

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

(19) World Intellectual Property  
Organization

International Bureau



(10) International Publication Number

WO 2022/109466 A1

(43) International Publication Date  
27 May 2022 (27.05.2022)

- (51) International Patent Classification:  
*CI2N 9/22* (2006.01)      *CI2Q 1/6888* (2018.01)  
*CI2Q 1/6869* (2018.01)
- (21) International Application Number:  
 PCT/US2021/060547
- (22) International Filing Date:  
 23 November 2021 (23.11.2021)
- (25) Filing Language: English
- (26) Publication Language: English
- (30) Priority Data:  
 63/117,441      23 November 2020 (23.11.2020) US  
 63/118,307      25 November 2020 (25.11.2020) US
- (71) Applicant: THE REGENTS OF THE UNIVERSITY OF CALIFORNIA [US/US]; 9500 Gilman Drive, La Jolla, California 90093-01910 (US).
- (72) Inventor: AKBARI, Omar; 9500 Gilman Drive, La Jolla, California 92093-0910 (US).
- (74) Agent: GREY, Kathryn et al.; Fish & Richardson, P.C., P.O. 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, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.
- (84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH,

(54) Title: SYSTEMS AND METHODS FOR IDENTIFYING NOVEL CRISPR ASSOCIATED PROTEINS

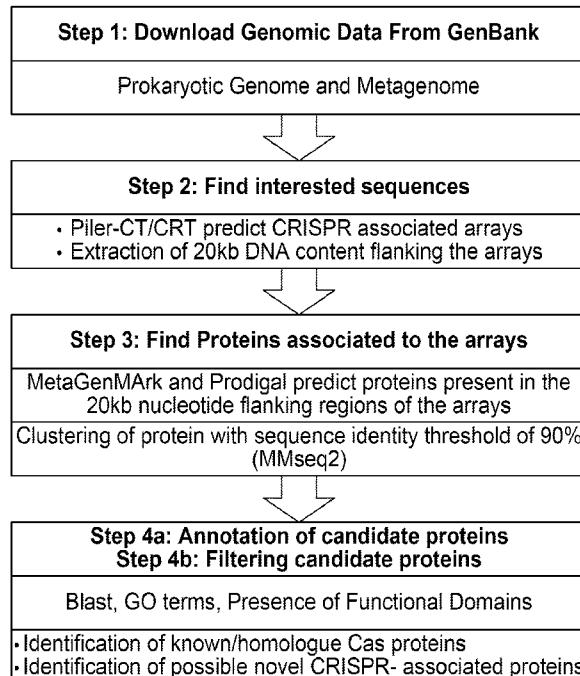


FIG. 1

(57) Abstract: Provided herein are systems and methods for identifying Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR)-associated proteins. For example, a method of identifying Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR)-associated proteins can include: (a) obtaining a plurality of genomic sequences, wherein a genomic sequence of the plurality of genomic sequences comprises a CRISPR-associated array; (b) determining a subset of the plurality of genomic sequences comprising a plurality of coding sequences within a 20 kilobase (kb) sequence flanking region either at the 3' or 5' end of the CRISPR-associated array; and (c) analyzing a coding sequence of the plurality of coding sequences and thereby identifying the CRISPR-associated protein based on the coding sequence.

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

**Published:**

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

## SYSTEMS AND METHODS FOR IDENTIFYING NOVEL CRISPR ASSOCIATED PROTEINS

### CROSS-REFERENCE TO RELATED APPLICATIONS

5 This application claims priority to U.S. Provisional Patent Application No. 63/117,441, filed on November 23, 2020, and U.S. Provisional Patent Application No. 63/118,307, filed on November 25, 2020. The disclosure of these prior applications are considered part of the disclosure of this application, and are incorporated in their entireties into this application.

10

### TECHNICAL FIELD

The present disclosure relates to systems, methods, and materials for identifying candidate CRISPR associated proteins.

### SEQUENCE LISTING

15 This application contains a Sequence Listing that has been submitted electronically as an ASCII text file named SequenceListing.txt. The ASCII text file, created on November 22, 2021, is 531 kilobytes in size. The material in the ASCII text file is hereby incorporated by reference in its entirety.

20

### BACKGROUND

The systematic interrogation of genomes and genetic reprogramming of cells involves targeting sets of genes for expression or repression. Currently the most common approach for targeting arbitrary genes for regulation is to use RNA interference (RNAi). This approach has limitations. For example, RNAi can exhibit significant off-target effects and toxicity.

25 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  
30 examples of which have been engineered and adapted to produce important biotechnologies. The components of the systems involved in host defense include one or more effector proteins capable of modifying DNA or RNA and a RNA guide element that is responsible for

targeting these protein activities to a specific sequence on the phage DNA or RNA. 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 that consist of a single effector protein that complexes with the crRNA to target DNA or

5 RNA substrates. The single-subunit effector compositions of the Class 2 systems provide a simpler component set for engineering and application translation, and has thus far been important sources of programmable effectors. The discovery, engineering, and optimization of novel Class 2 systems may lead to widespread and powerful programmable technologies for genome engineering and beyond.

10 There is need in the field for a technology that allows precise targeting of nuclease activity (or other protein activities) to distinct locations within a target DNA in a manner that does not require the design of a new protein for each new target sequence. In addition, there is a need in the art for methods of controlling gene expression with minimal off-target effects.

## 15 SUMMARY

This document provides compositions, methods, and material for identifying Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR)-associated proteins. For example, provided herein are methods including (a) obtaining a set of genomic sequences, wherein a genomic sequence of the set of genomic sequences comprises a 20 CRISPR-associated array; (b) determining coding sequences within a 20 kilobase (kb) sequence flanking either 3' or 5' of the CRISPR-associated array; and (c) filtering the coding sequences and using the filtered coding sequences to identify CRISPR-associated proteins. The present disclosure is based on the discovery that methods, including computational methods, can be used to mine prokaryotic genomes and metagenomes for novel CRISPR-associated proteins.

Provided herein are methods of identifying a Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR)-associated protein comprising: (a) obtaining a plurality of genomic sequences, wherein a genomic sequence of the plurality of genomic sequences comprises a CRISPR-associated array; (b) determining a subset of the plurality of genomic sequences comprising a plurality of coding sequences within a 20 kilobase (kb) sequence flanking region either at the 3' or 5' end of the CRISPR-associated array; and (c) analyzing a coding sequence of the plurality of coding sequences and thereby identifying the CRISPR-associated protein based on the coding sequence.

In some embodiments, the obtaining step comprises selecting, within the plurality of genomic sequences, a genomic sequence comprising a CRISPR-associated array.

Also provided herein are methods of identifying a CRISPR-associated protein comprising: (a) obtaining a plurality of genomic sequences; (b) selecting, within the plurality of genomic sequences, a genomic sequence comprising a CRISPR-associated array; (c) determining a subset of the plurality of genomic sequences comprising a plurality of coding sequences within a 20 kilobase (kb) sequence flanking region either at the 3' or 5' end of the CRISPR-associated array; and (d) analyzing a coding sequence of the plurality of coding sequences and thereby identifying the CRISPR-associated protein based on the coding sequence.

In some embodiments, the plurality of genomic sequences comprise one or more of genomes, wherein the one or more of genomes are selected from: a prokaryotic genome and metagenome. In some embodiments, the selecting step comprises using an algorithm selected from the group consisting of PILER-CR, CRISPR Recognition Tool (CRT), and combinations thereof. In some embodiments, the determining step comprises using an algorithm selected from the group consisting of MetaGeneMark, Prodigal, and combinations thereof.

In some embodiments, the analyzing step comprises filtering the coding sequence that comprises more than 500 amino acids. In some embodiments, the analyzing step comprises filtering a coding sequence that comprises more than 800 amino acids. In some embodiments, the analyzing step further comprises classifying the CRISPR-associated array based on having three or more coding sequences present in the 20 kb flanking region. In some embodiments, the analyzing step further comprises determining a relative position of the coding sequence in the 20 kb flanking region relative to the CRISPR-associated array.

In some embodiments, the analyzing of the coding sequence further comprises removing known CRISPR-associated proteins from the identified CRISPR-associated proteins. In some embodiments, the analyzing of the coding sequence comprises using an algorithm selected from the group consisting of HHMSCAN and RPS-BLAST. In some embodiments, the analyzing of the coding sequence further comprises determining the presence of a structural domain. In some embodiments, the analyzing of the coding sequence comprises determining the presence of a functional domain. In some embodiments, the functional domain comprises a DNA binding domain, a RNA binding domain, a nuclease, a helicase, a restriction domain, or a structural maintenance of chromosomes (SMC) domain.

Also provided herein are computer implemented methods comprising: (a) obtaining a plurality of genomic sequences; (b) selecting, within the plurality of genomic sequences, a genomic sequence comprising a CRISPR-associated array; (c) determining a subset of the plurality of genomic sequences comprising a plurality of coding sequences within a 20 kilobase (kb) sequence flanking region either at the 3' or 5' end of the CRISPR-associated array; and (d) analyzing a coding sequence of the plurality of coding sequences and thereby identifying a CRISPR-associated protein based on the coding sequence.

In some embodiments, the plurality of genomic sequences comprises one or more of genomes, wherein the one or more of genomes are selected from: a prokaryotic genome and metagenome. In some embodiments, the selecting step comprises using an algorithm selected from the group consisting of PILER-CR, CRISPR Recognition Tool (CRT), and combinations thereof. In some embodiments, the determining step comprises using an algorithm selected from the group consisting of MetaGeneMark, Prodigal, and combinations thereof.

In some embodiments, the analyzing step comprises filtering the coding sequence that comprises more than 500 amino acids. In some embodiments, the analyzing step comprises filtering a coding sequence that comprises more than 800 amino acids. In some embodiments, the analyzing step further comprises classifying the CRISPR-associated array based on having three or more coding sequences present in the 20 kb flanking region. In some embodiments, the analyzing step further comprises determining a relative position of the coding sequence in the 20 kb flanking region relative to the CRISPR-associated array.

In some embodiments, the analyzing of the coding sequence further comprises removing known CRISPR-associated proteins from the identified CRISPR-associated proteins. In some embodiments, the analyzing of the coding sequence comprises using an algorithm selected from the group consisting of HHMSCAN and RPS-BLAST. In some embodiments, the analyzing of the coding sequence further comprises determining the presence of a structural domain. In some embodiments, the analyzing of the coding sequence comprises determining the presence of a functional domain. In some embodiments, the functional domain comprises a DNA binding domain, a RNA binding domain, a nuclease, a helicase, a restriction domain, or a structural maintenance of chromosomes (SMC) domain.

Also provided herein are non-naturally occurring CRISPR/Cas systems comprising: (a) a guide RNA, wherein the guide RNA comprises a repeat sequence and a spacer sequence capable of hybridizing to a target nucleic acid; and (b) a CRISPR-associated protein or a nucleic acid encoding the CRISPR-associated protein, wherein the CRISPR-associated

protein comprises an amino acid sequence that is at least 80% identical to a sequence selected from SEQ ID NOs: 1-50.

In some embodiments, the CRISPR-associated protein is capable of binding to the guide RNA. In some embodiments, the CRISPR-associated protein comprises an amino acid sequence that is at least 85% identical to a sequence selected from SEQ ID NOs: 1-50. In some embodiments, the CRISPR-associated protein comprises an amino acid sequence that is at least 90% identical to a sequence selected from SEQ ID NOs: 1-50. In some embodiments, the CRISPR-associated protein comprises an amino acid sequence that is at least 95% identical to a sequence selected from SEQ ID NOs: 1-50. In some embodiments, the CRISPR-associated protein comprises an amino acid sequence selected from SEQ ID NO: 1-50.

In some embodiments, the target nucleic acid is an RNA or DNA. In some embodiments, the targeting of the target nucleic acid results in a modification of the target nucleic acid. In some embodiments, the modification of the target nucleic acid is a cleavage event.

In some embodiments, the guide RNA further comprises a trans-activating CRISPR RNA (tracrRNA). In some embodiments, the system is present in a delivery system. In some embodiments, the delivery system comprises a delivery vehicle selected from the group consisting of an adeno-associated virus, a nanoparticle, and a liposome.

Also provided herein are methods of treating a condition or disease in a subject in need thereof, the method comprising administering to the subject any one of the systems provided herein, wherein the spacer sequence is substantially complementary to a target nucleic acid associated with the condition or disease; wherein the CRISPR-associated protein associates with the guide RNA to form a complex; wherein the complex binds to the target nucleic acid sequence; and wherein upon binding of the complex to the target nucleic acid sequence the CRISPR- associated protein cleaves the target nucleic acid, thereby treating the condition or disease in the subject.

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 disclosure pertains. Although methods and materials similar or equivalent to those described herein can be used in the practice or testing of the present disclosure, suitable methods and materials are described below. All publications, patent applications, patents, and other references mentioned 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 disclosure will be apparent from the following detailed description, and from the claims.

5

## BRIEF DESCRIPTION OF DRAWINGS

**FIG. 1** is a schematic diagram showing an exemplary method for identifying CRISPR-associated proteins.

10 **FIG. 2** is a schematic diagram showing exemplary step 1 and exemplary step 2 of a method for identifying CRISPR-associated proteins.

**FIG. 3** is a schematic diagram showing exemplary step 3 of a method for identifying CRISPR-associated proteins.

**Figures 4A-4B** show the Cas9 size distribution by member and cluster count.

15 **Figures 5A-5C** are histograms showing number of CRISPR-associated proteins typically associated with the different types of Cas Type II effectors.

**Figures 6A and 6B** are schematic diagrams showing further annotation and filtering done on the 10,913 candidate clusters.

**Figure 7** shows a summary of the method as described herein.

**Figure 8** is a schematic diagram showing an exemplary workflow.

20

## DETAILED DESCRIPTION

This document provides methods of identifying Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR)-associated proteins where the method includes computation identification. In some embodiments, these computational methods are directed to identifying 25 CRISPR-associated proteins that co-occur in close proximity to CRISPR arrays. It should be understood that the methods and calculations described herein may be performed on one or more computing devices.

30 Various non-limiting aspects of these methods and systems are described herein, and can be used in any combination without limitation. Additional aspects of various components of systems and methods for identifying CRISPR associated proteins are known in the art.

It must be noted that, as used in the specification and the appended claims, the singular forms “a,” “an” and “the” include plural referents unless the context clearly dictates otherwise.

As used herein, the terms “about” and “approximately,” when used to modify an amount specified in a numeric value or range, indicate that the numeric value as well as reasonable deviations from the value known to the skilled person in the art, for example ± 20%, ± 10%, or ± 5%, are within the intended meaning of the recited value.

5 As used herein, a “cell” can refer to either a prokaryotic or eukaryotic cell, optionally obtained from a subject or a commercially available source.

As used herein, “delivering”, “gene delivery”, “gene transfer”, “transducing” can refer to the introduction of an exogenous polynucleotide into a host cell, irrespective of the method used for the introduction. Such methods include a variety of well-known techniques such as  
10 vector-mediated gene transfer (e.g., viral infection/transfection, or various other protein-based or lipid-based gene delivery complexes) as well as techniques facilitating the delivery of “naked” polynucleotides (e.g., electroporation, “gene gun” delivery and various other techniques used for the introduction of polynucleotides). The introduced polynucleotide may be stably or transiently maintained in the host cell. Stable maintenance typically requires that  
15 the introduced polynucleotide either contains an origin of replication compatible with the host cell or integrates into a replicon of the host cell such as an extrachromosomal replicon (e.g., a plasmid) or a nuclear or mitochondrial chromosome.

In some embodiments, a polynucleotide can be inserted into a host cell by a gene delivery molecule. Examples of gene delivery molecules can include, but are not limited to,  
20 liposomes, micelles biocompatible polymers, including natural polymers and synthetic polymers; lipoproteins; polypeptides; polysaccharides; lipopolysaccharides; artificial viral envelopes; metal particles; and bacteria, or viruses, such as baculovirus, adenovirus and retrovirus, bacteriophage, cosmid, plasmid, fungal vectors and other recombination vehicles typically used in the art which have been described for expression in a variety of eukaryotic  
25 and prokaryotic hosts, and may be used for gene therapy as well as for simple protein expression.

As used herein, the term “encode” as it is applied to nucleic acid sequences refers to a polynucleotide which is said to “encode” a polypeptide if, in its native state or when manipulated by methods well known to those skilled in the art, can be transcribed and/or  
30 translated to produce the mRNA for the polypeptide and/or a fragment thereof. The antisense strand is the complement of such a nucleic acid, and the encoding sequence can be deduced therefrom.

The term “exogenous” refers to any material introduced from or originating from outside a cell, a tissue or an organism that is not produced by or does not originate from the same cell, tissue, or organism in which it is being introduced.

As used herein, “nucleic acid” is used to include any compound and/or substance that 5 comprise a polymer of nucleotides. In some embodiments, a polymer of nucleotides are referred to as polynucleotides. Exemplary nucleic acids or polynucleotides can include, but are not limited to, ribonucleic acids (RNAs), deoxyribonucleic acids (DNAs), threose nucleic acids (TNAs), glycol nucleic acids (GNAs), peptide nucleic acids (PNAs), locked nucleic acids (LNAs, including LNA having a  $\beta$ -D-ribo configuration,  $\alpha$ -LNA having an  $\alpha$ -L-ribo 10 configuration (a diastereomer of LNA), 2'-amino-LNA having a 2'-amino functionalization, and 2'-amino- $\alpha$ -LNA having a 2'-amino functionalization) or hybrids thereof. Naturally- occurring nucleic acids generally have a deoxyribose sugar (e.g., found in deoxyribonucleic acid (DNA)) or a ribose sugar (e.g., found in ribonucleic acid (RNA)).

A nucleic acid can contain nucleotides having any of a variety of analogs of these 15 sugar moieties that are known in the art. A deoxyribonucleic acid (DNA) can have one or more bases selected from the group consisting of adenine (A), thymine (T), cytosine (C), or guanine (G), and a ribonucleic acid (RNA) can have one or more bases selected from the group consisting of uracil (U), adenine (A), cytosine (C), or guanine (G).

In some embodiments, the term “nucleic acid” refers to a deoxyribonucleic acid 20 (DNA) or ribonucleic acid (RNA), or a combination thereof, in either a single- or double- stranded form. Unless specifically limited, the term encompasses nucleic acids containing known analogues of natural nucleotides that have similar binding properties as the reference nucleotides. Unless otherwise indicated, a particular nucleic acid sequence also implicitly encompasses complementary sequences as well as the sequence explicitly indicated. In some 25 embodiments of any of the isolated nucleic acids described herein, the isolated nucleic acid is DNA. In some embodiments of any of the isolated nucleic acids described herein, the isolated nucleic acid is RNA.

