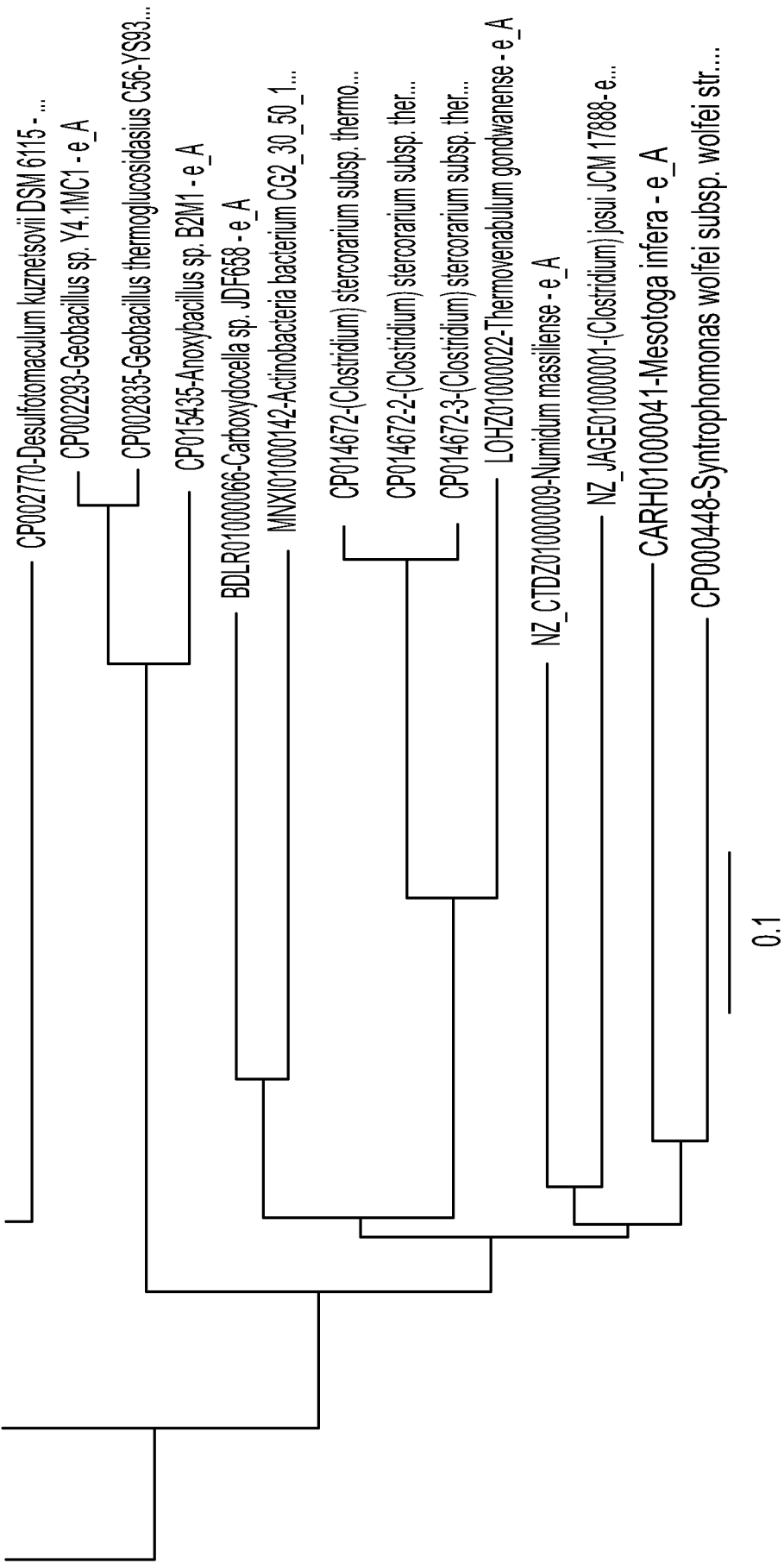


FIG. 2B

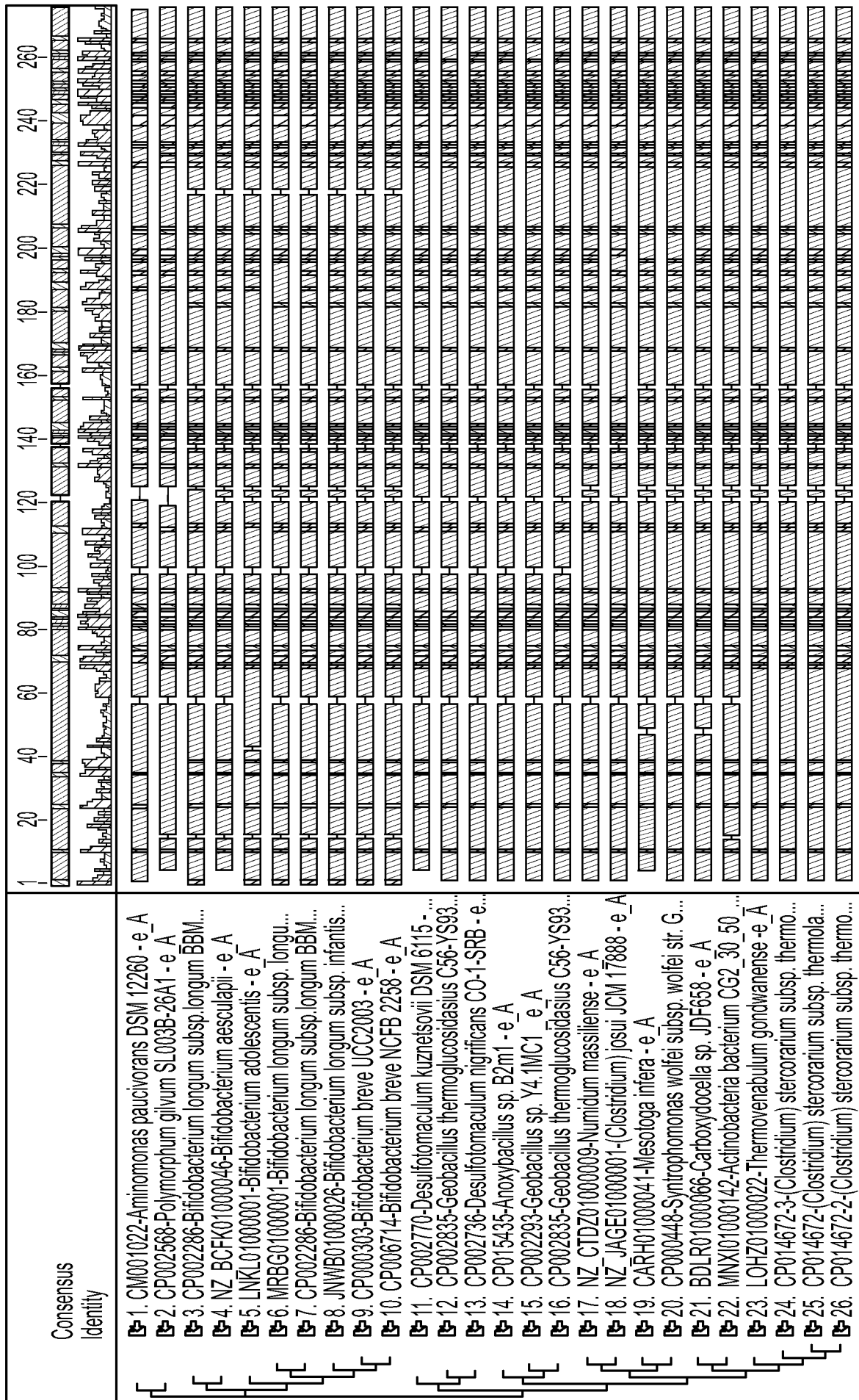


FIG. 3A

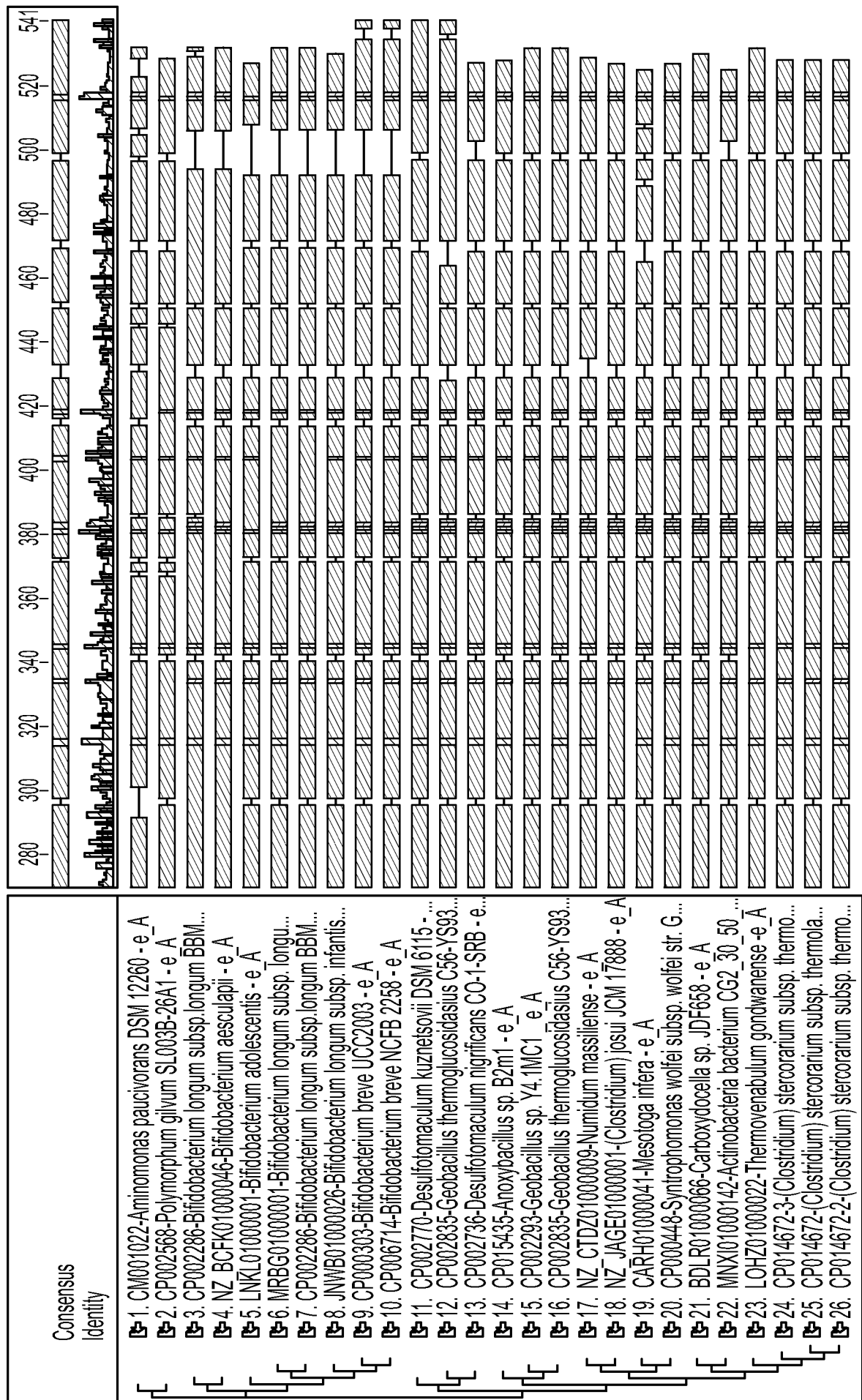


FIG. 3B

Scores for complete sequences (score includes all domains) :