Modifications can be introduced into a nucleotide sequence by standard techniques 30 known in the art, such as site-directed mutagenesis and polymerase chain reaction (PCR)- mediated mutagenesis. Conservative amino acid substitutions are ones in which the amino acid residue is replaced with an amino acid residue having a similar side chain. Families of amino acid residues having similar side chains have been defined in the art. These families include amino acids with basic side chains (e.g., arginine, lysine and histidine), acidic side chains (e.g., aspartic acid and glutamic acid), uncharged polar side chains (e.g., asparagine,

cysteine, glutamine, glycine, serine, threonine, tyrosine, and tryptophan), nonpolar side chains (e.g., alanine, isoleucine, leucine, methionine, phenylalanine, proline, and valine), beta-branched side chains (e.g., isoleucine, threonine, and valine), and aromatic side chains (e.g., histidine, phenylalanine, tryptophan, and tyrosine), and aromatic side chains (e.g., histidine, phenylalanine, tryptophan, and tyrosine).

5 Unless otherwise specified, a “nucleotide sequence encoding a protein” includes all nucleotide sequences that are degenerate versions of each other and thus encode the same amino acid sequence.

The term “plurality” can refer to a state of having a plural (e.g., more than one) 10 number of different types of things (e.g., a cell, a genomic sequence, a subject, a system, or a protein). In some embodiments, a plurality of genomic sequences can be more than one genomic sequence wherein each genomic sequence is different from each other.

The term “subject” is intended to include any mammal. In some embodiments, the 15 subject is cat, a dog, a goat, a human, a non-human primate, a rodent (e.g., a mouse or a rat), a pig, or a sheep.

The term “transduced”, “transfected”, or “transformed” refers to a process by which 20 exogenous nucleic acid is introduced or transferred into a cell. A “transduced,” “transfected,” or “transformed” mammalian cell is one that has been transduced, transfected or transformed with exogenous nucleic acid (e.g., a gene delivery vector) that includes an exogenous nucleic acid encoding RNA-binding zinc finger domain).

The term “treating” means a reduction in the number, frequency, severity, or duration 25 of one or more (e.g., two, three, four, five, or six) symptoms of a disease or disorder in a subject (e.g., any of the subjects described herein), and/or results in a decrease in the development and/or worsening of one or more symptoms of a disease or disorder in a subject.

The term “promoter” means a DNA sequence recognized by enzymes/proteins in a 30 mammalian cell required to initiate the transcription of an operably linked coding sequence (e.g., a nucleic acid encoding a fusion protein (e.g., a RNA-binding zinc finger domain and a fusion partner)). A promoter typically refers, to e.g. a nucleotide sequence to which an RNA polymerase and/or any associated factor binds and at which transcription is initiated. The promoter can be constitutive, inducible, or tissue-specific (e.g., a brain-specific promoter).

The terms “identical” or percent “identity,” in the context of two or more polypeptide sequences, refer to two or more sequences or subsequences that are the same or have a specified percentage of amino acid residues, e.g., at least 60%, at least 65%, at least 70%, at least 75%, at least 80%, at least 85%, at least 90%, or at least 95% or greater, that are

identical over a specified region when compared and aligned for maximum correspondence over a comparison window or designated region, as measured using a sequence comparison algorithm or by manual alignment and visual inspection.

For sequence comparison of polypeptides, typically one amino acid sequence acts as a reference sequence, to which a candidate sequence is compared. Alignment can be performed using various methods available to one of skill in the art, *e.g.*, visual alignment or using publicly available software using known algorithms to achieve maximal alignment. Such programs include the BLAST programs, ALIGN, ALIGN-2 (Genentech, South San Francisco, Calif.) or Megalign (DNASTAR). The parameters employed for an alignment to achieve maximal alignment can be determined by one of skill in the art. For sequence comparison of polypeptide sequences for purposes of this application, the BLASTP algorithm standard protein BLAST for aligning two proteins sequence with the default parameters is used.

#### 15      **Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)**

As used herein, the term “CRISPR” refers to a technique of sequence specific genetic manipulation relying on the clustered regularly interspaced short palindromic repeats pathway, which unlike RNA interference regulates gene expression at a transcriptional level. The term “gRNA” or “guide RNA” refers to the guide RNA sequences used to target specific genes for correction employing the CRISPR technique. Techniques of designing gRNAs and donor therapeutic polynucleotides for target specificity are well known in the art. For example, Doench, J., et al. *Nature biotechnology* 2014; 32(12):1262-7 and Graham, D., et al. *Genome Biol.* 2015; 16: 260. The term “Single guide RNA” or “sgRNA” is a specific type of gRNA that combines tracrRNA (transactivating RNA), which binds to Cas9 to activate the complex to create the necessary strand breaks, and crRNA (CRISPR RNA), comprising complimentary nucleotides to the tracrRNA, into a single RNA construct. Exemplary methods of employing the CRISPR technique are described in WO 2017/091630, which is incorporated by reference in its entirety.

In some embodiments, the single guide RNA can recognize a target RNA, for example, by hybridizing to the target RNA. In some embodiments, the single guide RNA comprises a sequence that is complementary to the target RNA. In some embodiments, the sgRNA can include one or more modified nucleotides. In some embodiments, the sgRNA has a length that is about 10 nt (*e.g.*, about 20 nt, about 30 nt, about 40 nt, about 50 nt, about 60 nt, about 70 nt, about 80 nt, about 90 nt, about 100 nt, about 120 nt, about 140 nt, about 160

nt, about 180 nt, about 200 nt, about 300 nt, about 400 nt, about 500 nt, about 600 nt, about 700 nt, about 800 nt, about 900 nt, about 1000 nt, or about 2000 nt).

In some embodiments, a single guide RNA can recognize a variety of RNA targets. For example, a target RNA can be messenger RNA (mRNA), ribosomal RNA (rRNA), signal 5 recognition particle RNA (SRP RNA), transfer RNA (tRNA), small nuclear RNA (snRNA), small nucleolar RNA (snoRNA), antisense RNA (aRNA), long noncoding RNA (lncRNA), microRNA (miRNA), piwi-interacting RNA (piRNA), small interfering RNA (siRNA), short hairpin RNA (shRNA), retrotransposon RNA, viral genome RNA, or viral noncoding RNA. In some embodiments, a target RNA can be an RNA involved in pathogenesis of conditions 10 such as cancers, neurodegeneration, cutaneous conditions, endocrine conditions, intestinal diseases, infectious conditions, neurological conditions, liver diseases, heart disorders, or autoimmune diseases. In some embodiments, a target RNA can be a therapeutic target for conditions such as cancers, neurodegeneration, cutaneous conditions, endocrine conditions, intestinal diseases, infectious conditions, neurological conditions, liver diseases, heart 15 disorders, or autoimmune diseases.

As used herein, a “CRISPR-associated protein” can refer to an enzyme that uses CRISPR sequences as a guide to recognize and cleave specific nucleic acid strands that are complementary to the CRISPR sequence. A CRISPR-associated protein can associate with a CRISPR RNA sequence to bind to, and alter DNA or RNA target sequences. In some 20 embodiments, a CRISPR-associated protein can be a Cas9 endonuclease that makes a double-stranded break in a target DNA sequence. In some embodiments, a CRISPR-associated protein can be a Cas12a nuclease that also makes a double-stranded break in a target DNA sequence. In some embodiments, a CRISPR-associated protein can be a Cas13 nuclease which targets RNA. Additional CRISPR-associated proteins within the scope of the 25 disclosure as identified by the novel method presented herein also include SEQ ID NOs: 1-50.

As used herein, a “CRISPR-associated array” can refer to a component of a CRISPR-Cas system, wherein a CRISPR-associated array can include alternating conserved repeats and spacers that are transcribed into a precursor CRISPR RNA and processed into individual 30 CRISPR RNAs. In some embodiments, a CRISPR-associated array includes between two and several hundred repeating sequences separated by unique spacers. Both the repeats and spacers in an array have interesting features, wherein each DNA repeat is a partial palindrome while spacers all share a common sequence called a Proto-spacer Adjacent Motif (PAM) that Cas9 requires to recognize its DNA target. In some embodiments, a

CRISPR-associated array has a 20 kb flanking region either at the 3' or 5' end of the CRISPR-associated array. In some embodiments, the CRISPR-associated array has a 20 kb flanking region at both the 3' and 5' end of the CRISPR-associated array. In some embodiments, a flanking region can include a coding sequence. In some embodiments, a flanking region can include a plurality of coding sequences. In some embodiments, a flanking region can include three or more coding sequences.

### **CRISPR/Cas system**

Provided herein are non-naturally occurring CRISPR/Cas systems including (a) a guide RNA, wherein the guide RNA comprises a repeat sequence and a spacer sequence capable of hybridizing to a target nucleic acid; and (b) a CRISPR-associated protein or a nucleic acid encoding the CRISPR-associated protein, wherein the CRISPR-associated protein comprises an amino acid sequence that is at least 80%, at least 85%, at least 86%, at least 87%, at least 88%, or at least 89% identical to a sequence selected from SEQ ID NOS: 1-50.

In some embodiments, the CRISPR-associated protein comprises an amino acid sequence that is at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, or at least 99% identical to a sequence selected from SEQ ID NOS: 1-50. In some embodiments, the CRISPR-associated protein comprises an amino acid sequence selected from SEQ ID NO: 1-50.

In some embodiments, the CRISPR-associated protein is capable of binding to the guide RNA and of targeting the nucleic acid sequence complementary to the guide RNA spacer sequence. In some embodiments, the target nucleic acid is an RNA or DNA. In some embodiments, the targeting of the target nucleic acid results in a modification of the target nucleic acid. In some embodiments, the modification of the target nucleic acid is a cleavage event. In some embodiments, the guide RNA further comprises a trans-activating CRISPR RNA (tracrRNA).

In some embodiments, the system is present in a delivery system. In some embodiments, the delivery system comprises a delivery vehicle selected from the group consisting of an adeno-associated virus, a nanoparticle, and a liposome.

[Table 1]

SEQ ID NO:	Protein ID	Amino acid Sequences
SEQ ID NO: 1	gene_5155 455	MTPYSIGLDIGTNSVGWAVITDNYKVPSSKKMVLGNTSKYIKKNL LGVLFDSSGITAEGRRRLKRTARRRYTRRRNRILYLQEIFSTEMATLDD AFFQRLLDDDFLPVDDKRDSKYPIFGNLVEEKAYHDEFPTIYHLRKYLA DSTKKADLRLVYLALAHMIKYRGHFLIEGEFNSKNNDIQKNFQDFLD TYNAIFESDLSLENSKQLEEIVKDKISKLEKKDRILKLFPGKNSGIFSE FLKLIVGNQADFRKCFNLDEKASLHFSKESYDEDLETLLGYIGDDYSD VFLKAKKLYDAILLSGFLTVTDNETEAPLSSAMIKRYNEHKEDLALLK EYIRNISLKTYYNEVFKDDTKNGYAGYIDGKTNQEDFYVYLNKNLLAEF EGADYFLEKIDREDFLRKQRTFDNGSIPYQIHLQEMRAILDQAKFYP FLAKNKERIEKILTFRIPYYVGPLARGNSDFAWSIRKNEKITPWNFED VIDKESSAEAFINRMTSFDLYLPEEKVLPKHSLLYETFNVYNELTKVRF IAESMRDYQFLDSKQKKDIVRLYFKDKRKVTDKDIIYEYLHAIYGYDGI ELKGIEKQFNSSLSTYHDLLNIINDKEFLDDSSNEAIIIEIIHTLTIFEDRE MIKQRLSKFENIFDKSVLKLSRRHYTGWGKLSAKLINGIRDEKSGNT ILDYLIDDGISNRNFMQLIHDDALSFKKKIQKAQIIGDEDKGNIKEVVK SLPGSPAICKKGILQSIKIVDELVKVMGGRKPESIVVEMARENQYTNQG KSNSQQLRKRLEKSLKELGSKILKENIPAKLSKIDNNALQNDRLYLYY LQNGKDMYTGDDLDIDRLSNYDIDHIIPQAFLKDNSIDNKVLVSSASN RGKSDDFPSLEVVKRKTFWYQLLKSKLISQRKFDNLTKAERGGLP EDKAGFIQRQLVETRQITKHVARLLDEKFNSNKKDENNAVRTVKIIT LKSTLVSQFRKDFELYKvreINDFHHAHDAYLNAVIASALLKKYPL EPEFVYGDYPKYNFRERKSATEKVYFYSNIMNIFKKSIISADGRVIER PLIEVNEETGESVWNKESDLATVRRVLSYPQVNKKVEEQNHGLDR GKPGLFNANLSSPKPKPNSNENLVGAKEYLDPKKYGGYAGISNSFAV LVKGTEKAKKKITNVLEFQGISILDNRINYRKDKLNFLLEKGYKDIELI IELPKYSLFELSDGSRRMLASILSTNNKRGEIHKGNNQIFLSQKFVKLLY HAKRISNTINENHRKYVENHKKEFEELFYYLEFNENYVGAKKNGKL LNSAFQSWQNHSIDELCSSFIGTGSERKGLFELTSRGSAADFEFLGVK IPRYRDYTPSSLKDATALIHQSVTGLYETRIDLAKLGE

SEQ ID NO: 2	gene_3815 793	MSIRSFKLKIKTKSGVNAEELRRLWRTHQLINDGIAYYMNWLVLLR QEDLFIRNEETNEIEKRSKEEIQGELLERVHKQQQRNQWSGEVDDQTL LQTLRHLYEEIVPSVIGKSGNASLKARFFLGPLVDPNNKTTKDVSKSG PTPKWKKMKDAGDPNWVQEYKMAERQTLVRLEEMGLIPLFPMY TDEVGDIHWLPQASGYTRTWDRDMFQQAIERLLSWESWNRRVRERR AQFEKKTHDFASFSESDVQWMNKLREYEAQQEKSLEENAFAPEPY ALTKKALRGWERVYHSWMRLDSAASEEAYWQEVTACQTAMRGEFG DPAIYQFLAQKENHDIWRGYPERVIDFAELNHLQRELRAKEDATFTL PDSVDHPLWVRYEAPGGTNIHYDLVQDTKRNLTLILDKFILPDENG WHEVKKVPFSLAKSKQFHRQVWLQEEQKQKKREVVFYDYSTNLPHL GTLAGAKLQWDRNFLNKRTQQQIEETGEIGKVFFNISVDVRPAVEVK NGRLQNGLGKALTVLTHPDGKIVTGWKAEQLEKWVGESGRVSSLG LDSLSEGLRVMSIDLQRTSATSVFEITKEAPDNPYKFFYQLEGTELF AVHQRSFLALPGENPPQKIKQMREIRWKERNRIKQQVDQLSAILRH KKVNEDERIQAIKDLLLQKVASWQLNEEIATAWNQALSQLYSKAKEN DLQWNQAIAHHQLEPVVGVKQISLWRKDLSTGRQGIAGLSLWSIEE LEATKKLLTRWSKRSREPGVVKRIERFETFAKQIQHHINQVKENRLKQ LANLIVMTALGYKYDQEKKWIEVYPACQVVLFENLRSYRSYERSR RENKKLMEWSHRSIPKLVQMCGELFGLQVADVAAYSSRYHGRTGA PGIRCHALTEADLRNETNIIHELIEAGFIKEEHRPYLQQGDLVPWSGGE LFATLQKPYDNPRILTLHADINAAQNIQKRFWHPSMWFRVNCESVME GEIVTYVPKNKTVHKKQGKTFRFVKVEGSDVYEWAKWSKNRNKNT FSSITERKPSSMILFRDPSGTFKEQEWEVQEWTFWGKVQSMIQAYMK KTIVQRMEE
SEQ ID NO: 3	gene_2964 877	MNKAADNYTGGNYDEFIALSKVQKTLRNELKPTPFTAEHIKQRGIISE DEYRAQQSLELKKIADEYYRNYITHKLNDINNLDFYNLFDAIEEKYKK NDKDNRDKLDLVEKSKRGEIAKMLSADDNFKSMFEAKLITKLLPDYV ERNYTGEDKEKALETLALFKGFTTYFKGYFKTRKNMFSGEGGASSIC HRIVNVNASIFYDNLKTFMRIQEKGDEIALEELTEKLDGWRLEHIF SRDYYNEVLAQKGIDYYNQICGDINKHMNLYCQQNKFKANIFKMMK LQKQIMGISEKVFEIPPMYQNDEEVYASFNEFISRLEEVKLTDRLRNIL QNINIYNTAKIYINARYYTNVSTYVYGGWGVIESAIERYLCNTIAGKG QSKVKKIENAKKDNCFKMSVKELDSIVAELYEPDYFNAPYIDDDNAVK VF GGQGVLYFNKMSSELLADVSLYTIDYNSDDSLIENKESALRIKKQL DDIMSLYHWLQTFIIDEVVEKDNAFYAELEDICCELENVVTLYDRIRN YVTKKPYSTQKFKLNFAASPTLAAGWRSRSKEFDNNAIILLRNNKYYIAI FNVNNKPKDQIIKGSEEQRLSTDYKKMVYNLLPGPNKMLPKVFIKSD TGKRDYNPSSYILEGYEKNRHIKSSGNFDINYCHDLIDYYKACINKHP EWKNYGFKEETTQYNDIGQFYKDVEKQGYSISWVYISEADINRLDE EGKIYLFEIYNKDLSSHSTGKDNLHTMLKNIFSEDNLKNICIELNGNA ELFYRKSSMKRNITHKKDTVLVNKTYINEAGVRVSLTDEDYIKVYNY YNNNDYVIDVEKDCKLVEILERIGHRKNPIDIICKDRYPTEDKYFLHLPITI NYGVDDENINAKMIEYIAKHNMMNVIGIDRGERNLIYISVINNKGNIIE QKSFNLVNNYDYKNKLKNMEKTRDNARKNWQEI GKIKDVKGYLS GVISEIARMVIDYNAIIVMEDLNKGFKRGRFKVERQVYQKFENMLISK LNYLVFKERKADENGGLRGYQLTYIPKSISKNVGKQCGCIFYVPAAYT SKIDPATGFNIFDFKKYSGSGINAKVVDKKEFLMSMNSIRYINEGSEE YEKIGHRELAFAFSFDYNNFKTYNVSSPVNEWTAYTYGERIKKLYKDG RWLRSEVNLNLTENLIKLMEQYNIEYKDGHDIREDISHMDETRNADFIG SLFEELKYTVQLRNSKSEAEDENYDRLVSPILNSSNGFYDSSDYMENE NNTHIMPKDADANGAYCIALKGLYEINKIKQNWSDDKKFKENELYI NVVEWLDYIQNRRFE

SEQ ID NO: 4	gene_4147 644	MKLSKEKHTRSAVANNGDIKSAEVNNGNTKSEEVNNGDIRSAVANE EQNIGGILYRFPGKSIDGVKDQMLRRDEVKKLYNVFNQIQVGTKPK KWNNDEKLSPEENERRAQQKNIKMKNWKWREACSKYVESSQRIIND VIFYSYRKAENKLRYMRKNEDILKKMQEAEKLSKFSGGKLEDVFVAYT LRKSLVVSKYDTQEFDVAAMVFLECGKNNISDHHEREIVCKLLELI RKDFSKLDPPNVKGSGGANIVRSVRNQNMIVQPQGDRFLFPQVYAKEN ETVTNKNVEKEGLNEFLLNYANLDDEKRAESLRKLRRILDVYFSAPN HYEKDMITLSDNIEKEKFNVWEKHECGKKTGLFVDIPDVLMEAEA ENIKLDAVVEKRERKVNLNDRVRKQNIICYRYTRAVVEKYNNEPLFFE NNAINQYWIIHHIENAVERILKNCKAGKLFKLRKGYLAEKVWKDAILN ISIKYIALGKAVYNFALDDIWKDKKNELGIVDERIRNGITSFDYEMIK AHENLQRELAVDIAFSVNNLARAVCDMSNLGNKESDFLLWKRNDIA DKLKNKDDMASVSAVLQFFGGKSSWDINIFKEAYKGKKKYNYEVRFI DDLRLKAIYCARNENHFKTAJVNEKWNTELFGKIFERETEFCLNVE KDRFYSNNLYMFYQVSELRNMLDHLYSRSVSRAAQVPSYNSVIVRTA FPEYITNVLGQKPGYDADTLGKWSACYYLLKEIYNSFLQSDRAL QLFEKSVKTLSWDDKKQQRADVNFKDHFSDIKSACTSLAQVCQIYMT EYNQQNNQIKKVRSSNDSIFDQPVYQHYKVLKKAIANAFADYLKNN KDLFGFIGKPKANEIREIDKEQFLPDWTSRKYEALCIEVSGSQELQK WYIVGKFLNAMSLNLMVGSMRSYIQQVTDIKRAASIGNELHSVQD VEKVEKWVQVIEVCSLLASRTSNQFEDYFNDKDDYARYLKSYVDFS NVDMPSEYSALVDFSNEEQSDLYVDPKNPKVNRNIVHSKLFAADHIL RDIYPEVSKDNIEEFYSQKAEIAYCKIKGKEITAEEQAKVLKYQKLKN RVELRDIVEYGEIINELLGQLINWSFMRERDLLYFQLGFHYDCLRND KKPEGYKNIKVDENSIKDAILYQIIGMYVNGVTVYAPEKDGDKLKEQ CVKGGVGVKVSAFHRYSKYGLNEKTLYNAGLEIFEVVAEHEDIINL RNGIDHFKYYLGDYRSMLSIYSEVFDRFFTYDIKYQKNVLNLLQNILL RHNVIVEPILESGFKTIGEQTKPGAKLSIRSIKSDFQYKVKGGTLITDA KDERYLETIRKILYYAENEEDNLKKSVVVTNADKYEKNKESDDQNK QKEKKNKDNKGKKNEETKSDAEKNNNERLSYNPFANDFKLLN
SEQ ID NO: 5	meta_gene _174274	MAKKNKMKPRELREAQKKARQLKAAEINNAAAPAIAAMPVAEAAA PAAEKKKSSVKAAGMKSILVSENKMYITSFGKGNSAVLEYEVNNND YNKTQLSSKDNSNIELGDVNEVNITFSSKGHGFESEGVINTSNPTHRSGE SSPVRGDMLGLKSELEKRFFGKTFDDNIHIQLIYNILDIEKILAVYVTNI VYALNNMLGEGDESNYDFMGYLSTFNTYKVFTNPNGSTSLSDDKKENI RKSLSKFNALLTKRLGYFGLEEPKTDRVLEAYKKRVYYMLAIVG QIRQCVFHDLSEHSEYDLYSFIDNSKKVYRECRETLDDYLVDERFDSIN KGFIQGNKVNISLLIDMMKGYEPDDIIRLYYDFIVLKSQKNLGFSIKKL REKMLDEYGFRFKDKQYDSVRSKMYKLMDFLLFCNYYRNDVAAGE ALVRKLRFSMTDDEKEGIYADEAAKLWGKFRNDFENIADHMNGDVI KELGKADMNFDEKILDSEKKNASDLLYFSKMIYMLTYFLDGKEINDL LTTLISKFDNIKEFLKIMKSSAVDVECELTAGYKLFNDSQRITNELFIVK NIASMRKPAAASKLTMFRDALITLGIIDDKITDDRISEILKLKEKGKIH GLRNFITNNVIESSRFVYLIKIANAQKIREVAKNEKVVMFVLGGIPDT QIERRYKSCVEFPDMNSSLEAKRSELARMIKNISFDDFKNVKQQAKG RENVAKERAKAVIGLYLTVMYLLVKNLVNVNARYVIAIHCLERDFGL YKEIIPELASKNLNDYRILSQLCELCDKSPNLFKKNERLRKCVEVD INNADSSMTRKYRNRIAHTTVVRELKEYIGDIRTVDSYFSIYHYVMQR CITKREDDTKQGEKIKYEDDLLKNHGYTKDFVKALNSPFGYNIPRFKN LSIEQLFDRNEYLTEK
SEQ ID NO: 6	gene_4200 106	MPAAEVIAPAAEKKKSSVKAAGMKSILVSENKMYITSFGKGNSAVLE YEVDNNNDYNQTQLSSEDSSNIELCGVTKVNITFSSKGHGLESGVINTSN PTHRSGESSPVRWDMGLGLKSELEKRFFGKTFDDNIHIQLIYNILDIEKIL AVYVTNIVYALNNMLGIKKSESYDDFMGYLSARNTYEVFTHPDKSNL SDKAKGNIKKSFSTFNDLLTKRLGYFGLEEPKTDRVSQAYKKRV

		YHMLAIVGQIRQCVFDKSGAKKFIDLYSFINNIDSEYRETLDYLVDER FDSINKGFIQGNKVNISSLIDMMKGYKADDIIRLYYDFIVLKSQKNLGF SIKKLREKMLDEYGRFKDKQYDSVRSKMYKLMDFLFCNYYRNDV IAGEDLVRKLRFSMTDDEKEGIYADEAEKLWKGFRNDFENIADHMN GDVIKELGQADMDFDEKILDSEKKNASDLLYFSKMIYMLTYFLDGKE INDLLTTLISKFDNIKEFLKIMKSSAVDVECELTAGYKLFNDSQRITNE LFIVKNIASMRKPAASAKLTMFRDALTILGIDDKITDDRISEI
SEQ ID NO: 7	meta_gene_524079	MLQQPYTIDYGSRKTGSKAAAVGDNNYPTFFLDLLIIIGRPLQPITQSN DRFFDNVTVTFKNWRASYSSKHVHGLPFDLHDHTFRLATAATREW YIVMHPTASTIDLPSSRRERRKRLEKSSQSSALQLHHAHFLAGYIKW VFLIDDLGEGVEPSWTINGPHLTKITFNKWTAFQNRFMEEWDSYVQ EYSCDNFWMENQPAFHAYDYGANIEIEIREESELSQLKSLPKETRLR RNNEESESEEEDTNILEDGTQLMSSRSNSREVSEAPEEINYQSLYTEGL RQLRTELERKYILNNISSISYALAVDIGCQDSNSPDPEDKQVYCLLADR NKVLGDFRGPRDFTFYPLAFHPAYGNFSSPGPPSFLIDNVLAVMRDN MSYQNDGADTLSYGYFQAYSNIKRSIRHKPEDLLATKGIAATAALALPE SEANASSHIKAKRQRLLQRLQQATPEDPDSSKPFERERQLIEAAIVAE KFDFRMEQVLTIQVSRLIDSRRNFSTVLNPQLVRFYLMESHRYTHLL RWFPSPVFPGLGSFARIFGLADEIYARFKAGGSKGLSIALAEGVSA RLGSYCFGTGPKSLMGSVLSPLGTIDGIEQGAWPYINPRMLDLQDG SLCLSQWPRGENKRPLLHVASICFYYGPEVAASRHNSNWFKEFGG MSIKGPSGAAKFLEDLFQDLWIPQTVAFVDHQLNRLRQGSGSADKT KEELLLEHQQALIRQWLQSEHPFSWAYVNDRRAVKCCS
SEQ ID NO: 8	meta_crt_a rray_WNG G01011662 .1	MRIPVTQARNNLGGERSWRDMVSPAQRFLQPRSARAPRSLAGSKM PNSPRETRLTHNNFRGSRLLAEHRHDCAGAGMARIRGIARVGLDDYD VKIIPGDDGPGLRDLARHDHGHRDGSRRWRAVARVCGPVHATGG GSRTVLEALERGWSRTRAQRVVLPLRAPEVHQLPDRSRDGANRLVS NDKGAARAEGEEQDAARCVSATAAHVTDVDFDGVGAKRRVRVPA HGEPAAGGNGSRRGGTAVTPVDGRCVVGHRRAVRVSIGEAGHHRIG RDVGAAQGLLAGCRQGGIGDGRARRSGGAVADVVDVGDRRC PLLSVGVRATHREGSTGRTGDGARGGNAAVAPVDRRRVFARRCLGV GICDGGDCAANGYSLSCRDRGSRGLDGRVCGGHPRLGSGGLQC INKGRRDGICPHSATVGVGERDLAVDVRRA DRAAGAHGAHVGP DAEGDGLPRLRAATLEDGRGHGVAAADWVSGRCRSQADVGL AAAEQDVWRARDGSPSVDVRRVGLGHGRVCSQCQHCP TREVVE EGMHRAAPVGQVGVEPGRPGRGDDGAA GTGGAPRCEL MRSVG AQSVHDWHRGSRWSSAIPVGVRGAELATRV DVVGAGRACHCT REGNAF SLEEI ASPGAACRARV VEGRVGADQNL VASTGH HRLPDGGV GLSGG VGLVALAGRI ADDERL IGGELACRHS VPLGV SRQGDGEA VGLS LLL QALPA AIGSG GARY AASQ DEHL GGGSMR VSDGA*
SEQ ID NO: 9	meta_gene_336895	MEIIDKVNANSFYKMRDKFLYSSDVRENIALRDNVFAPIFILCEVNEIR NFDGERNDKLSYFEMKLN YETKEIDLG SNYRSL VDRIK VIKKEMKIFY EEMLKNERD DVDS FINS NKEIKE KLID FIKE KEEFF EMS DDDF FIDLY KR FYRL LYC IN DEQS KLI I GENIK REINF YRN IVNT DKT KTR LLE KKYN VED DPY LVL F SFY NG IS DY EY S I F N I G ML K REVN K LF ML K LN N EF YN LL IDMH VMS MEE ILF SEN NL GI KY SE LET MV K Y SL H DR I GIVE N AD R LII KE DE KT KE KT E NT T E NI Y K FN N L E YY D K V K T L D L I N E F K I S E R T I N R G F I Y L K L S E K D F Q F E G T L E L F E I R L G L K I Q K R E K L S F T E E D I K S D K V E S V S K I L N S N D S N K Y F D G N T T F G L P R N Q V D F N V F R K F K E Y F S S N R S R K F T G Y D I K I N R D R T R E K N I P I L K R I N K S A Y C E K I R V N P L T G E I S N N K Y N K S L E D V M L H T Y L Y I Y M Q S L F L M V K T R L K N N G E F K T L D L F N F E F M G L I N N I V P D T R H G L F R Y I N Y Y F E K Y E P F I E N K I K Y K P I T K D G K I K N E I I N N M E D Y I L A K I S E F I Y L Q I L H N F S L E N I I D I E I A T N E E I K S I L N L Y N D V E L K K E S K Y I K T I M N Y Y A E F L E T F N N E I K I K E E N

SEQ ID NO: 10	meta_gene_321445	MVGEYLGTTFEKAIEQKPITLTKVDidKSIINDYKEINDFILAKKSTVS LLNEDNKKMVEYAKIHGIDTKEILKEIKSLHKAENKELEKDMKSSELD KNYAWYLENKENKAIKDVLETKKNTLSEWSKEIGNLESDEKLAYLK GTNDVNFKNEIYNFSDEKIKKVDIYTNFKEEYTANQKEYFNNNGVEI EKEEEKEKNFHISINEINQNYLSDKEKINDILSVLKITVEDIEKTEEQIKK RNPDLNDKTIADMITNHIYDNMIKENVALIDFSKKEFFNFGNENITKE MELERYFNLKEKNSIESFDEKDLVSFDKVDYLIKDRDENIINNERQYLLS NELKEHQDINYQAKKEIALILTNTNALDREIKAEMLEFKDDKLIVITPE YNAEKTEKLTEENNIDKIRNIILNGIENKTSLNDMEKLNKNEILFNLG QRDKLREFGSFYKDMIFDKENEIIKEKVEIVKEKYLESDLREKVLER AKELGITLEKEPVFTTKSITKDMEDAEPNIDKDNEVTINYFENRNDIYQ FFENSPLYLKNALRIYEDTLDFDYKEIYPYNNLEENAKFIVANILEIPEI KNNKEEFGIDSINLHEYLKEMSFDDLEWINNIDETIKEVIENKVEKDD DYVPEVTDKTEDISNNNEDENKRIENDKEKNKEKDDEYNF
SEQ ID NO: 11	gene_3820 393	MIEAPGDPVERQFDEWLTRWSRWAEPEAARRSRETLRRELAAAARQ LDLHSDTQELILGVGLLCWRSPRGDEVFRHLLTAPVQIVVDKQTGRV GVHLRDEGELEALEDQYFLTEQDGYVASRVEPLRGALSEVSDPLDDQA KALLHKWASHGLETPCKFEPVVWSTPETGGPHALVSLSPALVLRHRSS NRAEFYQGIHASLSDPEGVAPLGLWAQLMFPMEPEERLAWRATRG TAGTSRLSEEPLFPLAMNDEQRLAFDKLSKDTVLVIEGPPGTGKTHTI ANLMSALLAEGKRVLVTSA RDKALNVLFDDGMLPKPLQRLCVRLDD QRGNRKGELTRSVTALSASAERSKEEILERARMLTDRRSELKREISL VHRQLWELEAETTDLGEVAPGYRGRADIAERVADTASTHWIGIM PDSAAPVPLNSQEAQELAQLLRTPAQNDQPLTLRAGNPPPTDEFTAL VSAAHQTLPASGVGARLAERLSTLDEGAFRTVS AFWELACNALQGLR LPGDTASWSSIDWQGTAALSLQGGDVS AWKHLWEATRTAAPHAE LARLTGRYLQIPALHGAGAAEAASAAEAYSRFKAGGRPGKIKKSPE QRMAERSLAECFVDGRRPSTVADFMDLTTALRAVAVLSGLSNRWRR SGVKTNTPDNVSQNLEALVGREADLAHLIRFAE ALES LHQHLPDRSAI HASGSWDWPALVEGFTAAPAHMKSARARRNLDLSRARIADADHPLF REMTTAVERRDLAAYTTAEIWTQAHSQLRAERRSELVDRVAAVH PALAHLR LATATMDDD WTSRLETLDEA WA SAAA AVSSRSVESTAE LQRELD RLED ALMKT TAELASEQ AW WHCLQRMSV REASALRSFARE MKRVGRGKGRYAGRHRQGAREAMRLARDAVPAWVMPVRQVAETI DPRPDAFDVIIIDEASQLPVESAFLWLAPQVIVVGDDKQCSPPMRVS GELEPIYERIEEYLPDVPRAFRHD LTPKSNL YELMNVRFP GGQRLTDH HRSMPEIIAWSSRMFYDGSLTPLRQYGTDRLLPPLRVVDVPDGYREGR DQNVRNPPEAEKLVTELKAMIEDPAYSGKTFGIISLQGGERSGHIRLIE QLLDEHPDQALRERLKRIVGT PPDFQGDQRDVILLSMVATGTPRIQG GADFEQQRWNVAATRARDQMVLFASTTLTQLKSDDLRASLLKML DTPMRETTPQHLLHVEPQTKHPEFDSLFEQKVFLKIRERGYEVVPQYP AGRNMRIDLVIVGEKRLAVECDGRYWHSGAKQVQDLLRERILRR AGWTFWRLRESDFLLDPVSLRPLWALLDRIGHPAKGQ
SEQ ID NO: 12	meta_gene_180752	MAQFNFTKKLDIDETQIEQTDVMTGDNRRNRYLYYQLKLSMLHAKK IDIIVSFLMESGVRLILNDLKTALDRGVQIRILTGNYLGITQPSALYLLK NELGNRVMRFYNDKHSRSHPKAYIFHYENYEDIYIGSSNISRSALTS GIEWNYRLNSQDNHKDFVLFYDTFQDLFENHSIIIDDNELKRYSKNW HKPAVSKDLARYDAVEDNSDTPVRKLFQPRGPQIEALYALADSRSEG ATKGLVHTATGIGKTYLAAFDSAKYQKVLFVAHREEILKQAAISFRN VRQSNDYGFY GKQKDKDKSVIFASVATLGRSEYL TENEYFAPDYFDY LIIDEFH HA VNDQYQRIINYFKPKFLLGLTATPERLDGKDIYEICDYNV PYEISLKEAINKGVLVPFHYYGIYDTVDYSSILVRGHYDEKQLDKAY IGNKDRYDLIYKYYKKYPSKRALGFCCSRKHANEMAKEFCARGIDAV AVYSNTNGEPSEERNIAIQKLKSQEIKVIFSVDMFNEGVDIPLDLMVM FLRPTEPVVFLQQLGRGLRISKGKTYLNVLDFIGNYEKAGRVPLLL