--- full sequence --- --- best 1 domain --- --- #dom-

E-value	score	bias	E-value	score	bias	exp	N	Sequence	Description
-----	-----	-----	-----	-----	-----	-----	---	-----	-----
3.6e-66	234.2	0.2	3.9e-66	234.1	0.2	1.0	1	A0A010RDY6_PSEFL/125-253	A0A010RDY6.1 PF00665.25; rve;
1.6e-64	228.7	0.4	1.8e-64	228.6	0.4	1.0	1	A4VH97_PSEU5/113-241	A4VH97.1 PF00665.25; rve;
2.5e-63	224.8	0.7	2.8e-63	224.7	0.7	1.0	1	A0A109BGE0_9BURK/128-257	A0A109BGE0.1 PF00665.25; rve;
3.4e-63	224.4	0.7	3.8e-63	224.2	0.7	1.0	1	A1W408_ACISJ/124-252	A1W408.1 PF00665.25; rve;
9.8e-63	222.8	2.7	1.1e-62	222.7	2.7	1.0	1	A0A086PBW6_SPHHM/96-225	A0A086PBW6.1 PF00665.25; rve;
1.8e-62	222.0	0.8	2e-62	221.8	0.8	1.0	1	A0A167WS55_9BURK/124-251	A0A167WS55.1 PF00665.25; rve;
•••									
1.6e-09	47.3	1.1	1.6e-09	47.3	1.1	1.5	1	U9W0R2_9CYAN/84-324	U9W0R2.1 PF01610.16; DDE_Tnp_ISL3;
4.6e-09	45.8	0.0	4.6e-09	45.8	0.0	1.0	1	I3TIQ8_TISMK/4-42	I3TIQ8.1 PF13384.5; HTH_23;
1.2e-08	44.4	0.1	1.2e-08	44.4	0.1	1.0	1	A0A011TUV0_9RHIZ/2-41	A0A011TUV0.1 PF13412.5; HTH_24;
1.8e-08	43.9	1.1	2.4e-08	43.4	0.2	1.6	1	A0A0D2VNF9_9PROT/62-300	A0A0D2VNF9.1 PF01610.16; DDE_Tnp_ISL3;
3.9e-08	42.7	0.0	5.4e-08	42.3	0.0	1.2	1	D3DBW3_9ACTN/114-316	D3DBW3.1 PF06527.10; TniQ;
4e-08	42.7	0.1	4e-08	42.7	0.1	1.0	1	A0A0D7K479_9BURK/7-46	A0A0D7K479.1 PF13384.5; HTH_23;
6e-08	42.1	0.2	6e-08	42.1	0.2	1.0	1	A0A0A1BBC9_9GAMM/2-42	A0A0A1BBC9.1 PF13412.5; HTH_24;
6.6e-08	42.0	0.2	6.7e-08	42.0	0.2	1.0	1	A2SFT7_METPP/3-41	A2SFT7.1 PF13384.5; HTH_23;
7.5e-08	41.8	0.3	1.3e-07	41.0	0.3	1.3	1	G8B102_AZ0BR/129-375	G8B102.1 PF01610.16; DDE_Tnp_ISL3;
9.6e-08	41.4	0.5	1.4e-07	40.9	0.5	1.3	1	A0A0D2WDK5_9PROT/12-250	A0A0D2WDK5.1 PF01610.16; DDE_Tnp_ISL3;
1.2e-07	41.1	0.4	1.2e-07	41.1	0.4	1.0	1	J2M308_9BURK/25-68	J2M308.1 PF13384.5; HTH_23
2.5e-07	40.1	0.2	2.5e-07	40.1	0.2	1.0	1	W0TU24_9GAMM/5-45	W0TU24.1 PF11427.7; HTH_Tnp_Ic3_1;