		GGGDSNKNAPTDLSSIEYPDDCIVDFDMRLIDLFKKLDQKSLTAKERI THEFYRVKEKLDGKIPTRMQLFTYMDDVYRYCITHAKENPFRHYLE FLEKLHELSEETEETLCISGLGKDFLTLIETTDMQKVYKMPILYSFFNHG NVRLAVKDDEVLAWKDFFTNTGKNWKDFAADITYDEYKSITDKQHL RKAKSMPIKYLKASKGFFVEKDGFAIRDDLKDIVKNDAFIKHMH DILEYRTMEYYRRRYLEKI
SEQ ID NO: 13	gene_7714 18	MRRNPEFTFFSHKNVPEVSGYEGGLVNSTMNSLHTSPTLGIDIGSTTV KVALLDAEHNILFSDYERHYANIQETLAELLRKAREKAGPMEVVSVIT GSGLALSHLQVPPVQEVVAVASALQDYAPKTDAIELGGEDAKII YFSGGIDQRMNGICAGGTGSIDQMASLLQTDAAGLNDYARHYKAIY PIAARCGVFAKSDIQPLINEGATREDLSASIFQAUVNQTISGLACGKPIR GNAFLGGPLHFLPELRNAFIRTLHLTGSQIAPDNSHLFAAIGAALNP QEQTSSLLSMIERLSSGIKMDFEVKRMPLFRDQADYDEFDRRHA GHQVKTGDLARYSGNCYLGIDAGSTTKVALVGEGGELLYRFYDNN NGSPLATAIRAMSEIREILPPTAHIAWSCSTGYGEALLKSALMLDEGEV ETISHYYAAAFFEPDVDCILDIGGQDMKCIKIKDGTVDSQLNEACSS GCGSFIETFAKSLNYSVEDFAKEALFAENPTDLGTRCTVMNSNVKQ AQKEGATVADISAGLAYSVIKNALFKVIKITRPSDLGRHVVVQGGTFY NDAVLRSEKISGCEAVRPDIAGIMGAFGAACIARERWHMQPADSGR ETSMPLLDKITSKYTTSMTRCKGCNNHCVLINQFGSRRFISGNRC ERGLGIEKSKEIPNLFDYKYHRMFYTPLPLDKAHRGVVGIPRVLN MYENFPFWAVFFERLGYHVTSPQSTRQLYELGIESIPSESECYPAKLV IGHISWLIKQGVKFIFYPCIPYERNETPDAGNHYNCPMVTSYAENIKN NVEELAEEHVNFNMNPMAFTNEEILTAKVAEFANAFDIPAAEVRA AHAGWEELLQSRRDMEAKGEEVLDWLKQTGKRGIVLAGRPYHVDP EIHHGIPELITSYGFAVLTEDSVSHLGKVERPLVVTDQWMYHSRLYAA ASFVKTQENLDIQLNSFGCGDAVTTDQVSILTRSGKIYTVLKIDEV NNLGAARIRIRSLIAALRVRDQRNFERKVSSAYHRAVFTKEMKKDY TLLCPQMSPHIFDLIEPAIRSGFYKIEVLQNHNRSAVDVGLQYVNNDA CYPSSLIVIGQIMDALLSGRYDLNHTAVFMSQTGGCRASNYIGFIRRA LEKAGMPQIPVISVNANGMETNPGBTITLPLLTAKAMQGVVYGDIFMR VLYATRPYEAEFGSANALHEKWKKRCVASLSKRSSMMEGRNIRGI IRDFDALPLRDVRKPRVGIVGEILVKFSPLANNHIVELLESEGAEAVMP DLMDFLYYCFYNSNFKSKHLLGTCKSTTYLCNAGIALLEYFRRTARKE LEASKHFTPPAAIDEALARMAQGFVSLGNQTGEGWFLTGEMLELIHG VENIICTQPGCLPNHIVGKGVIKELRRHYPQSNIIAVDYDPGASEVNQ LNRIKLMATAQKNLKKGTN
SEQ ID NO: 14	gene_1433 645	MGSSEYGIKSLKNLDGIEHIRLRPRMYTDIGSEIGCHHIAQEVLNCG DEAIGGFCSRITVEIESDHVICISDNNGRIPVETDEASMSGVEMVLTQ DKAGGKFDHDSYQVSGGLHGVGVTVTNALSSFLEATVKRDGGEWF MRLEKGRVIEKLRRAADCGPRTGTSIRFSPDPEIYEQAKFRVQQIRQ QAMDKAIIPIGPLEVIFKAPGLEAERFCFKRGLAEYMEANMADSPVFEF SGALGDVEKVHWFFAAFDEPVDSFIRSYANTVPTPRGGTHEKGFA MLKAVREYLLRPELKTLGKNTIAPSNDVMSQMGLSVYIKDISF EGQTQKQLGSREATKFVGGVIHDAASLLHRDVELSDAWVKMVIDR ASARTALENGKKKVERKSYTGRTPPLPGKLQDCRFNGIEGTEIYIVEG DSAGGSAKQACNRDTQAVIPIKGKILNCEGINQEDAIASEAVADLVA VGSGVGDVCDPANRRYGVVIIMTDADVDGLHIQNLGTFFYRLMKPL IDAGCVYIVQPPLYGVTIGKQKHYAQDQEELDGLKAMALAEEKKISY TRYKGLGEMDPPELAETCMADAENRVLVKVLPNSDKRMDALMTKLM GDDADQRKNLLMGVEIEDAVHLEPVEEPCDVTEDVKELCQPNSYDS GNNKVAPFETVREMRYRGYGLQVVGGRAIPDVRDGLKPVHRRILYA MEMLKLRSRGPTKKAARVVGDVIGKYHPHGDSVYDAMVRMSQPW KMRYPYIHPQGNWGSIDGDSAAAMRYTEARLTPIAEAMLSTDLKEGI SEYQPNYDDEDIEPLLPAFPFPSVLMNGTTGNPGVGFKSEIPPHNLTEL

		MGACIALADKRIRTGEAESPQDFASVRKHITAPDFPGGGIAGSHDDLE KMYASGRGKMLLRSKWHVEKLERGAWQIVITEIPYGIEKSPLLISMG QCISDPTLPERKRLPMLEDIRDSEGTDIRIILYPKSKGLDPHDIMLHLF SVTNLQVTIEYASYALEDWVLAPNGDRYRLPRLFALDQMIRSLNNR EQIVTARSTVRLAEIEKRLHILDGLLLAYPNIRDIVEIILENDEPKPIIMK KYALSDPQVVAILAIRSQLRKLEEMKLQGEHNQLSAEAVELRQTIDD YTHRKKIKKELQHVRKTFGDERRTEVDPDAARARIMSKEQLVARE PVAVLSKAGWLKGMRGSNIDVENVKFREGDTILDHAAGHTTSRVV LIGRTGRAFNMLAADLPSGRGNGEPISKNFISIDEAPTRLFMINPDAE YMVVTTLGHAFRAKGEDMLTANKKGKAFINFPTGSKLLCIREIDPGH DAIAFITDDGCLGIVKLDEFPLLAKGKLTAVTMKKGVKLLRDAAPV NTSAAVRVGTEKRSTAFEPDEQAETYIIERGRPARPLPKACVNGMLII
SEQ ID NO: 15	gene_4426 209	MKEMNKSETKSSKLLGIVLFHSFIPGKLFKVKAIGHSNNTGDNGAGKS TLSLLPAFYGADPSKLVRDQADKVSFVDYLYLPTPKSVIVFEYEKLGE RKCSVMYRNGSSVAYRFLTGTAEQLFSQHLYDELIKQGSETRTWLKN LVSQSMSVSSQIETSVDYRSVILNNKKRLAQRRSTGKNLVAIAHEYSL CSASHNMNHIDILTATMMRHKKMLSFRKTMIVDCFLNNNTSMDDVPY KKEYSELINSLDVFVQLETKKSKFDEALANKDSLEEYIKQLNSYRAQI ASYLHQLALSQTSQDKIRSQKEQHEILVNERKGKLHTFNSELNNQRI EFERKSKIIDAIYNKRDKYENEDDILGKITLYNSLSDMLREVESARKHY DNLLEDVRTEETELKSQVQKLELECSDFRFRKQQEINSVLKAKEEIVE QKSERLEAMQSDLNFEKKKLQDAFDEQSERIKQEQLRLATLEGQLSD FTSEQKAELRILENDLDDRREFNASQNTVIYLNEQLRQATKTHEGSL SAYHACRDELKEISDEIISVSRAALLPTK GSLNEFLEQKVPGWRCNIGKV IDPNLLNSKNLKPFDFLDTTESMFGHLHDLDISLSPDFCLSEEKLSERLS TLKIKELETETREEAKSRAKSDEQETEKLQKEVKIQSQRSKVLEDEL SKLNLLKDQKTAQFESDAESRTYEVKKQKSVLESEFFAIKSELKAKLE NEEQRHQQERVQVKANFDYRLSEEDAKSAIEALIKDKEKVTSDRISD CKLAFNQALMNKGVDPVSIESTAKLKWELLERQCEEIAFKQALIIDYHT WLEAEWKYIDTYNSEKLDLERQIARGVAKRDDYEKSVGRKIDDVAT SIKLDEQELITVKEAIGQLTTCTNLEKAVDESDLASLEDVSVDFH VEHAVSLVTDKITAITNLKKEIVSKVKDVSNTILGLDDNNEIKMMWE QMRSATMTKLSKDQDYAINYDSPQFLACLGDEGLVNVIPDVRDV KIETLRSISTQISNYHQLKQVNSKVDSVSSTLDKSIETGNPFPASIHI KLSSKIHTFDLWKDLNLFSEVLDRWSGETSRGPLPSKAFLASFQKL FKEAQISKNLESLVEMEITIVENGRPAVVRNDEDLEKVGSEGISKL VVFCGMTRFLCQDEDVAIHWPLDELGKISISNLAILFDMMAQKGICL TAQPDLHPATYKYFATKNHVKNVGVKSFIGGRRSRVNPLLSESKLNQ STEVVE
SEQ ID NO: 16	gene_5411 831	MNKNLKKFAIEARQELREKTKAQLKRLGIEEKKIEEGKDMGSQVEIY GKLYSKSSYQHLLVKYHSLGYEELVEESEAYLWFNRRTALAYMELHD CFTEHMIFSKGNKGEPDILDEYFQADFFQKMPLEKQEEHLQLRDKNTS DSLETLYSILMEEKCEELSKIMPFLFSKKGKYADILFPSGLLMQDSVLK KLQVILLEIQEEDQSSIPVEILGWLYQYYNSERREVYDGSMKKS EFIPAAATQLFTPWIVRYIVDNTVGR LAEEQFSISKDIKKWQYYIAPEI VSKNEKMQIESLKILDPMAGSGHMLTYAFDILFDVYQELGWSK KESVLSILQNNLYGLEIDDRAGQLAAFALLMKGKEFPRLFQV LEREENFEMPVISLQESNAISKRMYTMLEECP TLQDLLKGFEDTKEYGSILKIDSFE ESILQEYHKLQEKIQNQGQFSLLNNNEFLEG DLEEDLERLEHII RQYKIMIQKYDV VITNP PYMGNAR MNP KLK TYIEK YYP NVKT DLF SVFF IKC CEMT TEKG YLG FMSP VWM FIKS YEEL RTL FIH SKTI ISLV QLEY SGFE DAT VPICT FIL QNT VIKK IGEY IKL SDFK GVKN QPIK TLE AIQN ENCTW RYQAN QKD FTK I PG S PI AY WV SDR IRE IFE KE KKL GEV GDA KVG LQT G DNN KF V RL W HE INF N KIG FG MQ N SEE ALK SK KK W FP Y NG QEY VV N WER D G YE I K H F C D T N G K L R S R P Q N T E YY F KK S I W G L I

		TSSGSSFRFYPEGFIYDVGMSYFIEDKFLTYLGILNTKIYSKLTKLINP TINLQIGDILNLPVANIQNPLFEQLVSLILWISFEEWASRETSWDFERLT LLNGENLSKAYKKYCTYWESKFFSVHSSEEDLNRLLESYSLQEEMDE KVDFSDITLLKKEASIVENTDSAASCGYLENRGVRLEFHSLVKQFL SYAIGCIMGRYSLDKPGLIMANSDDVLTMSSNKITVSGVNGAIRHEIL NPSFFPEEFGLSVTTEERFENDVVSRIAIFISAAYGKEHLAENLEFITE VLGKKAGESHEEVLRNYFIKFYTDHCQRYQKRPIYWMLHSGKKNG FSALIYLHRYEKDTIARMRSDYLLPYQEFMEQQEAHYSKIASDEISTPK EKKDAQKKVKELEHDILKELDYANKVKHIAEQRISLDGGVKVNY EKLGSILKKI
SEQ ID NO: 17	gene_9417 61	MALKGDKLLCTNFEFLKVKEFTSFSDACIEAEKSILVSPATTAILSRR ALELAVKWVYSFDEELGIPYRDNISSLIHNGSFIELDSEMLPLLKFWIN LGNVAVHTNKTVTREEAILSLHNLYQFINWIDYCYGDDYKEKKFNEN SLLQGEEKRVRPEELKDLYDKLSSDKKLEEIKEENEELRKVITQKRKE NIENYDFNIEEISEFDTRKIIYIDVELLAGWDFNKGDIGEEIELFGMPNN AEKGYADYVLYGDNKGPLAVVEAKRTSRDAKAGQQQAKLYADCLE KQYNVRPVIFFTNGLETYIWDDYNGYSERRIYGFFKKDELQLMIDRRT QKKTLRNIDIKDEISNRYYQKEAITACCEELERRRKLLLVMATGTGK TRTAISLVDVLTRHTWVKNILFLADRTALVKQAKKNFSNLLPDLSLCN LLDSKDNPEESRMIFSTYPTMMNAIDDTKADGKKLFTCGHFDLIIVD ESHRSIYKKYKAVFDFDAYLIGLTATPKDEVDKNTYGFDMENGVP TYAYEFDKAVEDEFLEVYETIEVKSKIMEDGIKYDELSDEDKEEYEK FDKDENIGEEIQSSAINQWLNFNANTIDLVLNKLMEKGLRIEGNEKLGK TIIFAKNHKHAEAIKERFDILYPELGSNYAKVIDNQINYVDSLIDDFSG KDKLPQIAISVMDLTGIDIPEILNLVFFKKIRSKTKFWQMIGRGTRL EDLLGIGQHKDKFLIFDCNNFEFFRMRNPKGKGNLQGTLSERIFNLKL DLVKELQDLRYSDEEYVSHRNELLKYLIEDVNNLNEDSFMVKMNLK YVQKYKNKNEWQSLGANAKDIKEHIAPLISKLNDEFAKRFDILMY TIELANLQGNNAATPRIKSVIETAESLSKLTIPQIQQQKYIIDKVRTTEF WEDVDFELDEVRSALRELLKYLGKTTQKTYTHFEDMIINEESHGA MYNVNDLKNYRKKVEYYLKEHENELAIYKLKNNQLTKQDLETLES IMWQELGTKADYKEFGDMPVNKLVRKMVGLNRNTTNEFLSEFLNN ENLNKQIHFKVLIIDYVVKNGFIDDNRILMEDPFRTVGNLSVLFKD MKEAKSIMGKISQIKENAEKIV
SEQ ID NO: 18	gene_1546 948	MRLIALELENFRQYAHQAQVAFESGVTAIVGANGAGKTTLEAILWAL YGARVLRDDTHTLRLFLWSQGGAKVRVLLEFALGSRRYRVRTPTDA ELAQLNPDGAWLSLARGANAVNRLVEQLLGMNHLQFQTSFCARQKE LEFLGYTPQKRREEISRMLGYERVGAAVEAIGRAERELKASVEGLRQ GVGDPRALEAQLDAVEQALQATEATALHAEQVALQRAVAARDAAARA HYDAQAALREQYQLQLHQQRPLLQNDRQHAERRIDECLAQWEQLKA ACDRYKVIKPDAERYRQLARELEAMEQLAQAAQQRAQLQARLDALG ERRAQLHAERDALLQKQAHLDALQPQRARAEQLARELQTLRHIARQ AAQRAQLEAQLQIAEQRQRLHALATERDALAQQAQRAEADLHARH TACAQTEAELQQTLQAWSQQRADLDAQLRAVQTTLQQQRARVQQL EALGESSECPTCGQQLGDAYQRVLTAQQQEAQATERELRALRQQRA LEQEPDAIRTLRQQLAQQQQARDDAQRQLAEQARLQLDAELRQT AALEHQQRDLEQRLAQIPPYDPEAEQRAQAEQDALQPALQQAHALEG ELRRLPAIERELSQTTERAQRIQRELDRLPDGYDPDQHAALRTQAEQL RPLYEESLQLAPIQQRDALRARIEDAKTALQRVIAQCEHLETQIAQLG YSEAAYQQAAEAYQQAEAQNVTLERSLAARQAELYASQTALRDQLRA QLERLLELQRALREQEHQLRVHSLLRKAMQDFRADLNTRLRPTLAAL ATEFLNALTNGRYSELIDEEYRFTLIDEDEGHRKQVISGGEEDIVNLSR LALARLITERAGQPMSSLILDEVFASLDAERRHNSVMELLNNLRSWFQD ILVISHFEEINESADRCRLVRNNPQTRASEIVEDALPDPATLATAALDD ALAGDEETGLLPPP

SEQ ID NO: 19	meta_gene_15450	MAKKKKTPVAQIEPISLPDEDLAKARA WLEGLNADIAYSQAKRQLAE ACGWERSKSNAVIVALHEEGFMAGEKNYFCNPNAPEPGVVRGARE VSNFTIMLQSDPEVSVPLPYAIHCLPGDVFMLKTVTGNWRVSNFVA RHQTRWVCKLRRGRIRRSGIAQVVPINGFAPVEMQMIDLADVPAE VDLEKAAFEVEFLPESMKPEPYVEIFVRFVKEIGNRFDPLGEIAIASAE YDLPEFSAAALDEAQALPDEVDPKNMGRVLDRLDIPFTIDGEDAR DFDDAVYCARVEDGRTRLVAIADVSHYVKPGAPLDVDAQQRATSV YFPASVVPMLPEKLSNGLCSLNPGVDRLTMCDAVIDPEGRT EAYQF YPAVIHSHARLTYTQVGAMQGEEGLA AVGDRLDDIRALYELFKT LRKARDARHTLDLETKETMAVFDDKGVICEFKVREHNDAHRLIECM LVANVCAADFVIQKKRGALFRVHDAPSQERLETLRTVLKSFNEKLESP TPEGFAELISRTKENEFLQTAILRSMSRACYSPDNVGHYGLQYEAYAH FTSPIRRYPDLLLHRAIKGILSRRIVPQVFDDSSLMVSQRAGLGSR PEAGDGDKPATQAERHSVWERLGILCSAERRADDATRDVMNYLK CDYMLRHGKGRHEAVVTGMIPAGVVALKDIAVDGFHIHSNLGWGY YEFDEKNLTMTSREEMTQVRVGDRVIVRLEEVLENRRMSFVLESNL ERRLIKGGKGGSSRSSRRGSRLYGRQFDPFIDDDDFDELFGQEGDDD WDD
SEQ ID NO: 20	meta_gene_73412	MSVARKTGSQPRALHAADSHDLIRVQGARVNRLDVSVVLPKRRLT VFTGVSGSGKSSLVFGTIAAESQRMINETYSAFVQGFMP TLARP DV D V LDGLTTAIIDQERMGANARSTVGTATDANAMLRLFSRLGQPHIGSPQ AYSFNVASISGAGAVSIERGGQTVKERRSFSITGGMCPRC EGRGA V ND IDLTALYDDSLSLNEGALTIPGYSM DGFWGRIFSGCGYFDPDKPIRKFT KREL RDLLYREPTKIKVDGINLT EGLIPKIQK SMLAKDIESLQPHIRSF VERAVTFTTCPECHGTRLSEAARSSKIAGISIADTCAMQISDLAEWL G GHYDPSPVAPLLEALRHTVDSFVQIGLGYLSLERPSGTL SGGEAQRIM IRHLGSSLTDVTYVFDEPTIGLHPDIARMNHLLKL RDKGNTV LV VE HKPEMIAJADHVVDLGPGAGIAGGEVVFEGTL DGLRASDT LTGRHLD YRAAVKETVRTPTGALEVRGATANNLREVDVDIPLGVLCVITGVAG S GKSSLVRGSIPAGADVSVVDQGAIKGSRRSNPATYTGLLDPIRKAF AK ANGVKAALFSANSEGACPN CNGAGVIFTDLAMMAGVATSCEVCEGK RFQASVLEYHLGGRDISEV LAMS VAGAEFFGAGEAKTPAAHKIL TH LVDVGLGYLSLGQPLPTLSGGERQRLKLATHLG EKGGVYVLDEPTTG LHLADVEQLLALLDRLVNSGKSIIVIEHHQAVMAHADWIIDLGP GAG HEGGRVVFEGTPAELVAARCLTGEHLAAYVGTGPRKVRTS
SEQ ID NO: 21	gene_3074 07	MQQT LGNEATTRALRGKRPMA PRPAIDERAEQGLVLP PYLMELEA GGLSTAYGLTQEFVSTA VAAVVGHGGGT VAGISAELAGR PESFFGR GRAFAVEGAEGGDGFDTV SIAPAPDLPPTFHPA ADLASAPPDP GG APLAAVDDAEGKETKVDVQHN SGATASSTVGNSSSKGAGGTA FGLA PVPLGLWLGAATGSVQPWQSSRDSRSQRGV AEPV RLRS DKGSVEV PRRVLYVVRVRPQAGGDEQVFRGSGGLTQRV PTEH LIPAGTEAPTLA APASGAPGRSQVDPDLARRVALADSLAPVG VSDTAGPHQGGGLF DAVASVLHPSLTAPGAPGRSRLYEATATPTVLEDPLRLLGGDGV TGD DLYSKDGT SAGSYRMRAV VTGLTPAWGTGKTQLRTHQQAQHTATES AGKGRSVAGGIGPAIGVGAAANAA VVRATAMPVAAARKARFSVNEQ TVSSRQGAEV RGEKVL LGTAQFTVEGTGPRSVRAILNPQARVATHA MRVWIGLRADEARELGLPLPPGVTA GEFIKKPEPQ QPAADADSDT DT DTESESEGGGDARHLPFGAMGSSVTIGRLDTAPMVKA VREMFA TDP RLAGYLPAFGATPPP PADLSREEDEAQRT NYRELMA ALSEANLRVNKEQ LLSTGIRVRLRRKTTMHS HDVQLRVHGT M GATRHLGEIDDWL VRAH SGVAANAQSGRSSRSIGGMVLAQARLIPGVLTGSARYERQSSGTRR NQGGPTTRTDVLTNGSEKASAFGAALRLNVDTM TSQRKLARALT PGGPGRDVPEAKLLTGLHMEEQDVRLLTPSEFTVGTDEKARLDAGAD QAPGP ARPVAGAAGIGDLAGLA PTPAAGQVVRDWQLVETLGDGQPV RDLAL ALLSRAAARGEAGRQDTALATEGLAPR LAVEERFGPRAITAA

		LRQAASSGWVVKNLRYPRRLAALNGAVGTRLALAAPQLVHEAAGPG TETFVMGGHQAGQQGEGTSTVQVGTVQNGTEWRVGEGLSGY RSTSRSDETESATVSGTVERNAHTPKAPLYLVRCDLLVTMVAEVKVT GGGPYVASAARTLPGAAAVWLTAEQLRAAGVDPESARKALKVEDR RPAEAERTAGGSGGERAEEASTAAASTSTVPAPSARASASTATGGQ AASPVRQGPALAREPLGFMIEDLPDFVPLLDRLRGNLAITGQQDLA DDILPRQQLRDRNDNVQRLRVLDRDGSTGLLASAMDGGVTVELLD GRNTPYWAVFKIVRSGDGVRGEADDGRDMYEITSAAAQQATSHGE GETTGVEGILAGSGKPDAGAGQVKSAGAAAGLVASGSRRGGESA RGQLGMKTVAEAKTAKSAKMRVPIVASLELHKDRRLLAGSGRTS LVHRILESDLTALHRSVAPRRAPIPGVPTSGAACGLGAWRAAGVPL PMEEAQANGFQGAHVRELVNTAVRAAGGGDRFRQKGQAAAYTLGE AVSTEWLIAALPLLTNAGAELPPVHASGAAGQDLQASVHARLAGRI LGAGDKMTFETAAQSSLGAPRPTQTEGQSQAESRQARGLFGAGVL NADQFRLNQLMGNVDGAGSASGAAANGAGSMPLHKPKFTSVLVQF TLDVRVVARVTNRVRTSRTEVAERDLTLPRPVIRMPLPVAGRLLAA HPTEITDQHDLGLRAAAVPPPTGV
SEQ ID NO: 22	gene_1432 510	MTTTQKNKPGSLDKKGMSDYTETQCSRQLYIKLGEHDPRWIQRDIQK NTHFTGSALT LAASGKRYEQKVYTILRRLFRQQTHCTLKPPANKEVIE TFLDPRLAKRLHQEV RGEAQLL EYE WPLCDQFVRRVFGQQPDEEIA TLGNQYGRVLRPDIMLLHPIPKGQKAPLKCLLPGGKAASFSP TALQGR FGISILDIKYTPDERVGRRHFAELLFYIHALTEWLHETQLDEF FVPC GHGILGFLEEDTLYDLT LDDLLWRSPDEL SGKHTPKISPLLWEDTHQL FTHAEKTVRTLWQLAKQRTPIEEIPLC VQPACGRCPFIDDCISTLK GTT PTQSDSWDIRLIPYLKTAVAQQLNEHG IYTVGELLQGIEEIP LGNTPV LHAEPALKLRAQALSTQRAVYPEGEHTSLSLPKYIDMALVFNLEV DH TNELVFAFGFYLDTKQPSPKLQRLHNDWWRMWSVLRGERELQDIS SVLDLEALELGWHKGDDFS DKL SLLLQEMERL RTLEADGV LIL RAV GESYQFGSQEYTTQYPLVRCQSYVS GGI EPEHEYML KNM IQQLH RVMRMCSL TELLVTTKHETYDSLYHENFAGFYWSDEQVDHLRAL VE RHLPALQQDHALS KTF YELVDW MTPADSGV RH ALHKKM YDLREF VGSSVGLPQIINYTWHQTRPLWKKD FEA NPYFWT PHFNQMDFGIWHS TIEEIDT NERSQKESDIRDQLV LKM RTLHEILRHFHKEASD VIPK ESKT MSSQDFQRDRRN RQYHQL GSLW QGYHQL NAAISALT ND AARL TWPE QSIAKLQAGKLSGM TIKID DRDGKD YEV VNFS LLG LSS HM KIS VKDR VLLLPRTMRD SHAFPFHNMGRL SKLIVEDL VWE PSEQGYCV TAV REL KKRKEGD KETLHSFT ELYALYDAE DWFV YPT DLDV WT GRLA LNG DA LL RRYQLGYSWLAERLMFLHGLG GEHLEAPKT LNVA HAAE LYTYAPQ LLPQKR DCTGEDV LTPIRFPDSS QQEGILHALSSSISCLQGPPGTGKSQ TIIALIDEFIDRHKGPARI LISA FSYSALQVVVQKLLDSRYGDGPAPD PT QLSDASRLPIFYASSSESES FVHD PNQ QDV MHL SS KG VHL DGERIDF RRGSRKD KIFERMFA HKGLEG DGSF VL FANA HTLYHL GTLS KANK RR LVHEDFGFDL II IDEAS QMP ASY FT AIAQFVHPFEAR LVLPK DED AL KR EIRC GAPELSIEGVPS DDL THV VL VGDQ EQL PPV QQIE PPKL KPMLD SVF RYFLEVHHVPK HQLSY NYR SHK DIVRC VR RLAIYDQL HAF HQDD AYLSAIPDV LPDTIEAPW LRQLL GRRQ VV STLI HGRQ WDTA LSP FEA K LTADVVL AFFAQM GVDS DERER QFW QEDV GVV SPH NAH GRL I VRE I AERLLSGVGARTYLPETELMECLSTTV SVEK FQG SD RRLI VGS VG VS SVDR LAAE EGFLYDMSRL NVL ISRAK HKM LLIC SQQY LDYV PRD RDV MTVAARV REYAYDLCNESQVYDVPFGSGSE FIELRWMVSKDP

SEQ ID NO: 23	gene_5570 191	MQSGSGVDLFRDFNEGEVSEVLRCAGCSRFVLIGPPGSKTFKENVYLEGRLGTGVIVDEYTLGISTTAKIESEEARKGSGISKKAMKYLKRMIPLIEKLRETAEVDEELRKVLGDRAPKHIVEGARRSIGDSPHRAYYIPWKCVDEPNACTFDANVSRALELIKVFDDKKIRWFKAEYVPPGLVKDVIDLIRVKGEDGAREELKGWVEAYSEADETRLRKILGLSDLLEWEESFVEYLSNFVINYASYVISGLVVDPYLIGASALALISVLTYMAFKREGEGYIKGIIELKRGLERLRRSDGEFNELGKLLVYRVAYAMGMSYDEAKEALMDITGLSIDELKRRVNEIEWRIKELEKKIELFRLEVPAGIVTADVNEFAKGRTYPNIKVENGELRIRVEDGYHSIVRAGKFNELVNEVRDGLLKQGFVVVVGPKGIGKSTAAAVIWEFMNSDIGLVARDVLDLNYSLEATFVENYGEKFSEHFGKLLILYDPVSTKAYEKVGIDTEAPIQSNIERTIKNLVNSKSSKASKPFTLIVLPSDVYNALSGEVKNALEGYRLDVSQVLINTEFLAELIREYSKTDKPNGCALSDDVLSQLAGELAKFDGHALIARLIGEELARSNCVGKVEELINSAGKAEAFIILHINGLFKVHENPDTAKALVEIFALRRPFISAVESDDSPDTSKFLVKVYVLRSPFISAVKPGDPILTPGIVELIGEAGGVKILYGAEGEEELRSWLAIWLDLIEEAIGKLLDCIEGKGECKVLDALKPWKTGVIELLRKVSEKVNDVDSAHEYFASNYGERLTSALKVFSNECWKRASYIIGHALAGDPLLPRRKYLSAFMSMNLSTGIESPSDALSRLGANGDKNPQRMSLAKYYASIVESLGDALKECGVDNYLIVGDKIPSMMGLIGNHACALAGVFIDKYNEAIAEIKRLLNIKNGEFYYEEAAYGLLATIIAKAAESGRPVGHSDADAALHIASFAMSHVQSTLHIIRLLTALAPLRDKAPQRYLELVCALDKFTRLGTCHDWDTVMNILNEDYILNKYGVEVKGHARTLVDVINTLTHSLYKCLERCVDYWFEHRVASFRAKFERMISELADLLDKTNRWSPNLGIIAYASLSALDSKNNKNCVRMLIESELGIDVVVNKTKEVAGELSELRGSVRELLRDEDLMGFVRSRLAEADEKAAKRGILEVTSILKHTLAQYKFVNDELDEAGRLFNEAAEESKVIDYLNLDNRDWALRVEAIKSPLAGDDLVKLVNGFRQLYEEALNAERFMSASPDYGTWKNILRDILGGYLVSLALTGGDEEIRRIEELLKEQWQLKYEPRPILTRLTNALLSPRVELSELRDWLVVKPGELIVAFGHGYLIDYLPALKATYGTIKPGDGKRCSSVYLTFMLYALINGNEKLAKAHALMGAMNHSGKLPARLFLEAYRACCDPNNEEFRRAIAKLFYTRALKSKTSGFWASLSS
SEQ ID NO: 24	gene_2435 065	MDRLKTDRKAVQHAEDLGYQVEVLRAKLHEARRALATRPHSYDTADLGYQAEQMLRNAQLQADQMRSDAERELREVRAQTQRILQEHAEQQARLQAELHTEAVNRRQQQLDQELAERRATVESVNENVAWEQLRARSESQAQRLLDESRAQAEQLASARAAEAQRLTEEARRRLGEETENARTEAEALLRRARADAERMLNAASQQAQEAQDTHAEQLRTSTASEADQAHRRSAELTRAFAQRMSEADTALREATSRSEKLVAAEAATAAKRMAAAEAGEQRTRTAREQVARLVEEATKEAEAVRAEAEELRERAVAEEAKARSEAEEAKARAAAEDSAAALAKAARTAEELVKQASKDAEETRSSASEEAERLSEAAEADRLRAEADHLAEELKGAAKDDTKEYRAKTVELQEEARRLRGEAEQLRAEAVAEGERIRSEARREAVQQIEESATTAEELTKAREDAAEAREAGEADGERTRAESAERAALRKQADDALERA RTEAAKLGEAEAAAARTREEAEQAARELREETEEGVRARREEAETELVRLREEAEQRVVAAEAEALTEARAEAGRLRKEAAEEAERTRTEAAERARTLSDQAVEEAEALTATAAEEAASRAEAEVAVRLRADAEEAE RLKAEAQEAADRLRAEAASAAERTEAEATEALERAQEEADRRRRSAE EALESARTEAGQERERAREQSEELLASARKVVEEAEAAEALVVEAADARATELVSAAEATAQQVRDSVAGLQEQAQEEIAGLRSAAEHAAERTRGEAQEEADRVRSDAHAERERASEDAARLRSEAAEELETARALAETAVAEATAESERLRADAGSYAQRLRSEASDALASAEADASKARAEARQDANRMRTAAEQADRLVSQAATEAESLGARSTEEAERLRAEAEAE RTVTEAAEEAERLRAEAARAVAEEAERAARAREEAERVESQALAAA EELTSQARAEADRTLDEARADANKRSEAAEQVDRLLSETAAEAEKL

		TTEAQQAALKATTEAESRADSMVGAARAEEAERLVAEATVEGNSLVE RARADADELLVGARRDATAIRERAELRERVTAIEIELHDRARRESSE AMRNAGERCDALVKAEEEQEAKARADAKELLADASSEAGKVRIA VRKAEGLLKEAEQKKAELVREAQIKREAEAAAERVVAEGQRELEV MRRRADINQEISRVQDVLEALEGFESQAGKAAPGGSGTGVKAGASA GSSRSGGKQNDN
SEQ ID NO: 25	meta_gene_343942	MENSGLSLDAEQKITVAEKVRKEPNKNYFISASAGTGKTYTLTNYYIG ILEQHEKTGESDIVDRIVAVTFTNKAANEMKDRIVKEIQKKLESLS DRAYKYWKDVYKNMSRAIISTIDSFCRRILIEQNIEAGVDPNFKIINEL KQKKLIDKATQRALQAFDVYDAIESGENYTEKVTNYLYGLTTERK RIRELSDELAKSKEDIFRLFEIFGDISDVAEKIESVVTNWRLENEKVS ERLLEVFEAAGGALRAFRNISLIAAEFYSETLDNFYDFKGVLKTLK VLENSVIREYYQKRFKYIIVDEFQDTNELQKKIFDLIHTNDNYIFYVGD RKQSIYRFRRGGDVSVFIKTMNEFEKIKSGRTDYEMLSLNINYRSHPEL IDYFNYISENTIFNNHVYEALSESPTSKTTNNKSKSKKDKNKSQAN GEDIVLNEALQNVNDIFSTERDENIYIHEVFRLRYPELYQKLWFIKKDD ESNAAFSPDSNEFLPGDLRRVNYITISKASLLENTQENDETAKEIGLDE DNQSPGKMKKLKDMDERELEALHVAKVIKSLVGKEMTFYEKDGKF VPISRRITFKDFSILSYKLEGIEDVYREVFAREGIPLYIVKGRGFYRREPI KAVISALYAIQNPNSNYYFTQFFFPTFDNL_EQNPEVGVRNGVKIFH KIVMRYRESKGQGLKKSLFQCAKERAEENELPENVTKMIKLIAKYDE LKYYLRPAETLKLFWKESGYLRKIPHYPNSSLRNRVKLLEQATEFD DQAPTFELTRLLERISEVQEVEASEISEEEDVVRMMTIASKGLEFNI VFLVNNDGVDKAEEEKTFPESEDGNGRYVYISQFLDKALKKFETSRV TKELEKELKKLEAEVIYDKTEILRKVYVAITRAKEMLFVVDLQRKNT KGIPAIKYLTPKGFEERIKISSLDEIDKLAGSGVESVSGKQFAESIQL LDLENVVDKGGLIFSDFTPCKAYKRYISPTLLYGIKDEKSDLESVDESED FDSAETISITSTSNEASKAKARLKVLNSLLEKATEITRGKQIHSMLASI TKYEQLKLLVEKNALPEDILNVRVLESLFNESEKIFSEWRLAKSIEIYD EKLKERKNYILFGVPDKVFLKDGFYVVDFKSTDLYKEAAEIERYMF QVKFYMMILLSDLGKVHCGYLVSVPRGQALRIDPPGEELDEIIYKIKQ FEELMSI
SEQ ID NO: 26	gene_1456 430	MLFGMTGCGTSSVTSSADA VTDTESVDDVKTESSGKTDEEKLEKIG ELTSAGKKGKDETUVVISSADGSKKSVIVSDHLKNGDGKDTLEDK SELKDITNVNGYETFKKGSDGKLTWDAGSDIYYYQGTTDKELPVDVK ITYLLDGKEVTPDEIAGKSGKVTIRFDYTNTEKTVKIGGKDEKIKVPF SVVSGVILPIENFDNVTVTNGRIISEGKNNIVVGLAFPGLKESIDLDDLK NEAVSEDAKKEIDDIDIPDYVEITADAKNFKIDTTMTVAQSNLLSSVN LTQDVDTKELTDKMDDELQDGADKLQDGAGKLKDGTESLDGTEKLK DGSGDLKDGTKKLAGGTDDLKDGADEKLKDGSADLKDGKTLADGT DDLSSGVSTLKDGSKKLAGGTDTLASGASQLKGSSKLAGGTDDLSS GVSKLKDGSKKLAGGTDTLASGASQLKDGTSQLSGGLKTLKAGTSQL KAGTDQLSAAKPQLDQLQDKLQDMGTQLKEAENGSAKISDGIGKLG DALTAFAKTALNMKAMDEGVQKLSAGISQAANGIKELKTFDNGV VGIHGQVNQLIADLKDYSKDEASGIKGIGYRGIGKAAYNTGINQAQR AAQSADENLQKAQEAVDEAQKAYDEALKAQQNSADAGNSLQQQND DLAKENAKLQQKIDELQNSADQEKKTNNVASPADNGSASSGNASAE KAGTQSTDSEGSKAAGTEPAETPAQNDAAADASSQSAAPADNTSSED TNAGNSAADTTENVQSTQASLAGLAVSKLNEMKNALYESTVLVAKA GESSETVAQAQQALEKAKESLQSAQQAKVAADATVSALKDMKSSVD SAEKWKGTNLKRVEKMTRIMGEAEAINSSLILQSVDAALDSLSS GLDSAKTGLDKIHNGIDQSLNSDETKAEQQQQLNESLTALKGGAGQLT TGLDSGLQLTDKSAATTKNIGDLKNGIDQLSRGANSLDDGAGKLA GAEQADNGAGSLAGGIQELGKGAHDLDNGIGTLKSGASDLKNGAHQ LDDGIGTLKSGASDLQSGAHQLDDGVSKLQSGASDLQSGAHQLDNG