FIG. 4

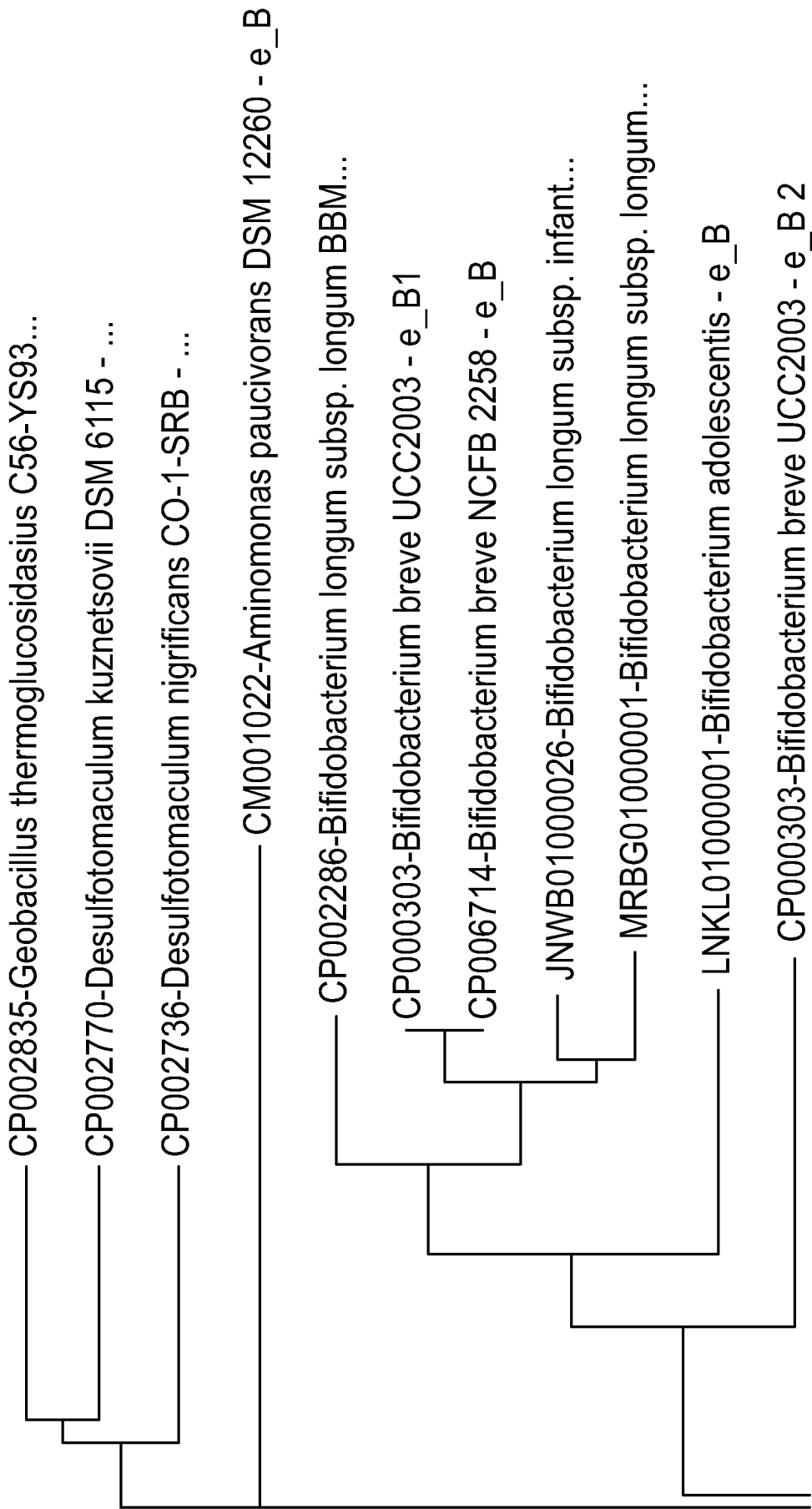


FIG. 5A

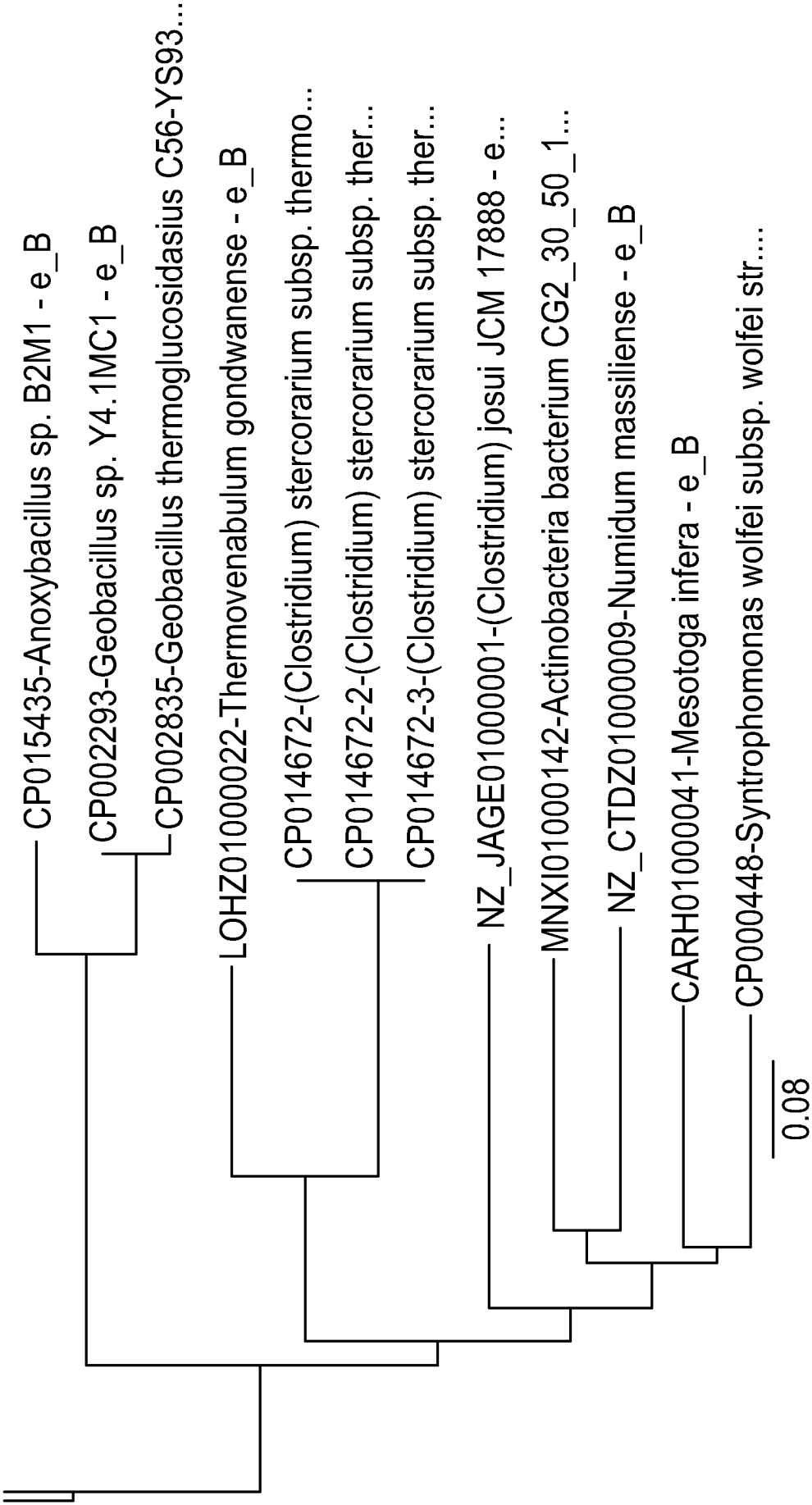


FIG. 5B

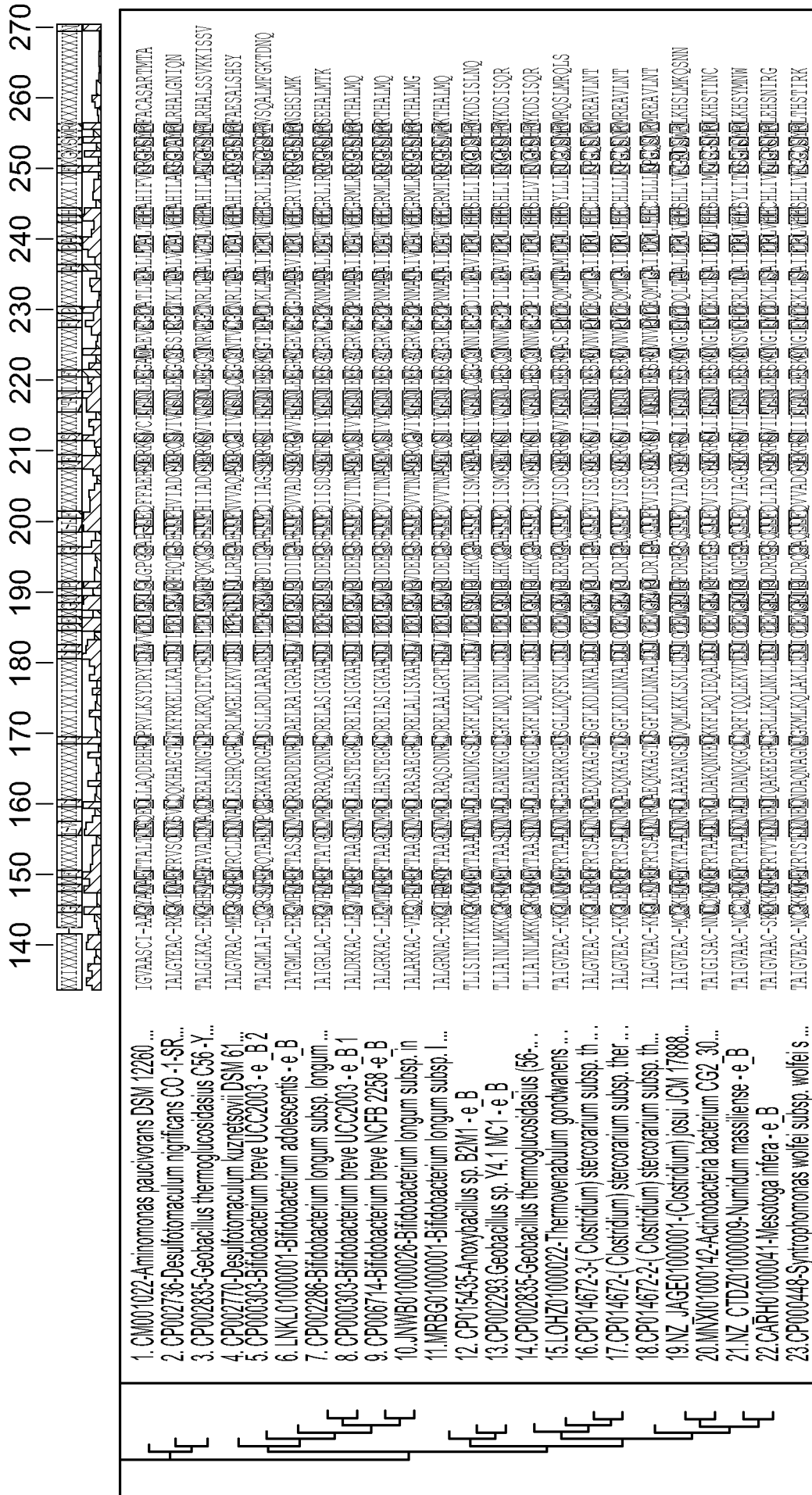


FIG. 6B

Scores for complete sequences (score includes all domains) :

--- full sequence ---	--- best 1 domain ---	#dom-	E-value	score	bias	exp	N	Sequence	Description
-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
1.5e-78	273.5	0.1	1.7e-78	273.4	0.1	1.0	1	A0A135HVB3_9RHIZ/57-234	A0A135HVB3.1 PF01695.16;IstB_I S21;
2.1e-78	273.0	0.1	2.3e-78	272.9	0.1	1.0	1	N6UZV1_9RHIZ/57-234	N6UZV1.1 PF01695.16;IstB_I S21;
2.3e-78	272.9	0.1	2.6e-78	272.7	0.1	1.0	1	M2S7P6_9PROT/57-234	M2S7P6.1 PF01695.16;IstB_I S21;
2.8e-78	272.6	0.3	3.1e-78	272.5	0.3	1.0	1	Q11ZW4_POLSJ/55-232	Q11ZW4.1 PF01695.16;IstB_I S21;
7.1e-78	271.3	0.3	7.9e-78	271.2	0.3	1.0	1	C4KBJ2_THASP/51-228	C4KBJ2.1 PF01695.16;IstB_I S21;
7.3e-78	271.3	0.1	8.1e-78	271.1	0.1	1.0	1	Q1N7F7_SPHSS/45-222	Q1N7F7.1 PF01695.16;IstB_I S21;
8.3e-78	271.1	0.0	9.2e-78	270.9	0.0	1.0	1	Q1GTQ7_SPHAL/57-234	Q1GTQ7.1 PF01695.16;IstB_I S21;
•••									
2.2e-17	73.1	0.0	2.7e-17	72.9	0.0	1.1	1	DNAA_GEOMG/115-333	Q39ZS3.1 PF00308.17;Bac_DnaA;
3.2e-17	72.6	0.0	4.6e-17	72.1	0.0	1.1	1	A0A0D91AL52_9HEIL/105-321	A0A0D91AL52.1 PF00308.17;Bac_DnaA;
4.1e-17	72.3	0.0	5.4e-17	71.9	0.0	1.1	1	A0A0F9ZZ92_9BACT/180-403	A0A0F9ZZ92.1 PF00308.17;Bac_DnaA;
4.8e-17	72.0	0.1	6.7e-17	71.5	0.1	1.1	1	A6Q664_SULNB/104-320	A6Q664.1 PF00308.17;Bac_DnaA;
5e-17	72.0	0.1	6.2e-17	71.7	0.1	1.1	1	A0A0C2Z1L8_9PROT/105-321	A0A0C2Z1L8.1 PF00308.17;Bac_DnaA;
5.1e-17	71.9	0.0	7e-17	71.5	0.0	1.1	1	A0A0C2YSS0_9PROT/103-319	A0A0C2YSS0.1 PF00308.17;Bac_DnaA;
6.8e-17	71.5	0.0	9.8e-17	71.0	0.0	1.1	1	A0A0G0QQI0_9BACT/119-336	A0A0G0QQI0.1 PF00308.17;Bac_DnaA;