		AGDLNDGIKLDNGAGDLQKGAGHLDGGTQLIDGINSLNDGAHDLD DGMATLQDGVIKLNNEEGIRKLTDLFGDNVQDVDRINAVVDAGDDYT SFAGTGDQENSAVKFIYKTDAIKAKED
SEQ ID NO: 27	gene_3178 27	VKKILFPKLDGPPSDDENYMFLGTFEDENGSLTTAKFFVRVSJVHSP GGCYEVEGDWKRTAKGEFFNSWCLIPSVPDTFALSCVYLNGLFPPMEM CGTSALSRRRLSALTREYGPDVLVRALATPTILTRLSDQPEIFAANILRL WEAATRESHMALMMHRAGFTTGDDMVWRGCAFKVAERIGGDPY QLVAIPGIDVAKADMLFRTLGGNPYDPRRIAGIIRRSLMASEGLSATN DDGEKIGFTAHVEFGSTAVDVTIDLTSGKAEPRLDDLSGIDPKIGMR LDVLRDFLSKPQEALKFGLRIRKTRDGRTLVARERVYQAEVRVARNI ARLLQAPPLKDKATVQATCRNLNFNQPDFQRFDAVQRTAVEMACYER FCVITGGPGTGUKEKSTILDAVIAARVAMGTEKRSFLGAPTATAALRME TTGLDAATIQSLLKCKGEKAGGEQWFDFNRNNPLPSGCTVYVDEGS MVDIFLSDHLLDAIPTDASLLILGDDGQLMSVPGAFLENLLNTRTMA GDRVVPAAICLQNTYRSNPKSNLAIQAKEIIRYGGVPTINGDSSGGTSMQ SVVPEKISNFIVYAMSNVMPALGIQNPLKDVAVLGPQNPGVGGLWEI NSQMSRYFNPNNGAKIPGLSAPRFAKEPMVPRVGRVMRRKNVKGDK LCVNGSRGFIEAYIPPSPADPAKKGKIKIRFDNNEVRTEDVSWDWHK KFELAYALTIHKSQQQYQYVLMVITPEHANMLDSLTVYTGWTRAK EGVAVVGSFDAFAGAVQRSRMNTRLTMLPDLLSEILVPGIADEFRSR WYKKPPMDDLPRPGGREKWFQTKYGNASGHKIRTIEGIKVEAPANG VQAGLRRGGFPSPPSQPHSSGSGPTTPTASGSHQAPPVRYAVNQPTSSPP RPMFTGGIGYRPNIPVSSALPNPPATPSYDKGVINHVQENAPPRQPN TSHQDATSPTHPKNSNALQPSQAVPLQAVGLQSPRRFGWSPTIRQPSA APATSNAQPTARSAAPDHVPATSRPAQPHRPVRPTTPVESPSARPVPA SRPSFGFIGWRPNIHPIKQTCHEPQPEMDSEMGMEDQHSSYYEDAPSP
SEQ ID NO: 28	gene_4421 494	MTNKVESNVSDQTEKRLSPEVSEQFQQDTRVVAKQAAEFIEEIHPARL LQTQEQIMDLSYAKSDELLDSFAFFRIVSCTTDEVDDMFDFLNEKMD KFYTALYAVGKPVVYGIVSYGETTNLVVGLLDTEDNSDLLKSIMEGL LDGIELLPYKTNFAARTACEKEVGLISAIPSVKIEEKQIFSLAPLMKSL NGQDYTVLFISRPLSQDIISKRRALIQKDQCFAVSKRNISRQQGISRS KGNTTEGRDTITKSTSNTISESGFWALGFTSESYSETTSESSSASENYS QTITDAINQSEGISAEVQNGVALEMODYTDKAIERLRQGRSNGMWET VISYSTDSDKLAAGIIRACISGEFAKPNPVILPQVVFHSFHLKTEAEGKSL LVPEILDAEPELSPLCTVVTSEELGFMCTLVDVVPNFELKKGKTYPLI TDNAVGVVEGHICEGRRILEMPFSLTHKDLARHTVCGITGSGKTTT VKGILKEADTPFLVIESAKKEYRNINLKDKKRQPIYTLGKPEINCLRNFN PFYIQCQGVSPQMHIDFLKDLFNASFSFYGPMPYILEKCLQNVYKKKG WNLTLGFHPYLVNTANSAKFFDADYMQKKYASAHHKYLFPMTQDL KLEIERYIKTEMDEYEGEVAGNIKTAIMARLESLSGSKGYMFNTYEYA DMNALLNHNTIFELEGADDSDKAFCVGLLIIFINEYRQISQEMLDNM RTLSHILVIEEAHRLKVNSTEKSSEDLGNPKGKAVERHFANMLAEMRS YQQGVIVAEQIPSCLKAPDVKNNSSNKIIQRLVSADDQAVMANTIGLTG EEGLDLGSLKTGTALCHKEGMSLPVRVQIAMVDDIKVTDDLLYGD KKRLYQINVSLAKEVLADSLPLMGGMKMLNTILVQDCNHVSHAVTVC RQSFRSSLKKNNVTLVMCDNENEIYAEELLYEGVLRYLLNGCYILKQM IPDELCSDIYQLMLSPDNDKLVLVKEQLQAEEENLEDQGCFIVAQLI YKNAFERTDIVQTICKNYFFEISDEDILKIKAEWRGSD

SEQ ID NO: 29	gene_3011 455	MSSWDPQTSGLTVRLDNPGRVGHTTGRWKAGSLTLVEAFGPNE KQFKNQELLEQVHSSEDPLDLLLGGKLGLPSDLRRVLAFEKVRGELT NIFYSMESSNTDFYAHQFKPVLRFVESPLGRLLIADEVGLGKTIEAAIY WKELQARYGARRLLIVCPAMLRDKWRRDLQAKFNIKAQVISASDLL VKAREIVTDGALESFVAISSLEGLRPPADFEDDRKASRRAQFARLLDQ NPTSADFAFLFDLVIFDEAHYLRLNPSTANNRLGRLREASRHLLLTTAT PIQIGSQNLYQLLRLIDPDVYFNEAVFADVLTANAAIVSAQRALWANP PKIREAEAAVRSARANSYFQGDPVLQRIEALLPEADTQTVMRIEALRL LESRSLLAQHMTRSRKREVLRDRVRRAQSVLAVEFSSLEKEVYDQVS AAIRAKAKGESWAVVFSPLICRQRQMASSIVGALESWKNTDFLEELVV DDLGVLQPQDLFGDRGDNNQQEVAAPTTINLTSVDLARLEELDTKYRQL IQFLKAELKRPHEKFVLFCAFRRGTLTYLHRLQADGVQAIVLMGGA DIDKDAVVTFSKTTGPTVLLSSEVGSEGIDLQFCRFVINYDLPWNPM RVEQRIGRLDRLGQRAERISIISLAVENTIEDRILMRYERIAVFRESIGD MEEILGDVTEKLVQLFDPSSLTEEREQRAAQTELALENSRQQQGELE QEAINLVGFSDFILDQINESRAQGRWLSGAELLALVDDFFARHFAGTR IEPLDHEVTSASILLSEEAKLSLGQFIADTAPAVRTHLHQSLRPISCVFD PRRVNRSVKGAEFIEPSHPLIQWVRQAYELEPAQIHRASALHRSGET DMPEGFYAYSIHRWSFQGIKRESVIAYAAQMLGQARPLTSIEAERLVG LAASRGQPLANVFASGVDRHELSQAAQACEEQLGLEFEKRLVDFLVE NTVRCDQQATSATKFAARRIAELQDRVERFQLEGNDRLVPMTEGLLK KEESELKFKLQVVDKKRNVDPTMVHLGLGLIRVA
SEQ ID NO: 30	gene_2590 511	MSNFNFLLTDISPELAQFGKSAELYCHDDKQVALVLRCTEVVVGEIY SRLSLTPPVRRDDLYNRLRSYEFKDVSDKGIWAKLDVLRHKGNKAA HSSNGSDEISLNETLWLIKEAYLVARWYAQAILNKPITPPEFVDPVKPI DHTSRLEAEELERQRQELNKREAELKTQLADNSDKYQQQTSELIAQLD EKNDTLSNVKKEQALLQIELEQKQKDLVASQQAFFDYRTREEFKQASI SSASSFDLDMEVTRRNIDIFDCFEGVSLTKGQNQIVKQINEFLDTKQN VFLLNGYAGTGKTFITKGITQYLERIGREFAIMAPTGKAALKVISDKTM QPASTIHRVIYNYDNVKEYKVDGVEGSETYRCYADLKVNVDTAEAV YIIDEASMVSDRYSDGEFFRGSGYLLKDLLKYINIDHNDHNKKVFIG DNAQLPPVGMNTSPALDASYLKENVQAVASGYLTEVVRQKGDSGV LNNAAMLRDGLEQNLFNKLKFEVNDHDVFNLSSENLLSTYLDSCDRK VSRTEGESIIASSNRQVAEYNRLVREYFFTQQQMVGDKVISVANHY RADACITNGEFGMIKEVLSPHSELISVDISVKGDTGDMVKRKVNLCSR DVLGFRNDYGEPFEEAKIVENLLYNDQPTLSSDEHKALYHFLNRH PELRRKGNEQKLRIALLQDPYFNAFKLKFGYSITGHKAQGSEWKT LQCQTHQKALTQDYFRWLYTAITRTSGILYVMNPPQLRLGDGMKIAG AYQPKAVNLDNSAPEGVENVVPSTEATNSVATAKFDFTDIPQLK YQLVDACIEGTGITVVDVLHYNYQDRYILQRGNEQASISFNYKGNWK VSGVKSITQDGFDVELMALLGQLEGTLVDPEPSKDTQFHSEPFLEE FYLNVMDQINSVGADISKIESRSFCERYAFVKGNELA VIEFWYNKSSQ FTKVQPMPLQLSNSTRLIDEIICQIGVLL

SEQ ID NO: 31	meta_gene_463174	MVNNKKVMSDNTQPKASVAEAFGNAKKAKTINGIJKKIIQNAESGFT VLNVFSNDKFITASGTFFDKPLMDSKIKLKGEFTYHKKYGYQFNFTQY EVLSNTKTAIIYEYLSSSIFKGIGKAIAREIYDKFKEKTLVDIDPEKLK DVNGIGAIKLAVIDEGLKESYGLRKTVMFFKPYQFSDYQIKAIYNRFK DKSVTIAKENPYLFTDIKGIGFKKADIMSEKLGKDDPNRIKEAIKV VNQICESSGNCYIYYQDVKKGIGEIIDLEETDLKKYLNDLIKERKLL DFKGIYGTDNYLSVVRDRYVSSKSIDLKGEVLDFTSAKEGKRLGCA RIYMPVYYHCELGAAKELKRIRESASPASDKIESLDDDKFLELGNNH VSLETNEQKTAVLNALKYKISIISGGPGTGKSTIIKTIVHLYSGEKIALTS LAGKAAQRALADIVNSGQTLSSRNDHSQEKMGRNLNISTIHRLKAQYD RQTGESYFTYNERNRNRLPHDLIVIDEMSMIDIIFYKLLKAIKDDANIVFV GDVNQIPAVSPGDVLRLDIYAGAGNMDGQDKTPFPSTFLTKVFRQ NEGLLINLNAHNILNNKKFVTLRKDCKEKNISTAEKDDSTIKYRKEY DIAVGKHELLIDFTRFIKRVENRINKRDVGLKSANMSIPTMLFDDIQ VLTPMRRGDLGYFNLNNILQDIFNPISPLHLSASVENIFICNGIQFRLYD KVIQKRNNYDQDVFGDTGYIVDVNVHNEKYLTVDFSNYSDLSKKCN EIGTGAESCANLTAQEGKMTNKAIKLVKYNFLDVYENISTAYALSIHK AQGSEFNNVIVLFHQTHYMMKKNLLYTAITRGKKNIVIFGTFKAIGI AMGSKETVRNSGLKDRLSEEFLDAN
SEQ ID NO: 32	gene_7738 46	MIENLPPFSIILAPAYLHPILRADIMKQTSGCMGLQLLSPQTFFASFTQK QARDHVEISFLYKQNIEKIISQLQTYQAIALTPSLMECYDFIESMKFY HISVDELPDKTQAQQEIKTILNNIFPIQTAQDIWNEAVRVSDCSNVYI YDAFYSLKDEKILNLTSGAHTIPLPKPQQQKEFYHAINPRQEVEAIA QYIIQHDLDADDIITLASSTYKPLIEQIFKRYEIPYTLQKNKASIVTQR FVNLIAYALSFQEDLFACMDAGVFQSEHLDELREYIEIFNCDIFQPFH HLMNVQANGHILDEVEITKLKELEEIAESGRQELCETLSLFIEDDLHQL VTHLLDILHNGMKEASMEDISVLSNIQDVVSSSWNYLNTKDDLAFLL PFIEQISISKSVREIHGVIVGDLKQIIPNRTHHFLVGATQKNYPAFPSESG IFDEIYLRTTLPDMETRYQYYIAQCEKQLHTNSHLIVSFLPLGTYEGKG NEEAALEIEEMKCDPTAFPIMENYEKITQTYIIQPETAKALFVKGHHIK GSISAIERYIHCPSYFLRYGLSLREPMQHGFDNSYMGTMAYALETL VDELGKQYTAKAMERIEEVNQEVEAIAAVFPNNADLMEVIKHRFLV SFAQTLKRLDDFETHSSMGPYLQEYEFHEEFPITEDISFALKGFIDRIDA SGNFHICLDYKSSAKSLSEDKVFAALQLQLLTYSIVAKQLHKDILGA YYISLKNQNIPYIAGMKRRPVGFVETEKDDYEENILKAHRISGWTMR KDIDMLDDNGSHIIGVSMNKDGIVKARKYYRYETIYEWFISLYRTIGN RMLSGDIACSPDADACTYCAYYEICRFKGFASERKPLVDIDDSLYWE GGVDDADME

SEQ ID NO: 33	gene_1188 229	MKGSIKSHKSAIAVLLALALSGQSSWAAQNSAAVQGNDFLSSIQQIEVKQIDFPAPTHRQQTPSASRAQINDLQQEIRLKKQLKAAEQEKKSLA PGDLQAQNQTQLLKDNSALAKENDRLSRSLQNAQREQGAASTQQAAR IEALEQKTAELQASLASKTEELAQLKKSSNSQAASESALQKQIARLET EKAIAERNTKDTARFNRMQALRNELNKRADELVALKNAGDKRA QSQTALQKQLAQLEKEKAALTATQSAQSIDVANKVQALQAELDKRS AELAALQKTGSEHEKSQSDLQKQLTQLEQEKAALTATQNAQSIDAANK KAQALQAELDKRTAELTALQKAGSEHEKSQSALLEKQLAQLEREKAA LTAQNEKSIGALNKQLAQLEEEKASVTEQNSLLMKNSSLSEEKA KL QKAQAEQTALLEKNQAAEAALKAQIAALTEKLNASTTLAATSQEKV AALASELASLKGQSKEKAQALQSQQQQAAQIAAAKEALTQQLATAQ ADIALTKQSLAEKENRLQQSDKALLALKEEAQSAKALTASATSQQK TQAELDTLKRANEELNAKLASLAENTAQKAQAEKEKAELLAQAEK EKAELLAQAEKLKADAATQVQTVAAATKAEPEVSAALKDKANKQS YANGVMFSRLVQKSMDQMADLGKTNLPILLAGIKDGLAQKVAVEP KTLSSLHESMLKELSSREEKKYQAGIDQLEKATAKKLLRNKSLFF VQAKAGKKAIAPIGETVNVTFKEATYEGRVINNNANVPVTYDENLPYI FQQALELGKRGGVMEVYCFAGDLYNPDTMPPDLFNYSLMKLTVIS GGK
SEQ ID NO: 34	gene_8002 33	MDYDVYSISIGTTANLGDDKANKAVQDLGRSIDKLPPQLPGGVGGT GAGGSATPSYVGTPSTSGSMTWRLDGMTELGTALGQTEAAVKQVDK SITITSQRLDKNSSWLSRSISTLASLPGKIQSWSGNTMQAWGQFNGPL QNVKNMISVGKQAWDLGWSLGESLNEAFGVTKQIDAKVAGIIQAA QDKLARWQDSINSARAQHREDAFLKQEAAGVKQVNDAYAARLRTIE AIDRKAMAGLELQQKLLQIENEKNRSIIRQRQIRGEISDAQARDELAKI DAKDAGERMDIERKQAEQAAATSQAKAEAEEERYRKLMEQS QGM ARQAVQDLKPMIDILNKADSLKRAEEDLAKWRSIQQRQKEAQKEIQQ AIKDQARASTMLPLVGAPIALARKQAEDQARQDYEAAVAQHEFMH DKGMSFNETDKGNEDALKKIVEQRRKALDSMLGKIDKTGLVGNMDG MAEDQRLGEYLRLKLVQDAMAQDAAQLESIFLETEALKQQAAEDK ERVQRVMQEHQSQQAANDAVTKETAATNARQDADKHADVMVGAQ EERLRKEIETKQRQQEKQKEDLSKTNERLNANMERFQQYAESFEGND ALSAKLKQFSDIFTRLKGRPRDTWNKKDLVDAKAAEKFAKELVEASK NSTNQDKKGIAQAAQAIKAWQESIKKERAIKKNDKALRELERTAQ DVANLSGKLHDGQSKVLELDDWLAKMRRKVLGSSGEIANKAPI GAL PQAEEVLKKVLSEQGDGGTAVTQGERKLLEHLKNKLKNDDRLEAG NEFDEMIGLIDQILTRYSSAQSTHSKLSGEVARLKARLDKIDSQGKFGP HR

SEQ ID NO: 35	gene_1538 800	MTDPTSSVQTQGVRKIYAYTTPAEEVDLNGKGNRVKIGHTTRS VAERIREQFGASSTDRTWYPRGEWDAQAEDGTWITDHMVHRYLSKR YRRVPDTEWFEVDPPEAVWEAVEVLKNDPKARPKGKDCYELRGEQR AAIDAAMTYYEADPSNRWFLWNAKMRFGKTFTAKLAERLRSKRIL VLTYFPAVDDGWSEEIEDHVDFEEWQYAENGASYEDGDQVQVSFSSF QMLEHKTLGKNRENANTKADALRREIAAVNWLVIIDEYHHGAHP ARREFVSSLKTQRILALSGTPFRAIAKGDFATENKFDWTYLDERQALA DWAKKETCEANPYEELPAIHFGYRLPPHAALTGVGDCDLTYSPTTI FKADKDGFKNPEAVKDWLQLSLSLGSARRAGGVPPPYPDFTGDLV SHVLWLLPSKHSCDAMKRLLEDGWFPGGEGEVIQVSGSEGETGKPAQ ITKKVRDKIAAKRSITLSVGKLTGVTVPETAVFHLKAGSSLESYL QASYRCQSSGSINLRNGDREVKTNCVFVDYDPDRMLVVMGDYIKSLK GSGAPLDRSNAAPTVIFDDEKNGHIPLNVSDIENNNYLLKRRPA ELMSDSMRLLEDAISAGGLNKALCAQLVKAGKS KSPHHAMRDLDPS LFKSKTPGTIIGDNGKQSAKSTEVADENPKNDEIRAIKEAMLVFIKSLG HLAYIGDLREASVEDLFKVDDDELFEKIMGNDKDQVREIIDGAGLDRV QLNAMIQKILMWEWFVSGCRNRDELGECKWYPSDEEATYAFSL QPNR
SEQ ID NO: 36	gene_5543 656	MQRTLGNAAATARAVGRGKRPAGRSPPAIDERAEQGLVLPPYLMDLE AGGLSTAYGLTGHEFVRGAAAVVGHGGGTVAGIAELAGRPESSFF GRGRAFAVEAGPGGGSAGAQGGGGYDVTVSIAAPAPDDRPTFHPAA GLGTAAPDPGGAPLAAVDDPEGKETKVDVQHNTGATASRSVGNAS KGVGGTAGFLAPVAPGLWLGAATGNVQPWQSSRDSRSQRGVAEPR VLRSDKSVEVARVVYVVVRVPQAGGDEQVFRGSGGLTQRVPTEH LIPAGTGAPARPEPVDAGLARRVALADSLAPLGVFDEAGPHRGGGGL FDAVASVLHPSLTAPGAPGRARLYEATATPTVLEDPLRLLGGDGVTG DDLYAKGSSAGSYRMRAVTGLAPA WSTGKTQLRTHQQAQHTAT ESAGKGRAVAGGIGPAAGVGAAANAAVVRATAMPVAAARKARFSV NEQTVSSRQGAEVREGEKVLYTGTVRTVEGTGPRSMRMIRHPEARVA THAMRVWISLRADEAQELGLPLPPGTAGHFIRPPRGAAPTSAGGE GEASTPAAAGSERHLPFGAMGSSVTLGRLDTAPMMKAVRELFTDP RLTGYLPAGFTTPVAGLSQEEAAQRANHRELT ALSEANLRVNKD QLLSTGIRVRLRRKTAMHSHDVQLRVHGTMGAEAGHLDIDDWLVRA HAGVASNAQSGRSSRSIGGMVLAQARLIPGALTGSARYERTTSGTRR NQAGPTTRTDVLTNGSEKAAAFGAALRLNVDVTMTSRPRKMTRALT PGAPGRDVPEAKLLSGLHLEEQDVRLLTPTEFVGAEKRLDAGAG RAPGAESATTATGIGDLAGAAPTAP TGQHLLSDWQLVETVGDRGPIR ELALSLLSRAAARGEAGRDPALTTEGLAPRLAVEERFSPRAITASLR QAASSGWVVKNLRYPRRLAALNGAVGTRLALSSQLVHEAGPGTE TFVLGGHQAGGQQGGGTSTTVQAGATLVQNGADWRVGEGLSAYGS TGTGDSEAATVAGTVERNAHTPKKAPLYLVRCDLLVTMVAEVKVTG GGPYVASAARTLPGAAA VWL TAAQLRAAGV DLP RSARKE LKAD GTP APTTTSAAGVSGAGSGV RSAARS DHGP GTRSGAGSGPGEASGGSR PRPTLSRGLPLGFGMIEDVPDFVPLLSGLRTTLALTGHQDLADELLPR QQLRDRNDNVQRLRLVLD RDGSTGLASAMDGGVTVELLDGRRTPY WAVFKVDRVGDGVWDGEADDGRDMEYITSAVAQQSTA HDEGE SVG VEGVLAASGRPDGGKGQVKSTGAAAGLGLAKGS GR RRG GAT RG QL GMKTVAEAKTAKAARMRVPVPSLEHRGD RRLA VAGL GRT TLVH RVLEADLKALSRVTPRPAAHPRPDAPQGSDAALGA WRASGVPLP MEAQVNFGQGAPRVRDLV SRTVRAAGGNPRFREKGQAAAYTLGEA VSTEWLIAALPLLTHAGAPLPPVHATGAKGQDLHASVHARL RAGRIL GAGDKMTFETVAQSDLTAPRPTQ TDQAQS AAEKSRQARGLLGAGVLN ADEFRLNQLMANGGGAGSATDASAGGAGSMPLHKPKFASVLVQFTL

	DVRVVARVTDRVSSRTAVAERELTPQPVVVRMPLVARRMLAAY PEAVADSRGELGV
--	--

SEQ ID NO: 37	gene_3943 627	MQQTLGNEATARAVRRGKRPANRPPAIDERAEQGLVLPPYLMELEA GGLSTAYGLTQEFGSAVAAVVGHGGGTAAISAEALAGRPESSFR GRAFAVEGAEGGQGGRNGQGGNGFDVTVSIEPAPDDLPTFHPAATL ASAPPDPGGAPLAADVDAEKGDTKVQHNSGTTASSTVGNSSTG AGGTAFGLAPVAPGLWLGAATGSVQPWQSSRDSRSQRGVAEPRVL RSDSGSVEARRVVYVVRVRRQEGGDEQVFRTGGLTQRVPTEHLIP AGTEPLPSSGAGGQERPVADLARRVALADSLAPLGVSAGPHQGG GGLFDAVASVLHPSTASGAPGRSRLYEATATPTVLEDLPRLLGGDG VTGDDLYSKDGSSAGSYRMRAVTGLTPAWGTGKTQLRTHQQAQH TATESAGKGRSVAGGIGPAIGVGAAANAAVVRATAMPVAAARKARF SVNEQTVSSRQGAEVRGKEKLYRGTVQFTVEGTGPRSVRAILPEAR VATHALRVWISLRADEARELGLPLPQGVAGEFIKQPEAGAEERHLPF GATGSSVTGLRDLTAPMMKAVRELFATDPRLTGYLPAFGATPPPADL SREEEEAQRANDRELMAALSEANLRVNKDQLLSTGIRVRLRRKTAM HAHDVQLRVHGTMGEAHHLGEIDDWLVRRAHAGVAANAQTGRSSSR SIGGMVLAQARLIPGVLTGSARYERQSSGTRRNQAGPTTRTDVLTNGS EKASAFGAALRLNVDVTMTSRQRKLARAVTPGGPGRDVPEAKLLSG LHMEEQDVRLLTPEFTVGPDEKARLDAGAGQAPGAERPVTGAAGIG DLAGLAPPTAGQLVRDWQLVETIGDGQPVRDLALALLSRAAARGE AGRRDEALGTEGLAPRLAVEERFSRAITASLRQAASSGWVVRNLRY PRRMAALNGAVGTRLALSSPQLVHEAAGPGTETFILGGHQAGGQQGE GTSTTVQAGATLVQNGPEWRVGEGLSASWSTSTGDTEAATVSGSVE RNAHTPKKAPLYLVRC DLLVTMVAEVKVTGGPYAAGSARTLPGAA AVWLTAEQLRAAGVDLPESARKALKERPRPENGPTTSRAEGSGGGT QTPAREGVGATGGPSRPGPGLSRDPLGFGMIEDLPDFVPLLDGLRG NLATTGRQDLADDLPRQQLRDRNDNVQRLLVLDRDGSAGLLASA MDGGVTVELLDGRRTPYWAVFKVVRSGDGVREGEADDGRDMEYIT SAAAQQATSHDEGESTGVEGVLAGSGKPDGGVGQLKSVGGAAGLGL GSGSGRRRGAARGQLGMKTVAEAKTAKSAKVRVPIVASLELHQGE SRLAMAGSGRTSLVHRILESDLTALRRVTPRRA PRPAPGAPTGGQAG LGTWRAAGVPLPMEAQANGFQGAPRVRRELVNATVRAAGGDDR FRE KGQAAAYTLGEAVSTEWLIAALPLLTNAGAELPPVHASGAKGQDLN ASVHARL RAGRVLGTGDKMTFETAAQSHLGAPRPTQTDGQSAAEQS RQARGLLGAGVLADEFRLNQLMGNTGGSGSATGAATNAAGSMPL HKPKFGSVLIQFTLDL RVVACVTDRV RTSNTQVAERDL TLPTPVIRM PLPVAGRLLAAHPTEIADPHDRLGLRTGAVPPGP
------------------	------------------	---

SEQ ID NO: 38	gene_5085 315	MKPLKSYLAWVAVTLAVAGATTACQDDIDDPIDIAPVAKDQPNTSIL ELKTKYWNDATNYIDTIGTRDDGSHYVISGRVVSSDEAGNVFKSLVIQ DGTAALSLSINSYNLYLKYRRGQEIVLDVTGMYIGKYNGLIQLGQPE WYENGGWEASFMSPEYFTAHAQLNGFPDTSKLDLTVVNSFSELPD PAGLIKWQSQLVRFNNVSFANGKATFSEHKSVNQSLVDAEGSSIN VRTSGYSNFWNKTLP EGHDVVAILSYGTSGWQLILNDYEGCMNF GNPTVPEGSQSKPWSVDKAIEIEKAGTEKSGWVSGYIVGAVGPEVTE VKSNDIEWKADPLSNTLVIGQTADTKDIAHALVIELPDGSKLQTLG NLVDNPNGYKGQIALHGTAKAMGTGITGNNGTTNEFSIEGLNPGG EGIPEGTVKESPYNCAQVIAGVSGNAWVKGYIVGSSAGKAAEMTN ATGAAASTSNIFIAAKADETDYSKCPVQLPIGEIRTLANINANPGNGL KVVAVKGSLEKYFGQPGVKTVEFDLEGVTPTPPTSGDSENP YNPAEVIAFNPQSSQEAVKSGVWVTGYIVGWADVSAAFYAINAETAH FDASATMATNILVASSADVKDVS K CIGVQLPTGEIRSALNLQANPGNL GKSLQIKGDIMKYCGVPGIKNATAKLEGGSTPTPTDPVASINENF DASSSIPAGWTQKQVAGDKAWYVPSFNGNNYAMTGFKGNGPFDQ WLISPAIDMSKVSKKVLTFDTQVNGYGSTQSALKVFVLTAAADPTTAK TTQLNPLATAPATGYS DWANS GEL DLSAFSGIIYIGFEYTSPVADNY ATWCVDNVKLNAEGGSTDPTPTPSGDFKGFNSFNNQPLSKPY GTYTNNTGWTATN AII LGGGETDANPIFTFIGAAGTLAPTLNGKTSAP GSLVSPALTGSIKTLTFKYGAFNESKCQFTVNVDATGNVIKSEVVT LDKIEKAKAYDFS LDV NY NGNFTIEIINNCYSQ LDANKDRVSIWNL TW TE
SEQ ID NO: 39	gene_4028 206	MVGVNERARVPFALLGVVLLVGSASIAAGLGGTSPTREPATEAAIEQ GR TSLGGTVHDATRTAARNVAASPVVAPANTLGRVLAATGDPFRA ALELRTYLA VRDRLSAT ERGTVDPSPALRDSADIDAALSRTTVEP VGANATAVRTTVANVTLTAMRDGRVIDRYAVSPTMTVQTPVFA LHE RTRTYQQRLDSGATEPGLARRATARLYGVAWARGLTQYGGGPIANV VSNQHVAVATNHALLAQQRATFGATDDTGRRAVRVAAARAAGTDL LAATGQSGKQI QELLAGVDAATPGSTLDPVAAANPPITPESALNVSVG EQATTAFDRFVTTDLD AVLAAPYRVTVERRAV TDSATTAGRERPT GDNWTLVGTEQTDETTVTDG DATVGSPVNPWHTLATTGRRVAETTR TERRWRRNHTTHTT VETT TQTRRV SIRL VGRHDGGAAPPVGT SPIHER GGAIDGPNLAAVERRAKTRL LGD EQDLD ALAARTTSDGTTQTTIRGE QPLELRDWVYRDLVRLRER VANV SVAVERGAVGTYQVNPSDELAGA LRARRARL VDRPDEYDG VADR ARVAARGAYL DA VITE LERR ADD RD GVKERLAGLLAARGLSL GRLRSIMAARSQVTTPTSHSISGVGGSYSLD VEGPAYLTLASVNRTQ TDLSLREGSVRPLA ARNT NIFTV PYG DADGI VGKLFGGDRVRLRSAARALAAGEELA THET LEAD VET AVS RRR RGM RRV LRRAGVGDSRDRR RIVAAGL GA WET VAERAIAV T ENRG PDAV AAVALRRSPGSFDGPADRDDL RSLRAV ATDGRGVPESSVTPHVERA RQMVGKLVKQSVGRAANQTTAVRERLESKTGKLA AVPSGIPVTPVP SQWYATANIWDIEARGGYDRFAV SVRNGGP GRR LT YVRDG STV VID WN GDGE LERAGTAT A VT FAY RTA VVVVPPGQGVGDV DGNADER SAGWGER

SEQ ID NO: 40	gene_2773 99	MPTTFENIKLKEDGTEGQIITISTFYVWDCTNQRFSTSPPVVLRNTMLA ALYPAKEFIIIGEIPKTSTNPSLLDPFKVPAVSDPYFLDLASNSRTHGRFL FTPKRTIGRDYFPKKDDWKRIYGSILHTGCNRMFYREIKYIVVDDERR NPSDSSPQDDGVNNTHWDTGDCHAKLSKSLLTLESWETIGNEDNPT TIQIRAAIFKEWTIKGTASHSYKFETDPRFAGVDLVIPLSCFKGNKPAP GNYTGKVLIGVVHEAEERRAKPGWMLWQWFSEFTELLEDGIISKLHEK CQKLSTALDDIYKLADVLRIDLDEAEQELANLDDNPDAEVAYVDSVL KIIKADKKGVILHPYVLLKVKFRLREMWNLAKSAGVRFYSVMCTP DTSLKEYQKAYGNDVFVKPKVFCSPSFNEGQYIVFCNPMRHWGDVQ LWENFHEGRFRNTRGVLAATRELLSLGRDTDGDFIQLINSSRYPNLT MALLYMDAPPKVKKFPKVALT GSLQQIAINSMNDITGVVASLLGRAR AIGAELIVLDIPKEGEMRIIDFLSQELQIAVDSLKSAYPNNQDGLKVVK EFLDKSGADIQWLADLKSDDCYFTRPCLVNLLTDVTTRIVSLVNSY YRQPNLKEDTIPMDYRFTLFSLVVSDAQDAIALRERDAYRAEMGAA LAHKAANDDDRVLKEVTAKFRASTEVIMRETLPFRKPYPPKTWAAS YWRVNHLAKSGTAGLVFLFCDEIEELKNLENKKVWLITIYAVQFTA FARPQLNAWNGEELTVRSSFLNVNGKDKVSLEGKLDGQPGFINMGL VNEKDIAQVPNGWTGRVKIYAKTYENDKYPRKMSANDVCTSLYCFS VDMEQSDIDDFMNDHWSTNSRFNPI
SEQ ID NO: 41	gene_1961 732	MNRSLSAVVLTAVLFNCVKSAPDLPTQPFAYHEDFETADPVQFWV SNGEYEVNSKGGLTEEKAFAGKKSFKLDVTLKTATYCYWSVPVKVAC AGKLKFSGRISVSQASKARVGLGCNYVFPPTHHSGCGAFDTFDKATD DWQLQEQLNLVADGDERADGVLQRQNTSDATGANVVTFTDRWGIFLY GGEGRVVYVDEVRLGEVPDAQVYAAEADQRFEPAREVFRKRLT AWREELATARQGIDALGALPPVAQRMKEVALKAADSAAEADLTKFAE ASYASPTDITRLESSVRTVRYATPNLIDMSKPGVADRPFTVYIVKPITN ARLLPTSFFIVGRIASELSVTGCAGEYEPEASFAVSALKDVEKLVVPTD LNSGANLIPANAADVSVIKCWYQAGVSISDTRHCLLTPELLKDDALV RVDTEKKENYLRSGEKEYALISTKDSSTLDIOPRDAKSLQPVDLAA DTTRQFWVTVHIPDDATPGEYTGTLKAAAANAPAAELTLRLRVLPFK LEPPALCYSVYYRGVLTDPGKGSISSEEKSPEQYAAEMRDLKAHGVD HPTLYQSFNEPLLEQALDLRKQAGLPTDTLYTLLGLGTGSPTNAADLD KLRATATKWVEVAQRHGFGEVYGYGIDEATGDRLTAQRAAWQVLH DAGAKVFVACYKGTFEVMGDLDDAIYAGAPLADEAQKYHQAGQRI FCYANPQVGVEEPETYRRNFLLLWQAGYDGAMDYAYQHSFGHGW NDFDSPQYRDHNFTYQTVGVIDTIQWEGFREGVDDVRYVTTLVKA MEAAREAKPALVKQAQTWLDGLDVKGDLDEVRGKTVEWILKLT