FIG. 7

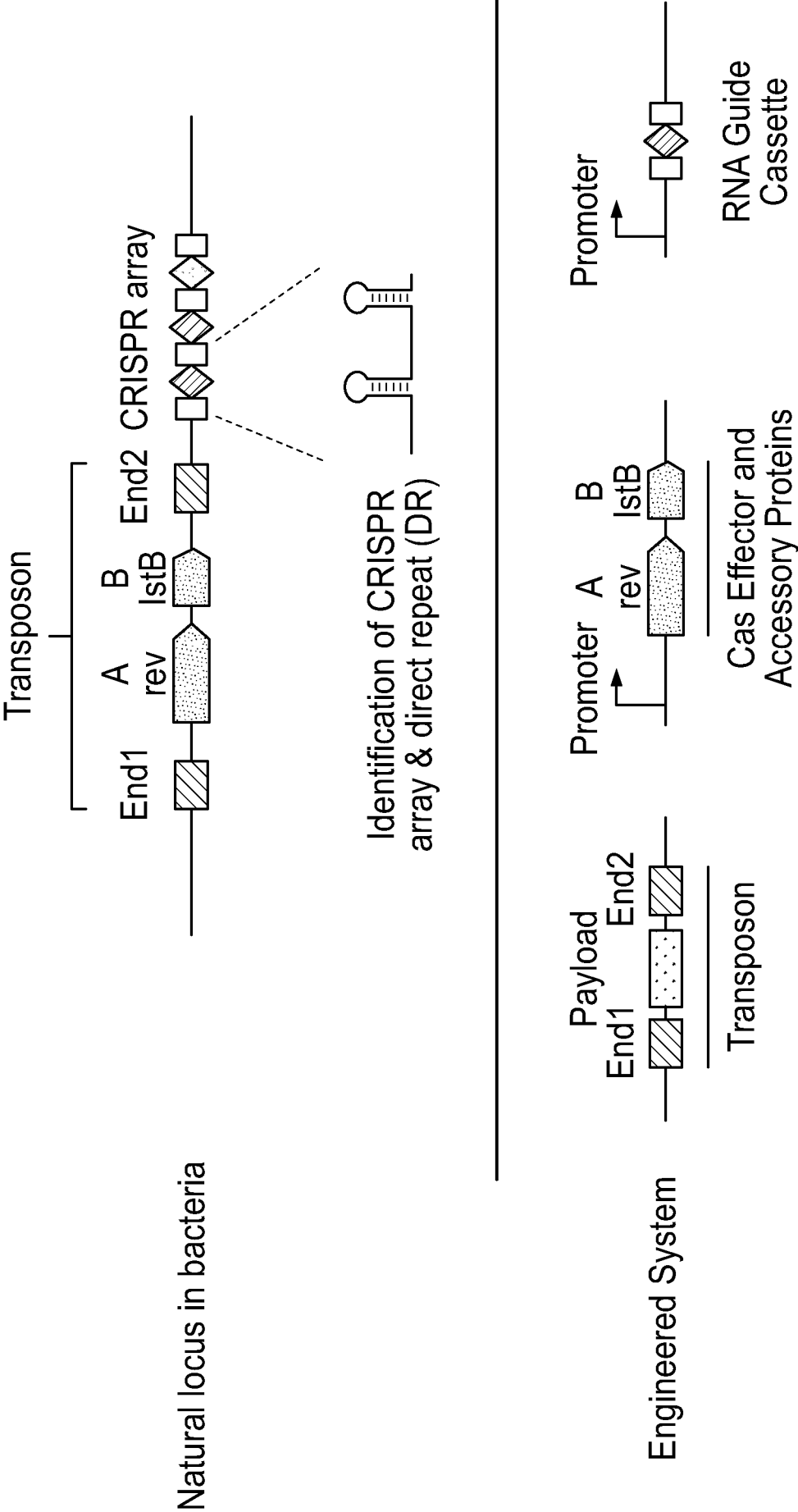


FIG. 8

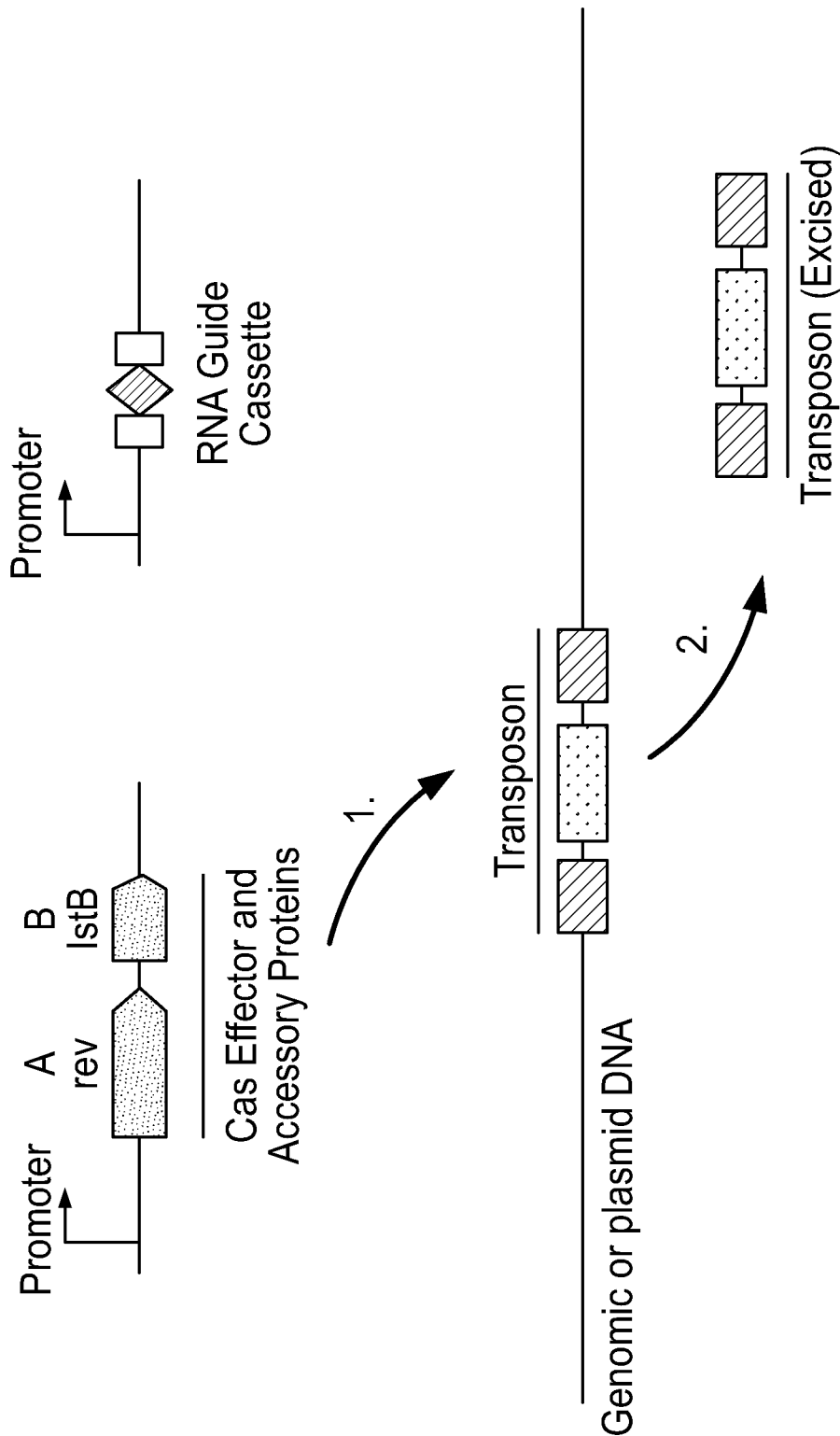


FIG. 9

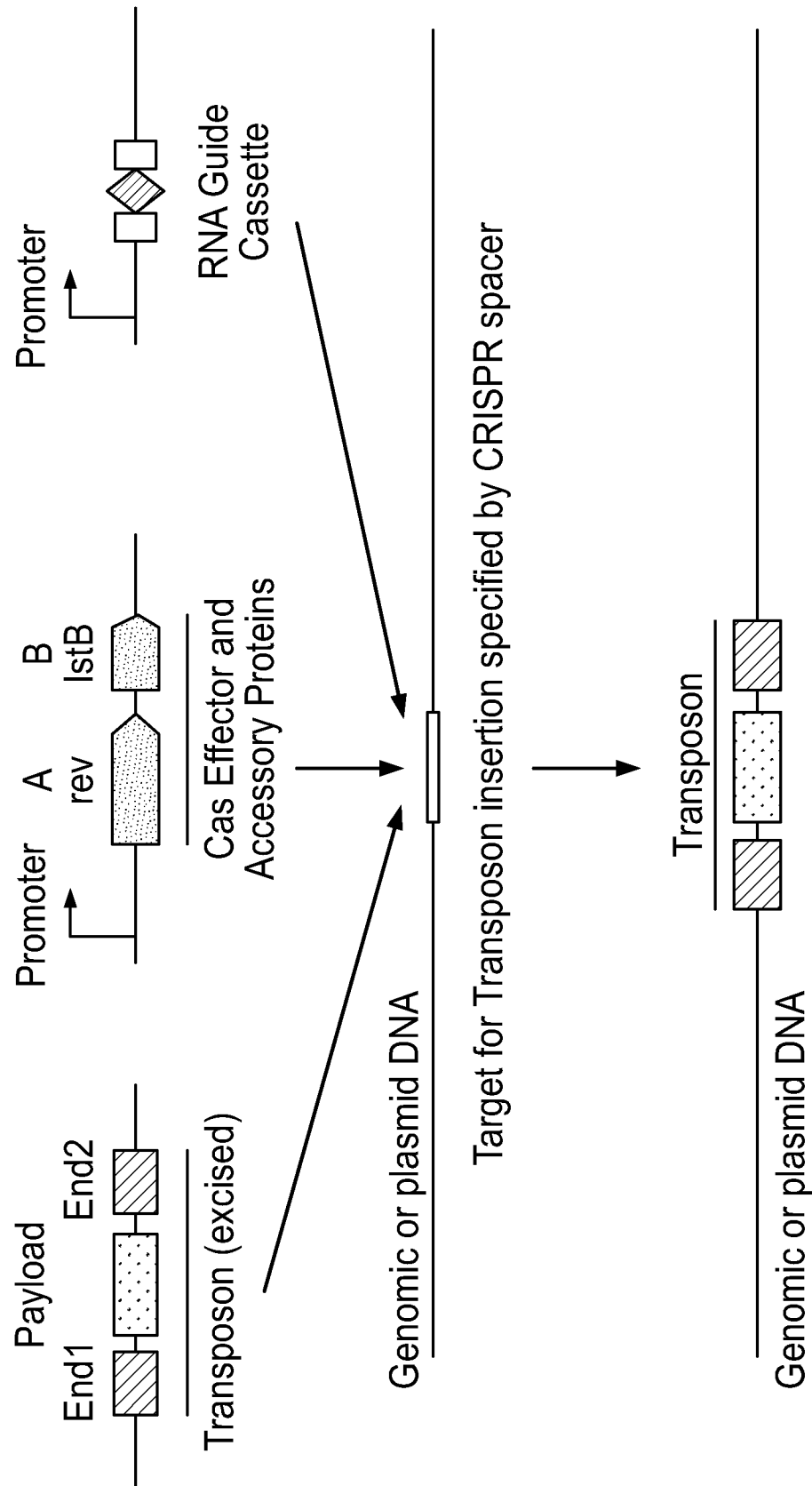


FIG. 10

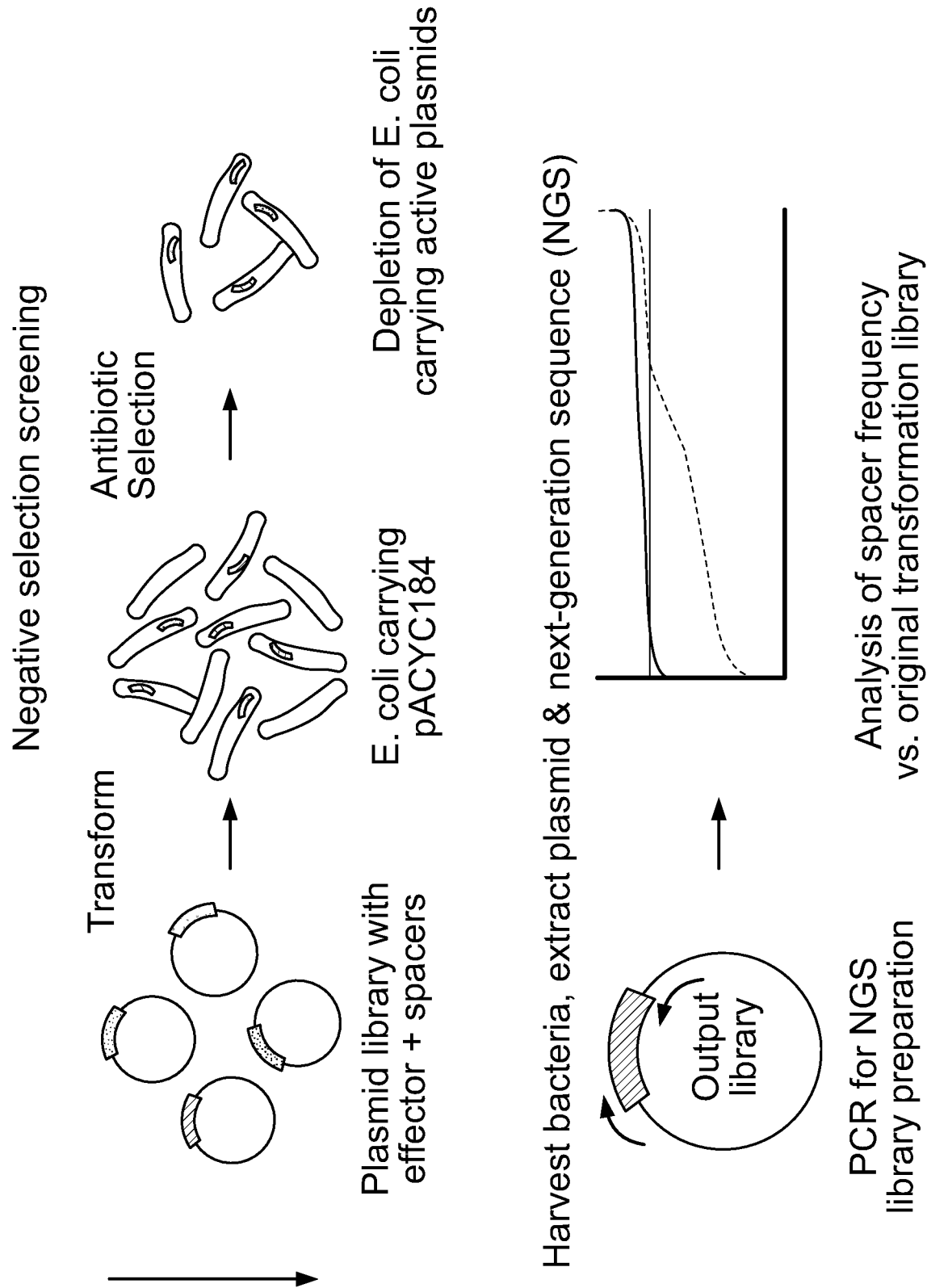


FIG. 11

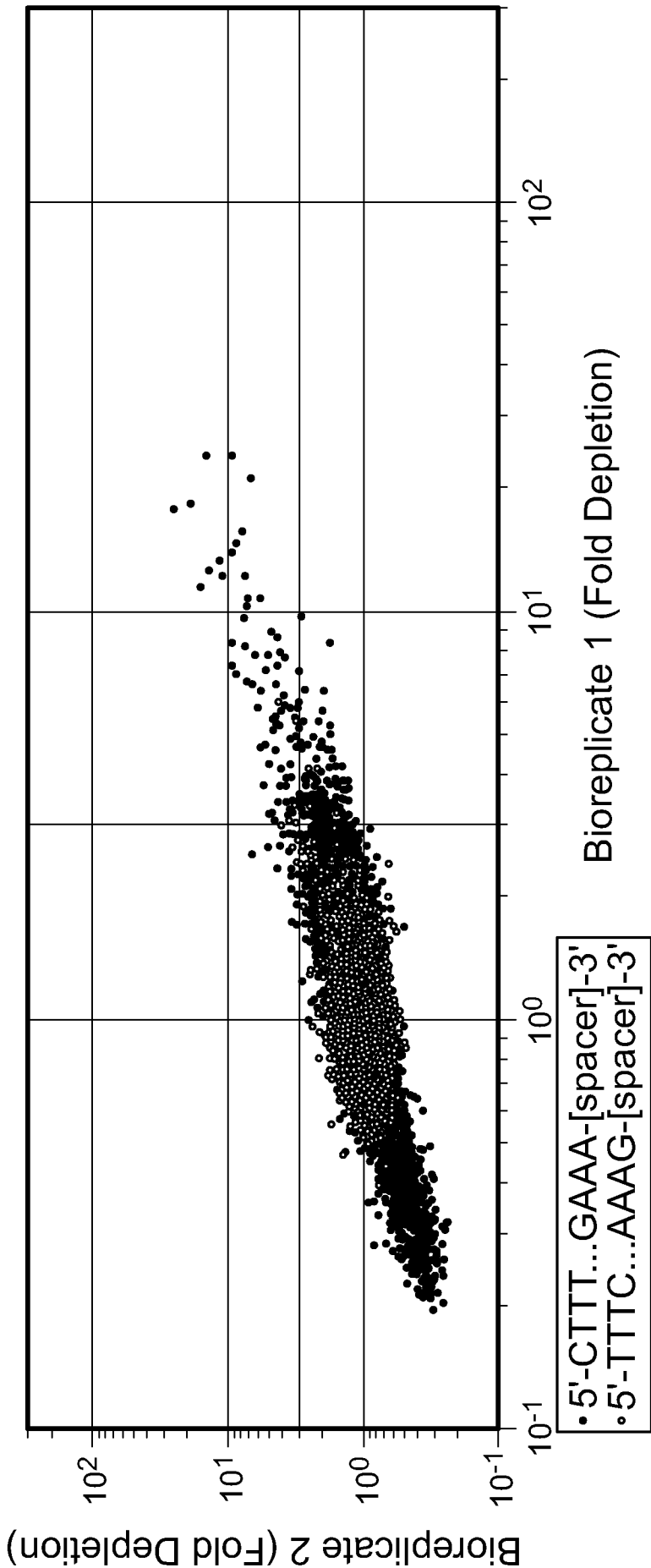


FIG. 12

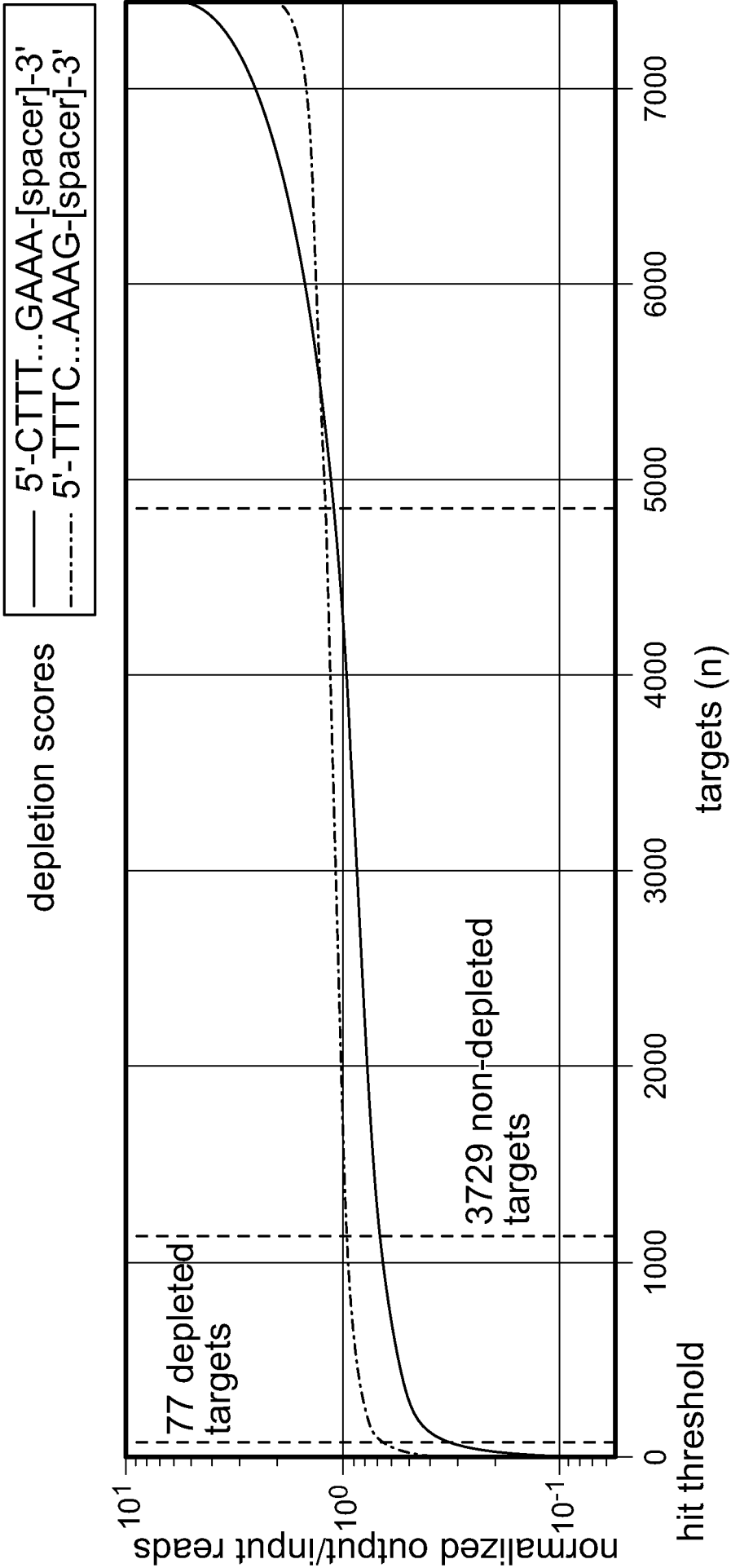


FIG. 13

distribution of targets on E. coli Essential Genes

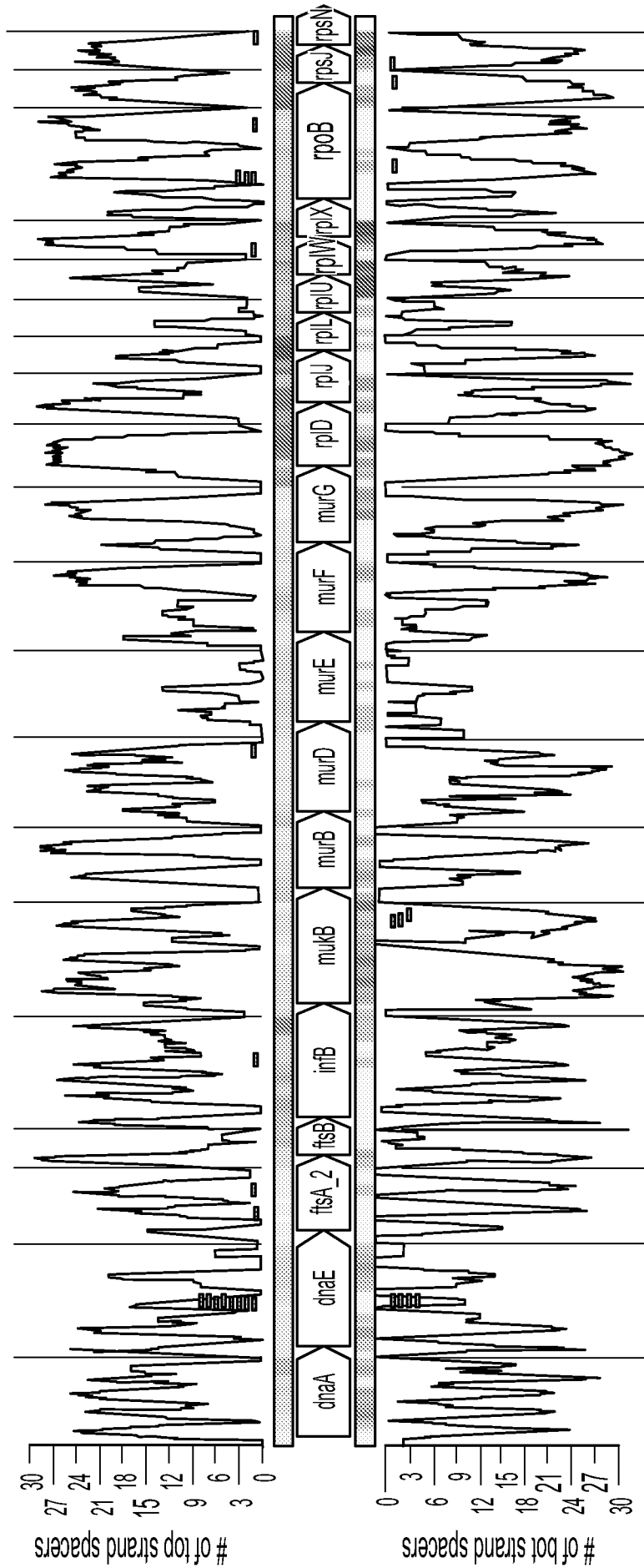


FIG. 14A

distribution of targets on pACYC184

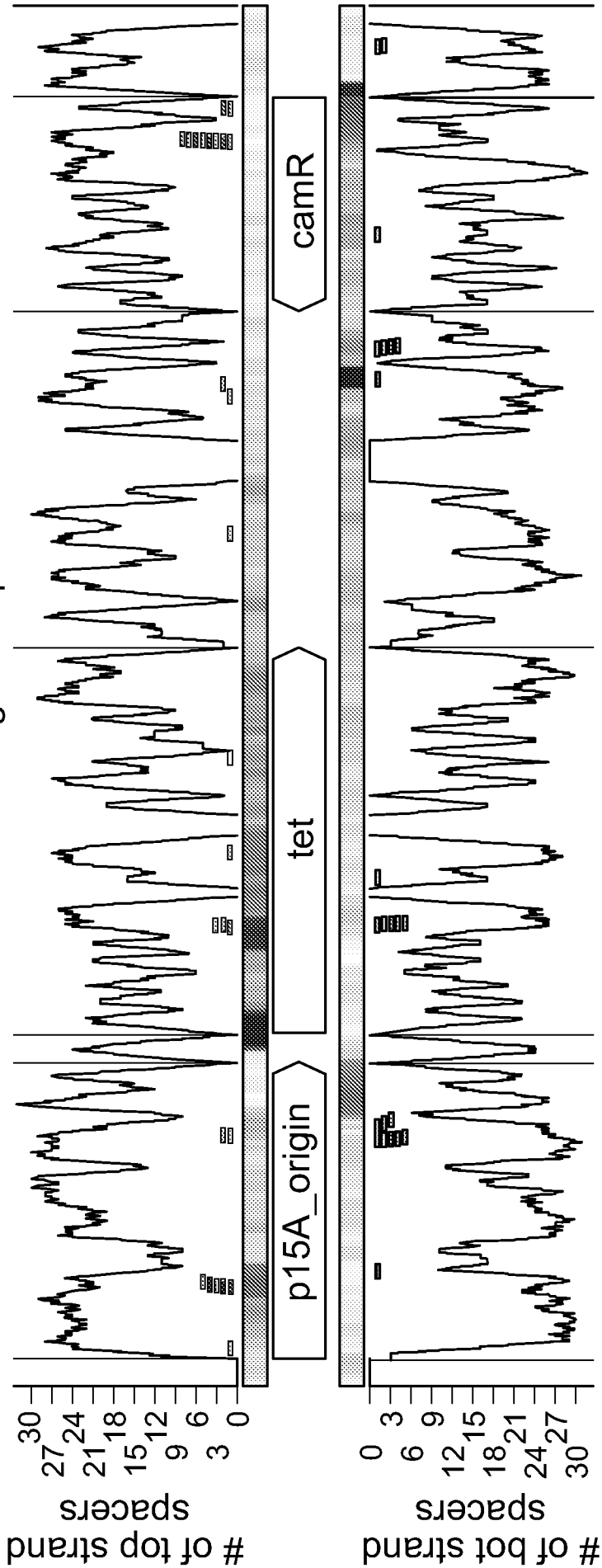


FIG. 14B

23/32

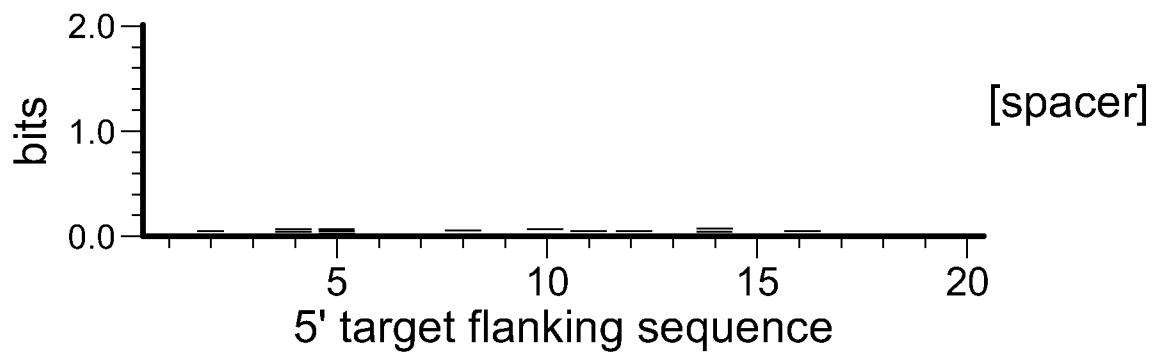


FIG. 15A

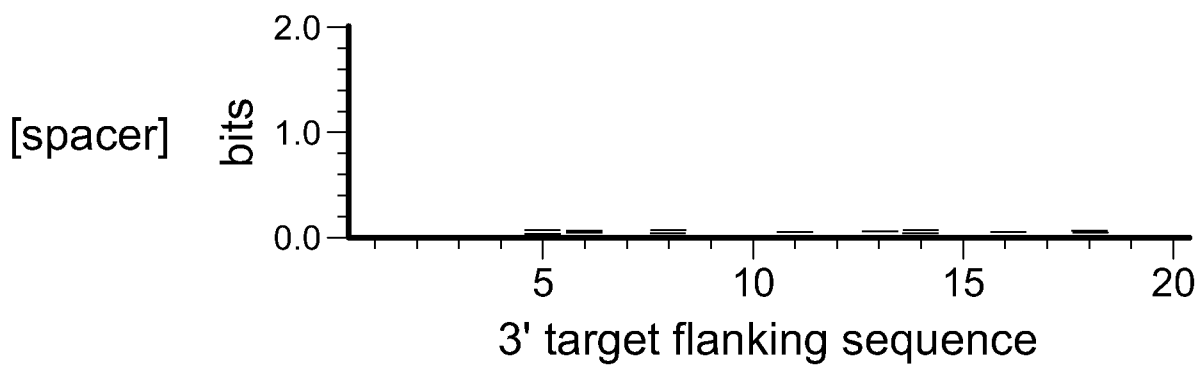


FIG. 15B

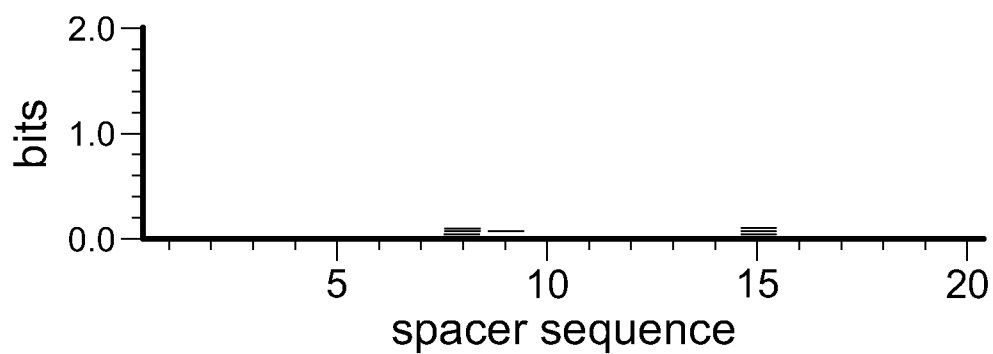


FIG. 15C

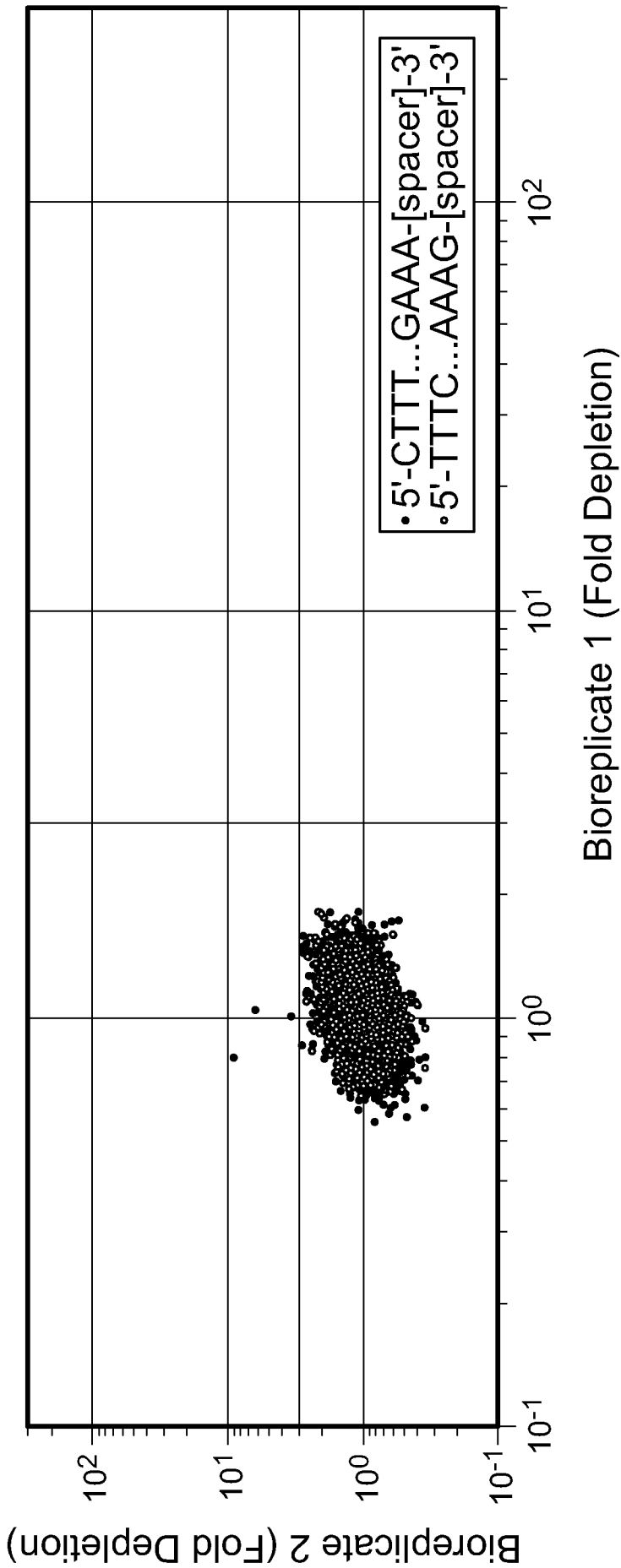


FIG. 16

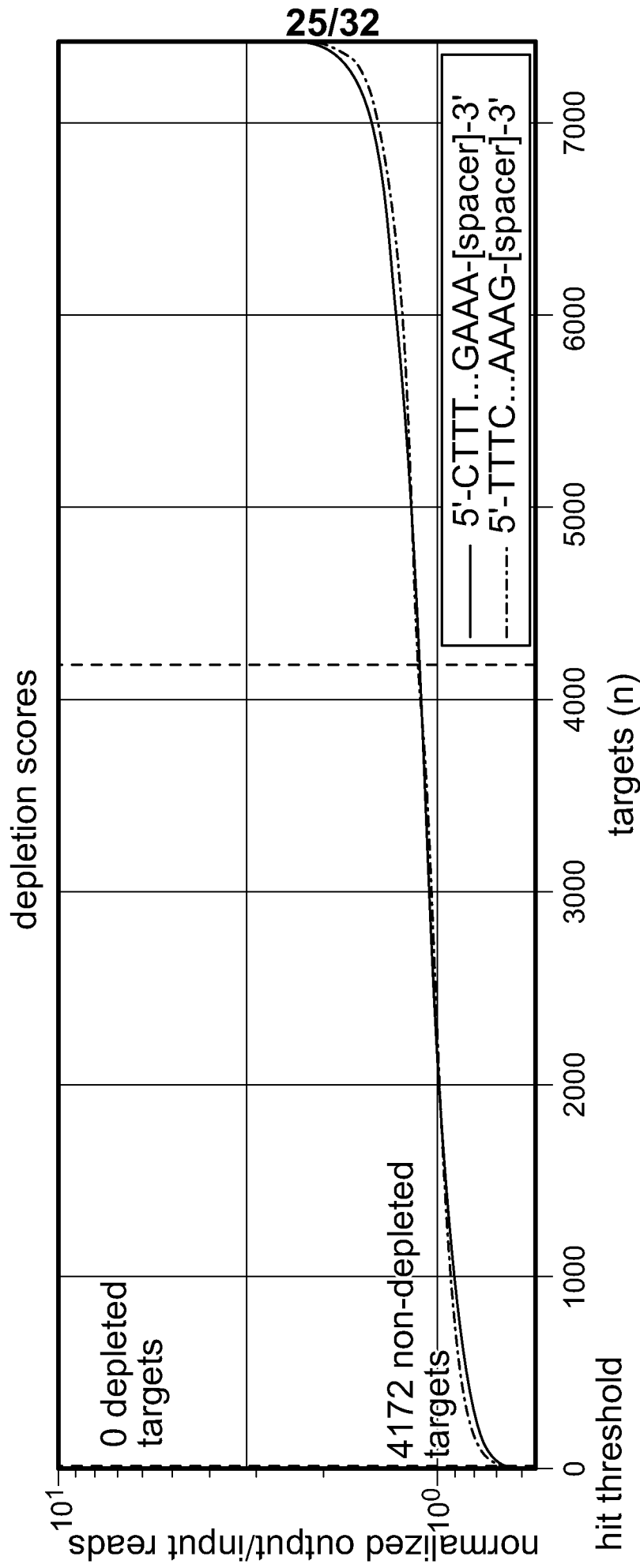


FIG. 17

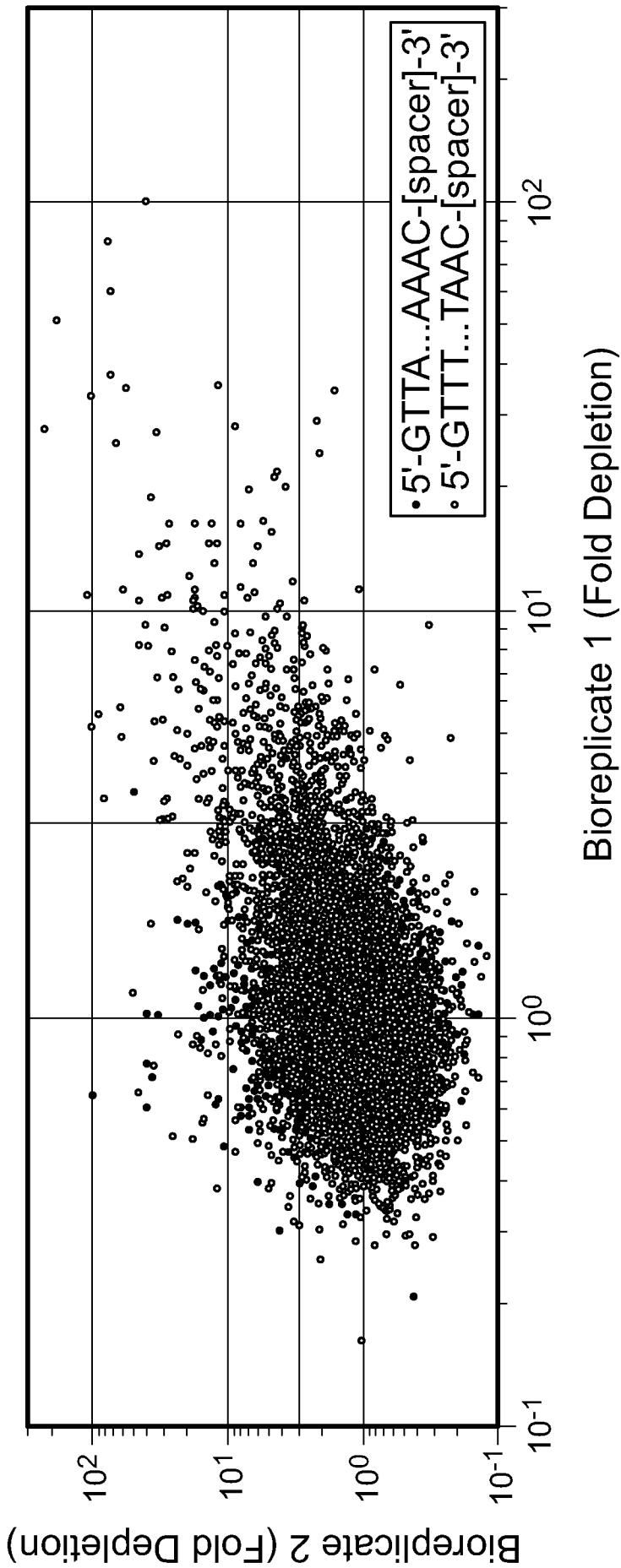


FIG. 18

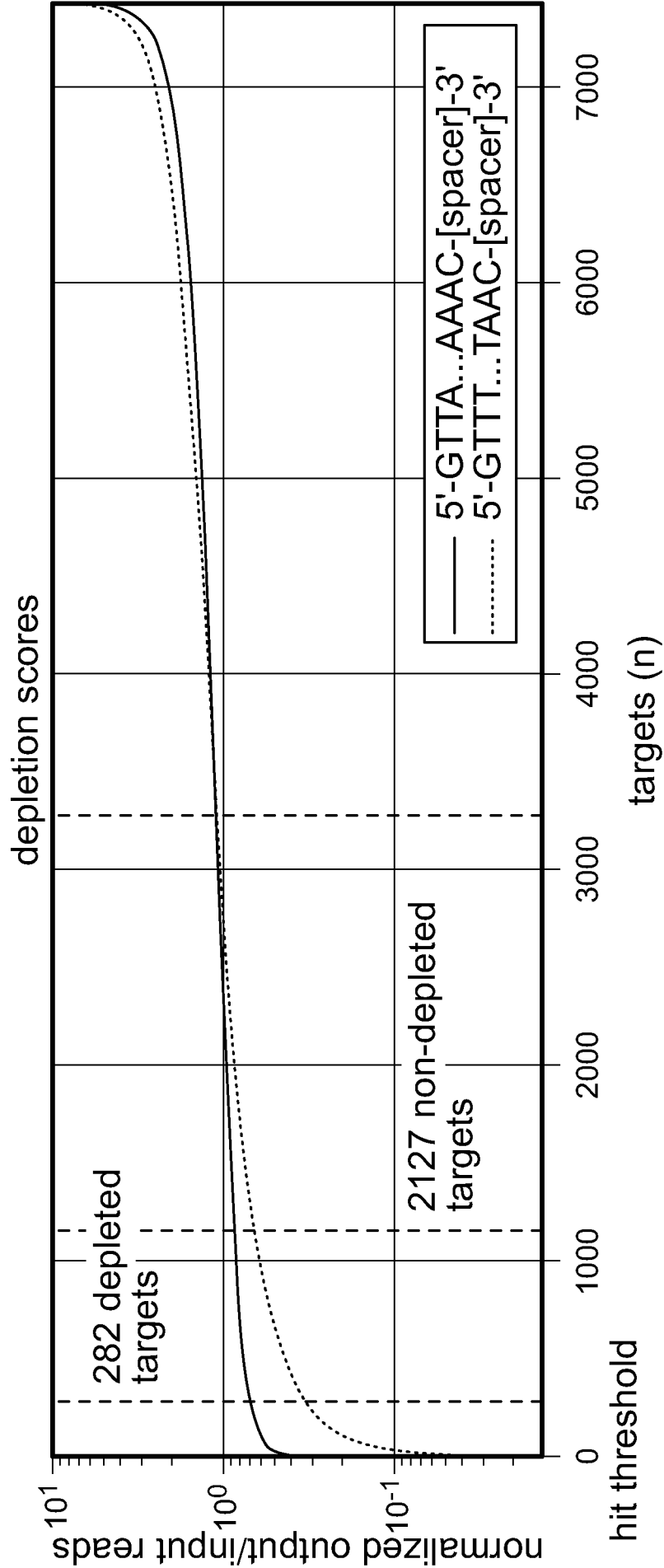


FIG. 19

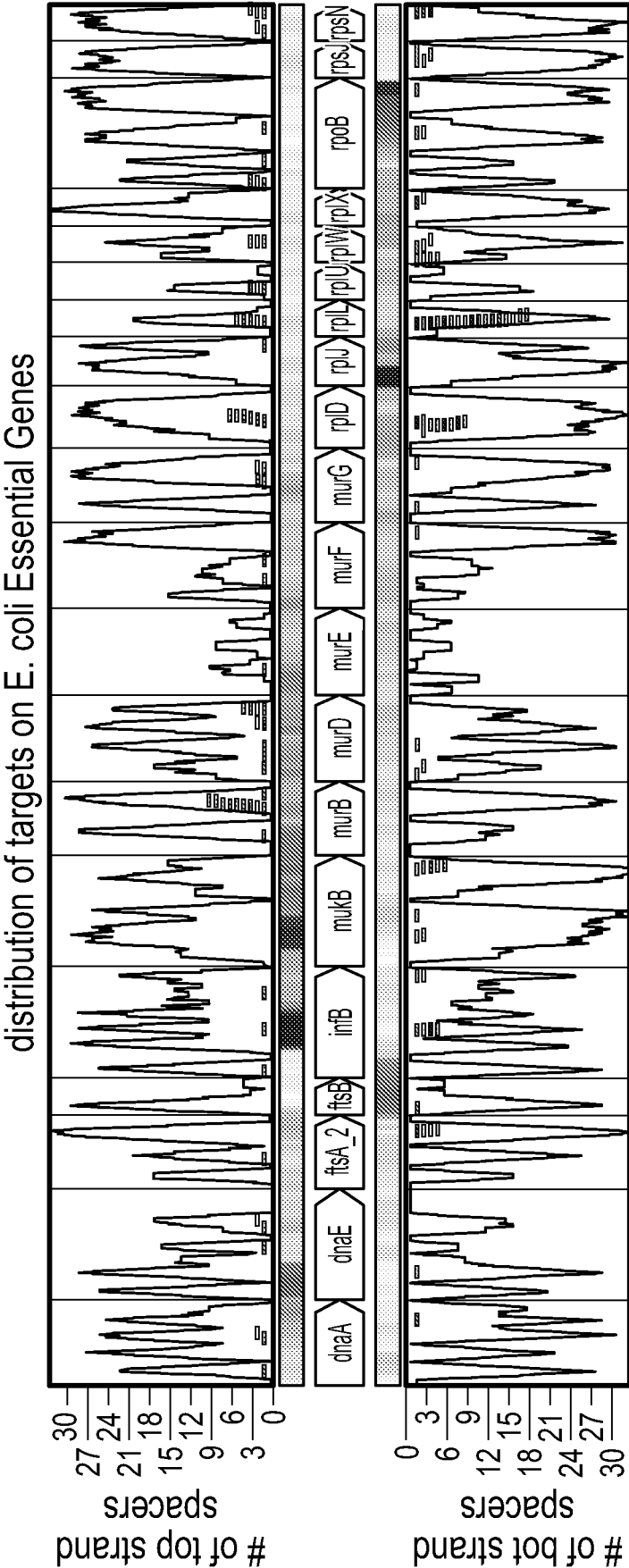


FIG. 20A

distribution of targets on pACYC184

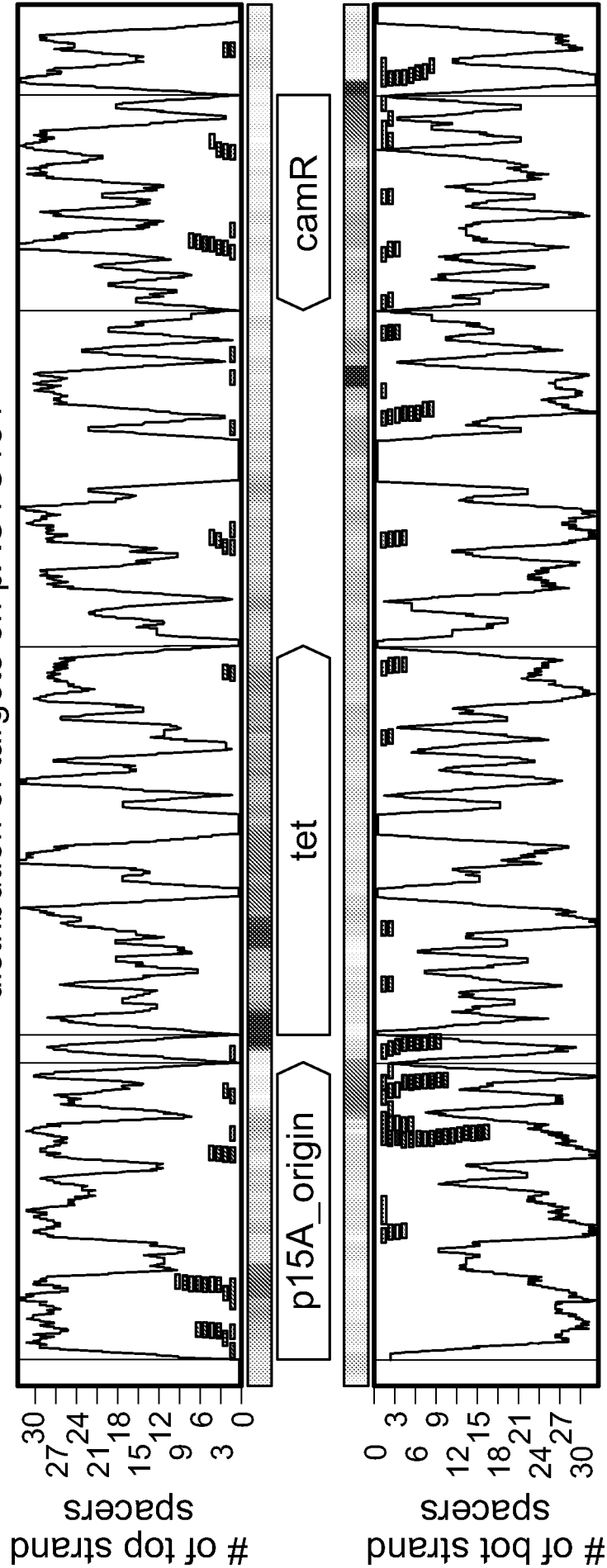


FIG. 20B

30/32

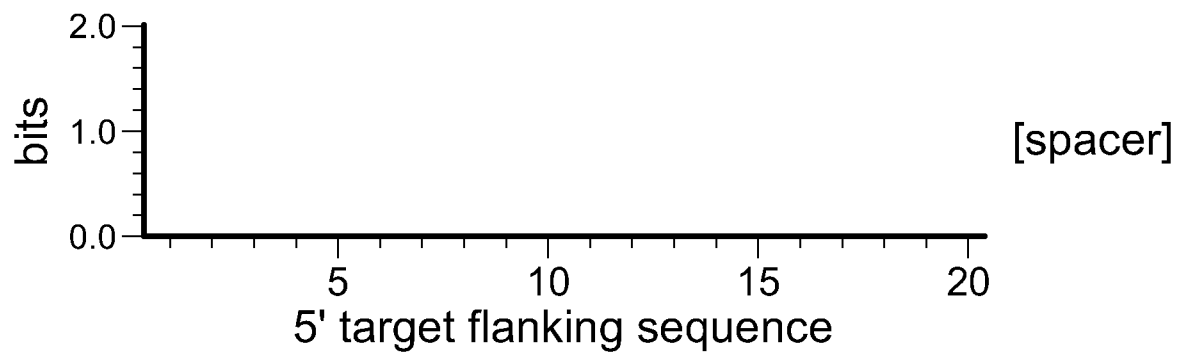


FIG. 21A

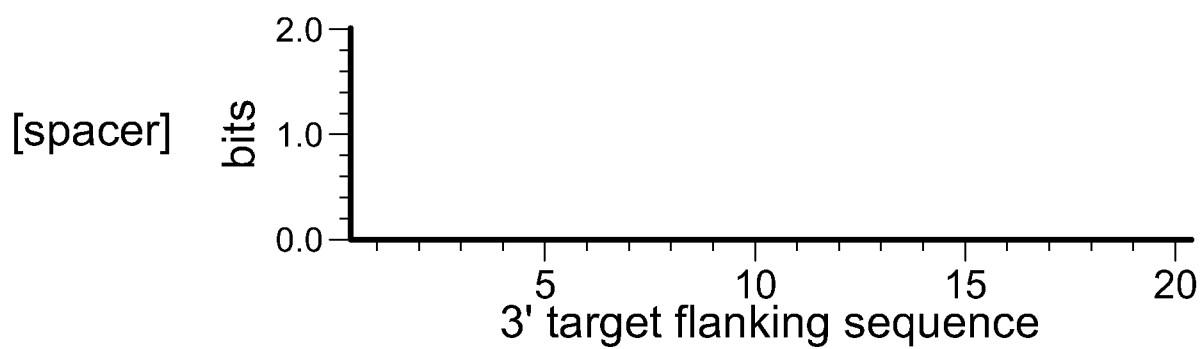


FIG. 21B

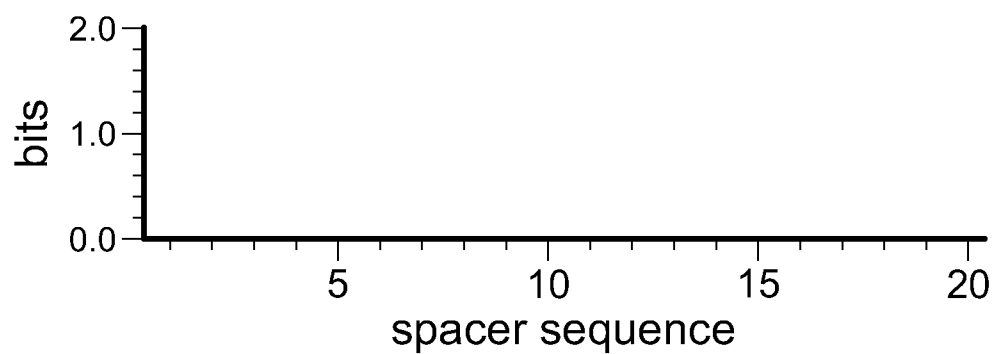
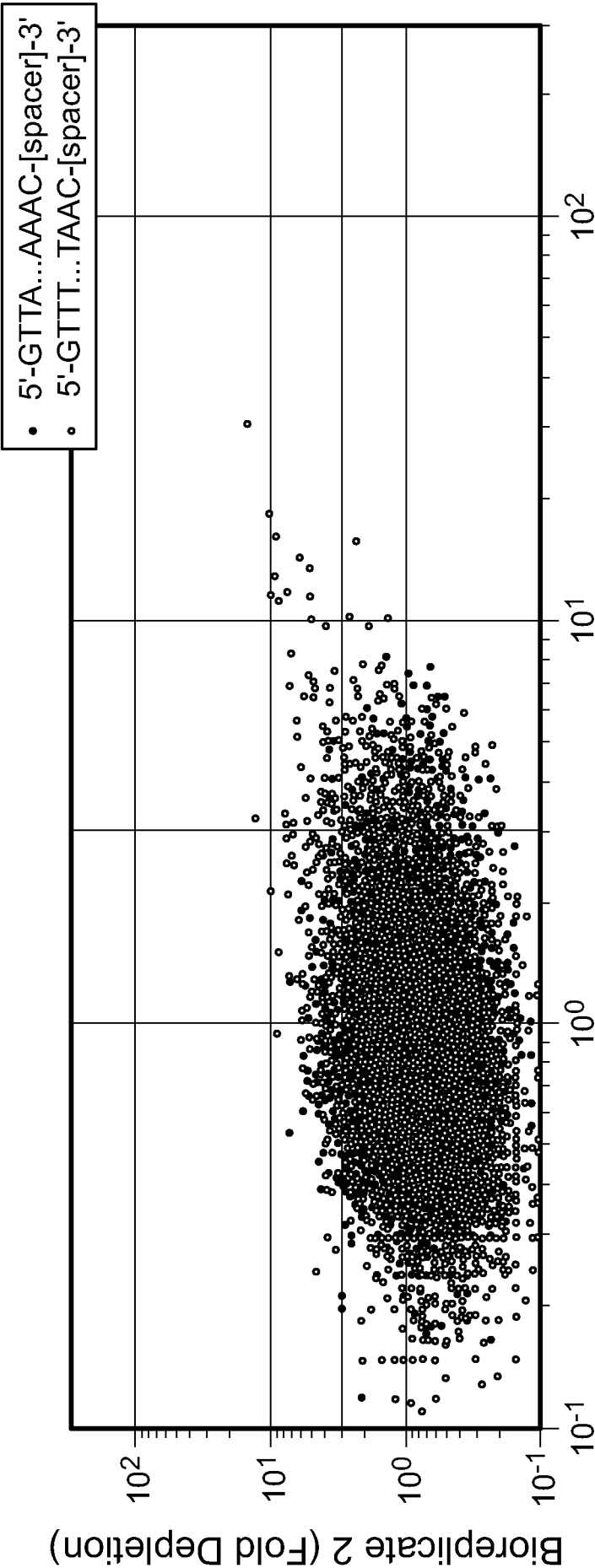


FIG. 21C



Bioreplicate 1 (Fold Depletion)

FIG. 22

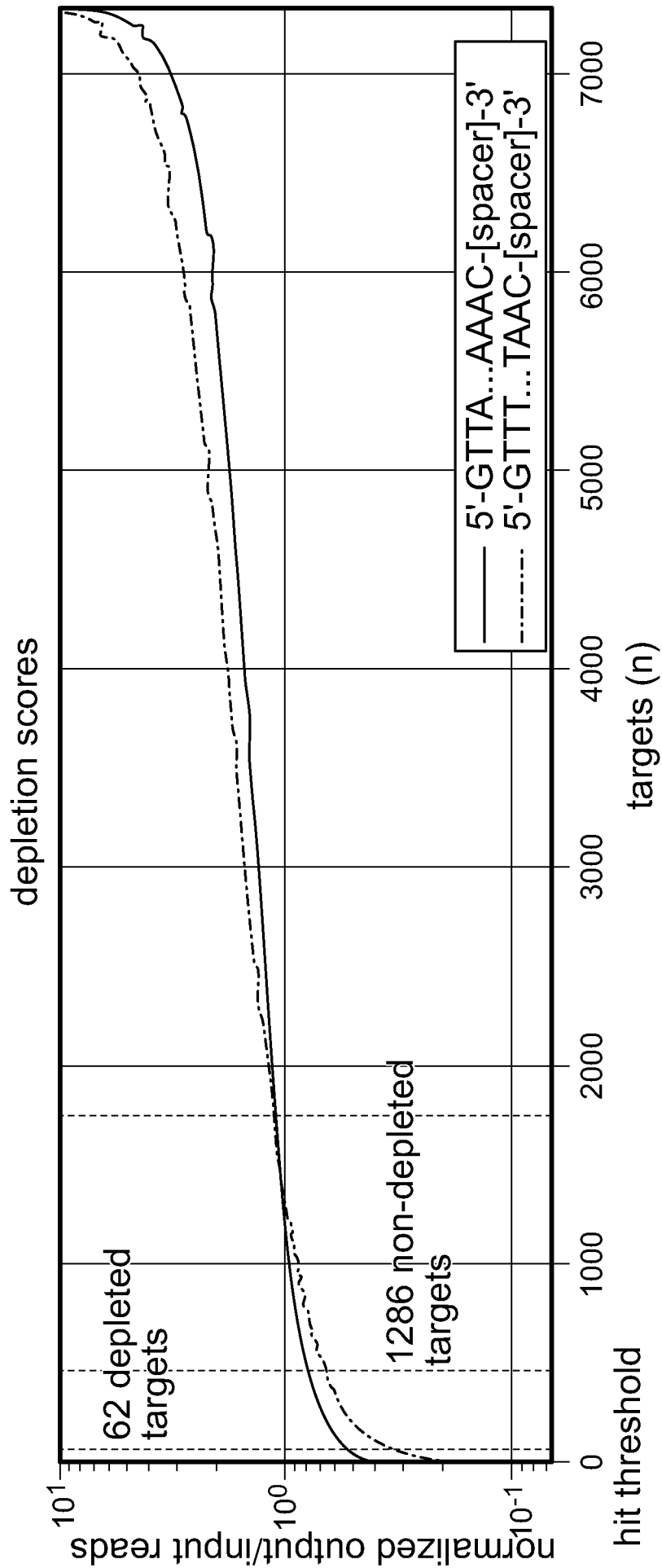


FIG. 23

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2018/059090

A. CLASSIFICATION OF SUBJECT MATTER

INV. C12N9/22 C12N15/10 C12N15/90 C12N15/113 C12N15/63
C07K14/195

ADD.

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

C12N C12R C07K

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

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

EPO-Internal, BIOSIS, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 2016/161207 A1 (EXELIGEN SCIENT INC [US]; SHEIKH FERRUKH [US]; KAWAMURA TETSUYA [US];) 6 October 2016 (2016-10-06) figures 2-10; examples 8-14 the whole document	1-39
X	WO 2017/137768 A1 (HORIZON DISCOVERY LTD [GB]; MAX-DELBRÜCK-CENTRUM FÜR MOLEKULARE MEDIZI) 17 August 2017 (2017-08-17) figure 13; example 2 the whole document	1-39

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

☒ See patent family annex.

* Special categories of cited documents :

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

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

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

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

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

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

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

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

"&" document member of the same patent family

Date of the actual completion of the international search

11 February 2019

Date of mailing of the international search report

25/02/2019

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040,
Fax: (+31-70) 340-3016

Authorized officer

Spindler, Mark-Peter

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2018/059090

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CHUNLONG XU ET AL: "piggyBac mediates efficient in vivo CRISPR library screening for tumorigenesis in mice", PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA, vol. 114, no. 4, 6 January 2017 (2017-01-06), pages 722-727, XP055547494, US ISSN: 0027-8424, DOI: 10.1073/pnas.1615735114 figure 1 the whole document -& CHUNLONG XU ET AL: "piggyBac mediates efficient in vivo CRISPR library screening for tumorigenesis in mice- Supplementary Information", PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA, vol. 114, no. 4, 6 January 2017 (2017-01-06), pages 722-727, XP055547501, US ISSN: 0027-8424, DOI: 10.1073/pnas.1615735114 figure S3 the whole document	1-39
X	----- MINGMING GAO ET AL: "CRISPR/Cas9-based Pten knock-out and Sleeping Beauty Transposon-mediated Nras knock-in induces hepatocellular carcinoma and hepatic lipid accumulation in mice", CANCER BIOLOGY & THERAPY, vol. 18, no. 7, 17 May 2017 (2017-05-17), pages 505-512, XP055547536, US ISSN: 1538-4047, DOI: 10.1080/15384047.2017.1323597 figures 1, 2 the whole document	1-39
A	----- JOSEPH E. PETERS ET AL: "Recruitment of CRISPR-Cas systems by Tn7-like transposons", PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA, vol. 114, no. 35, 15 August 2017 (2017-08-15), pages E7358-E7366, XP055547696, US ISSN: 0027-8424, DOI: 10.1073/pnas.1709035114 the whole document -----	1-39

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/US2018/059090

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2016161207 A1	06-10-2016	CN 108124453 A EP 3277805 A1 JP 2018513681 A KR 20180029953 A US 2018080051 A1 WO 2016161207 A1	05-06-2018 07-02-2018 31-05-2018 21-03-2018 22-03-2018 06-10-2016
-----	-----	-----	-----
WO 2017137768 A1	17-08-2017	AU 2017218651 A1 CA 3011886 A1 CN 109072258 A EP 3414333 A1 GB 2564296 A KR 20180127339 A SG 11201806400X A WO 2017137768 A1	09-08-2018 17-08-2017 21-12-2018 19-12-2018 09-01-2019 28-11-2018 30-08-2018 17-08-2017
-----	-----	-----	-----