SEQ ID NO: 42	gene_2755 817	MLLVHIAGHADLGAPSPFEDPDKGPLRAEELKNCMTPHEATRCLFDL SFTQTPSHKYTDATHSPHSGSALRKELTAVSQISAATSTDDETTEVLIIG VEGEDPTDRLARALVDALRMASSEAADLAGTSEIIIRDACILPSLAVS RESIELLERRIGAHDGHVLLAMAGGATTVLAEAAGVAAATHQDEWS LMLVDRVEEGSDGQLPLIPMSVDADPLRGWLMGLGLPTVLDDIYEQ SDRIDTEVKKAADAVERRVMGELDSEPSAEDFAQLQADVARGDLAA GMTLRAWILAKYKHLRDAHSYTNDSCQSNKQLRQELGRVIGRLRE SAKSHALEEPESWLVHQGDNLNDLGKYATHNLESPLRNLTNNLQERI KQAVGEPPPEWLSMPSGDVCLTAQGKAARNAPLTSGADAPDRKRRR PIIVSLLTSEPSDSVRQACAVHGPLTLSPIACSSSLSEGRRVADEVKN GEQPASHSPWTLDETSIVHDYGESITRPGVSSETISSSMKGLSRAAEH WLEERTSRPRAVVVTVLGEKAAAISLLHAAQIFGAKHGVFVFLSMV NSKDTETGESKESVQFHQLGDRDVRQALLKATTYCLNRFDLLSASR LLSLGDPAMEVLSNEANILADRLIESVNTNDLGASSTVLSAMNAVA DLVKIVPSDAQVRLTTIVGELLRTPDEKYRSPNFKAVALACASPDFD QGNDYKKKLKQLEPSESLLRLLIRVRNKIPINHGRNTLDVATELSL QNFPDGNRYTYPVLLQRAIAAVGSKHGARAGDWGHRFHSLRDQVEA LGKTGYGEKP
SEQ ID NO: 43	gene_2831 443	MTYHIRAGQLVLEINERGEARLQADKVGASEGLPMAMYPSPLLRLVQ DGEIQEPAGCEQEDRTGTLTYPNGTKIKGVAVRDSYAALEVLTIE SGSPDAVIWGPFRTRIGGSIGESGVVHDGRFAIALQVNAKTVGGWP LELDRLAYMAPSYSEGDAPDPNGRRGSDNKFEYPVCTAWPTVDGGS ALQAYARDRTKRSIRKAWNVPATEVRPFEGEDAVIVGSGIALFGCPVE EVLETIEQIELGEGLPHPTIEGWGKTSPAANQSYLITAFTEETIGEAVQ YAKLAGLSYYHPDPFEQWGHFKLKRGSFPSGDEGLRRCSEAARAE GVSLGIHTLSNFTTLNDSYVTPVDIRLQPLGAAVLAEEADERGDSLTI DEPWPFITVALYRKTARIGSELVEYAAVSETKPWRLLGVKRGMHGTA ASKHGKGETVARLWDHPYDVFPDLELQDEYADRLAELMNGADIRQ VSFDGLEGLYATGQDDYGVIRFVERQYRSWGREINDASIVPNYLW HMATRFNWGEFWGAETREGQLEWRLSNQRYFERNFIPRMLGWFLVR SASDRFESTALDEIEWVLSKAAGFGAGFALVADEEVLRNGNIEALL AAVREWETARRLGAFSAEQRERLVEPKGDWHLEPVGPQRWNLYPVQ ATKPLVCTPAEQQPGQPGGSWAMFNKYAEQLRFTMRVRPSYGNE DAAVQRPTFYTDGVYMTFDETEIAANQYLECDGTRTGRVYDANRNL RVVEASAEAPTVRHGGTLSFSAKFIGDPKPDVAVKVWLGYDPETVS ADE

SEQ ID NO: 44	meta_gene _118560	MPLSRLQNFLKSVRGNILYVNPNLDATDSIENQGNSLTRPFKTIQRA LVEASRFSYQTGLSNDRFAQTTVLLYPGEHVVVDNRPGFIANDAGGGS AEYTSRGTTGLSISPFDLTSNFDLESSNVLYKLNSIHGGVIVPRGT VGYDLRKTKLRPKYVPDPENSNIENSAIFRVTGGCYFWQFSIFDASPS GQGYKDYTNTFLPNFSHHKLTCEFADGVNNIAVKDSFLNVSKSFS DLDNYYYKISDVYDNASGRAIAPDYPNGVDIEPIIDETRIVGPKGGSV GITSIRSGNGVTGNTTITVETSTALSGITVDMPLRIIGVTASGYDGQRT KSVGSGSTTFTYEVDVTVPSTLFETPSNAKELQVDTVSSASP VFNC LLSVYGMGLHADGNKATGFKSMVAQA FTGISLQKDVKAFVKYNTS SGVYDDSTTVDNAADSLARYKPAYSNYHIRCSDAVLQIVSCFGVG FNGHFLAESGGDQSITNSNSNFGGAALVSDGYKEDAFSRDDVGYITHI IPPKEITTSDSADEFVSLDVSKTLSVGNTSRLYLYDQTNADVKPETVIQ GFRLGAKTDDKLKVLIPSLGGTTTEYSARIIMHNTAYASDEPSSVKRFTL NRSSVGINSITNSILTLTKVHNFLSGESVRVIS ESGHLPDGIDEKLTYNV IDANIDSSLATNQIKLAQNETDALADNFATLNNKGGILTIESRVSDKLA GDAGHPVQYDSGQNQWYVN VATAATTENNI STVIGYSTAIGSNT PRT YISRKSDDRSQQD TLFRARYVV PAGVSSARPP IDGYVMQESC GDIETT ANIQLVTL TNSVQQRNQTFIADANYLAATGIATITTEKPHNLEVGAQV QMLNVVSANN TTGIGTSGYNFKATVSGINS DRSFSVAL DDDPGA FQ DT STR TV DLP YYKKDY ATNFY VYRSTE I KKHV KDQQD GVY H L T L L N AS N A P N I T P F S G Q N F S Q N I I D L Y P Q T D R D N I N S D P D S A R F A T P D D I G E V L K K S I T K E N I I R F R D S K V G I G V T D C I S D I V V G T S H T I Y T D R D H G L F G I K S V G L G S T G F G Y G S G A A G T L Y N A T L T A V G S S T V G K S A T A E I T V D G I G G I T V C A R A M V F N A G P G I G I T Y F S Y D Y L S G I A T V G S G V T A H G L S V G N V L S F V G S S N T A Y N G D F R V T Q V V G L T F K V N A G V G T E S P S E A G G S F Y A L P R G Y A S N D G A I S L E N E N L S S R M T P I L S G I S T T L N S A V T T K T A T S V E I T N S F N S G L Q K G N Y I Q I D E E I M R V A T T P V G G S D A V T V L R Q L G T R R A T H I D G S V I R V V S P I A T E F R R N S I L R A S G H T F E Y V G F P G N Y S T S L P E K V D R V L T G K Q E L L A Q S V K K G G V N V Y T G M N D K G N Y V G N K K V N S T T G Q E E V V D A P I A T V T G E D L I A S G V A V G L D V I T P L E V T V R S L K V E G G T D A N I I S E F D G P V L F N K K V T S L G A G G I E A N T F F I Q G N A T V A R E V S G I S T P V N G N P G D I K F F S D P K S G G S V G W V F T V E N A W R R F G R I S L Y D F K D T N I F D Q V G I A T T P N N Y E L Q I G A G S I I N A S A G K L G V G V T T P V R K L D V Y G D V G A T G F V T A G T Y V Y G D G S R L T N L P S D S Q W T R T D A G I N T I S T N A G I T T N P A Y S L D I R G G K S G N G Q L Y V G G D S Q F T G V A T M A N V Q A T L S A T D V L I I D S D G Q A D V G I V T V R D Y F N V G G G T V I F T N S A G K V G I N S A T D N Q A A V D I G G R V R L D D Y Y E K V T T V T S S G V V T L D L A K S R T F N L T S E A V T Q F V L S N R L D S D D H T T F L K I N Q G S S A Y A V G I N T F K Q T S G G T A I P I S W S G G V V P S V V N G L K T D I Y S F Q T F D G G A S L Y G I V V G Q N F S
------------------	----------------------	---

SEQ ID NO: 45	meta_gene _324030	MNTSTVTNNNAETTAIESLFAKKLLRSKGIAVIPPSTGSGKTREIARFA SNPKEYIDNIKNSNFSNGLCEIDESKKIKTIYISPQIKHCQDFISDIANDES CKDFCYEKRACRILNIFEVAEKVVGAYEDTKEKLNTGKTPSLLNE RLLYKDGGIGENGENKQIEQFIEILNGLDKSSNMSEQIKEELQSKAKSQ FRDIKTMIAKNYLNLEENPDFKDIELETYLKEPSLNWFVLLPAHFWD EINTYSLTVKMSSFTIRDVIFSKDLSSLLKPEEEQSFVFIDEADTASEELI DTESENATKNSSIDVIKLLITLSRILDFKDVFPNYSSQKEKKQFEKAIER GRKKFIEHFGDVDSNSTLIPTKEVKKLANSKVLHNYILRDSVETRVIK QGESAKKYMKDYYLSFPKNSEESKETPAFLITKDQIEEFPSDKYKVFE YKSFLKIASGLLNYFCEFVYPAINELIKKNEEDDNEMRSLNGLTQTFK ELYNVDDEFIKLLHDYKTQNYKKKITGASSGLLSYCDIGYEIFQVKVP LGGRPAELSRLVQGTPEQTVELAENSRVVLSATANVPSLKNFNL DFLSNRFGDYFDNFTMEDKKDFEAKLNYSNHNKSIELISDVSYELKYY EKDKEPDDTEETWLERKVEENFSYLSKKMCKMYLTNELTKEGIHRAN YYILLIKYYLAMKAAKTKANLMIFQPNLEKEVIETLLNIFDPKLNEEN AIFCANTEKLKTDFIEKVENAYLEGKTIFLITSLATMGKAVNFTFKAR EDEKLIHITPNGWIDDATKPAKRTFDGIAIGDINFSAKDNNESSNNE SSALRLLIDKITEVERLYATNLISNQIKRRIIQEMIINYESLYSFRGEFSTI RKLQGFYVYKEISQAIGRLYRTPNFSEKMLVTTKNNHDNLSTIKDSIE RKSFIETPLMTALMNEVQKEEITKKNSIEAKTmplknsGelfsRLLGTL LSDALKFKDKTSIQILEEMRRICIKYGVFLTEETYNSITSEQKDVDITSI KERLYQKVETSDFIKNGYKYKSHDDHSIIDFIDPKSSSEQGIPVSPTNCT IQQFRNLEGFYDYKENCGYTYDKVFNGEYIYILNPTAYNNLFKGALG EFVGKYIFEILFKLPLSRITDPEAYERADFFFADHNSTAIDFKCYSNPKV EKESLLEGIKNKAKALDIKEYHVINVFPYSTKGVPFTKETLLNEDGTA LLNSNGEPVVVKIVQATARPTSNCIVTDEFHQYILDTFLNKGN
SEQ ID NO: 46	meta_gene _295919	MENISFSREKALPFSLEKLETIFNNLIQRDTYSNKILKEPLSEFYRREIES EGKYRDNFLQIVEYTLSSLETIVKNPKRELLKISELQSINEIRSTDYKTM IWLGNGKPGKTLAEKIGAKGKILAPKNKYSIDKKENRVVVVYFKEAYK ILEERYKRYIENSVDIPENLKKIYERFYRIKREMINNELFFLDRPIDFSPN NALIDHRDYSVVNRLKHLKKYLEKLDYSENILLEAKKIVFLKLSYF IARLENIDIFDEILDIEELLKTKKKIIFYSSKQLQYLIKVILNKSksKIRIEF QKIFFNRDTKEVEKRDKEILDIDIIDTYENSASYYKLKVKDVEYNFND DDLKKILFENIKIDNLIKNKKESMN SERIINKYIYMFNSQLFIDNKAL EIKSYNKKLDNFMDTKDYFLSHSEQQAHYHINEIVSSDETIDIFPKYLE YIKEKRNIDKQNICIYSSLEALDSDSQKMLSSIYDSNFNKSYPIWRSLA TYAIKNSSKKWLENKEKFFVLDNSEIPTINTIEIEKNINRHPVIILEES ENEELKELSLQAYLKEYLEKYLNVYSIEMDEVEKTNLISSGGKVYETIF KRKRYLITNMNFYLEKDEDIINKVGNKFYNSVQKFVSKFLIDKRKLL IISDYLGEKYSLNGVDVKVIKEKELSLGKDEIIEKIKNNKNLWNEYLPN LTLETVKDGHFYNLDLIRENEDVEVIFGVEQKININENLVLPGMDVI KFPLYSQDSNNKKLYFLEIKSELFPLKENLVNLELIYSYGSKEPYKIK LKANGIDSSKFSTKWTEINKLKIVSLDYPEKNNKKNNYLGKIKILEKI DLNNNTNLKDYLKRKNRFRNYIIIEIERGNLERIKEVLDRNSKILALLE ILNKQEKEKGLLNEMIAVFLASFGVLIYDRIKVDIRKFEYRKRTLFY SLNNQLKLEDVLKYNKKDPEIIETVAEISWLDKVFKINKLAKEPELLEG ALKFLKYTLKSLNQKFGEELYEKWSKENLLWMLANRFKNYLEFILAIL TIKDKEKILKVLNKRDILKILYDIKAIDRKIQIDYPKLKEEFNKRIKLKF DRVVEQKKEVGLEAMS DLAYTVYC YLSGNNGSEA KIKEV LDDFND

SEQ ID NO: 47	meta_gene_237613	<p>MYLHGHHYYNEQNERIEVHIVTHGDKTDNQEISADTGDIQWTDDPVEI      ESQVSDFDVLLPQQATIRLQVRNFVADLFCADLREAVVNIYREGECL      FAGFLEPOSYSQGYSEEFDEIELSCIDVLTALKSKYGDVGSIGRLYHE      VKANARQRQSFQEIIITEMLTSLSHTSHIDLGGHSMSLYYDGSKAIDNQTDS      RYRIFSQLSINELLFLSDEEDNVWTQEEVLTELLKYLDVHVQVGFTF      YIFSWESVKRAASITWQNLLTGQNSETPYRKMDIRTGDVIGDDTTMSI      GEVYNQLLLTCKVEKMEQLIESPLEDSALRSDFPAKQKYMNEFISWG      TGKRAIEGRDLDVFNSTTAYDAASIVDWYIWVKRHPHWTMPHDNSL      QAGMSLSDYFGQTGRNQQAYLQWLGSHLGAALVAYGVATEMAR      GDNSPIAKIDMDNVLVLSVNGNGQDDQAKTYPKETDLKAAIPYAVYE      GKKAGGVFSPADEFQTTNYIVLSGKMLNPIMTQTATFRDLRTKPWTA      KNIFSGQPIEGKACVYGNVVKDKNGSEKYYTCKYWQKTDNSPKLN      EEPQWDEQGDGGWYPFTGTAPESYEYNYSAVGDGTDKISKVGLVAC      MLIVGDKCVVEKGSGSQIEDFEWRKYKERSACSSDDEYYQQSFTIGF      DPKIGDKLIGREYSLQNNISWKRGIDTEGMAIPIRKRDHVSGAVRFVIL      GPVNVLWDITRRHPTFRHTKWTEHAVPLLAHVSSIQIKQFEVKLHS      DNGLIEHLGDEHDIIYMSDAKTSFCNKDDLEFKITSALTYDESVQLGI      VNTPCLSTPVNMASGDGVQLVCNTLTGQQAKAEQLYVDAYYREYHE      PRVVLKQTFAVRTNGIVDLFTHYRQAFMDKTFFVQAINRSLTEGSAEL      TLKEINND</p>
SEQ ID NO: 48	meta_gene_35066	<p>MPTNYKTIINFRDGIQVDANDLVSNNGLVGIGTTIPREELDIRGNLIVE      NQANFRDVNVVGQSTFYGDINIAVGNSVGIGTTVPEATFQVGVGTTG      FTVDSNGNVTALTFTGSGANLTNLPTAVWTNPYPGAGTTINAFRPG      VSVTLQPQADFAVGDLIKLDTSGVGTFEGLVAKNITAVNASGSGQGN      VNGEVGTFSTITATDTAVIDKLDGNLIGLSTIAGTASTANSVYVTDEST      DTLLFPLFVDGAVLSGQIVAGNKEVKAGTNLQFDSANGLSATSLSA      AGGISIGPGGIMTATTFSGTATTALNASVAYAIAGQPDIQADKIDSLGI      NSIFIRNTGVSTFGGEVKVGNFLGVGATSSAIGKGMGVIGAADFSGAG      TFGGDLLVAGNLSVGGTFGGAVNITDVTAGEIIATGILSATTSSSCVLH      DTTITGNVVQVSAGKNLTVGQNLNSIGGTTFGSQINFQDASTQVAAGT      LFANLSGIITTGGINVGDLDISGTSYTGGSIAUTFGSILLNSNTGFVSCSS      IEAGTGIISCTGLNARTGEITGGGLNLTGPTTSNNFFQSTSGVSTFFDIDI      TGGTNSNIQLTRLGFNTSLGALGITEGIALWDDAEIYVNDSPASGIGIG      TTSGKRDNSVALYVGYGRDGAGNFINGQSVFEGGVGIGTMMGNDDG      NMLEVYKETVFHSYHTGVGGTDAGPARVGFETNKPRTRLDGFVTS      GFLRIPSYYNDDPNNTVPTNDTGSQGSLFFDTAINSISIKDMNDNWVGI      KTELSTGDDPAQYVQELFIGGVTDQANRLSAEQGVANIIQPYDEVG      NQGIGWGTAHMWYNKTFNKHQYKTNQGIGVATHYRSYVSTGSAID      IELDSSGTKVYITLPGIGSATFNLV</p>

SEQ ID NO: 49	meta_gene _524019	MWWKFYLIPDYVSIRR DINGHPVFLLIKYAFNDQDRQENKNLPRGGG FMVFDVELSVREADYPKIIAELQQSVNSWQQLKALADAAGNDVRG YSVNSWHYLNNGNFQFSTLSVNDLQLGLHPERPEAPPGDAPPKVIISQP TWKEGKFHVSAPQSTDVAHRVSEGPVSLVGNNVVSANMDLTGGA TFMEKTLTNLDGSGATDLTPIQVYELTFWARVPPVHLLTVDSRSL YEATKNIYHDYEGNGCDEDSINHSEQNLEMAVQSGLINIQIDTGTLSL SDDFVQQLRSGALKFVQDQIKDNFFDKKQAPPADDPTKDFVGSDKE IYYLKSDIDFKSVSIGYNEQIDSIVEWKANPQGTLQTFLAGVSPSEMKR YVRDVDLRDTFFMTLGLTTVFADWEHEPIAFVECQISYTGRDENNQ LIEKVQTFTFAKDHTAEFWDSLIGSKREYEYRWVGFFGHDAGEFTS WLTETTPKLNISIADPGKITIKVLAGNIDFAQTTKQVQVDLKYGPGPL EVPEEGTTLVLVNGQLEGNYERYIYSTWDHPVLYHARFYLKNEQVVE SDWQETVSRQLLINQPFLDQLKVQLVPAGSWDGVVQTVNLRYKDE LHSYHSEEAYTIKSADEFKTAIVLRDPNQRKFQYKILSTFKDGSTPA QTDWIDADGDQAVLIRVQQHPELKVKLLAGQIDFKVTPVVECTLHYD DLQGHIQKVDTFPFSKAEDA VVDFPLASDSRRTYRYQITYHTADGHTI PMPEVSTDTSVVI PPLEIPVISCTIFPKLVNFVQTPVVEVDFEYKDPDH HIEFEDTAVFTDSNPQSFRVQVDKASPRNYNLAVTYYTADGKVIQRD PVTLDKNKVVIPMYVATS
SEQ ID NO: 50	meta_gene _523517	MIYRDHQDKGLFYI PERPRLARNDGVPEFIYLVYKRDITDNP AFDPE TKASLGGGFLAFTV D LGVDDQQLAEMKQELARFSDGEV K LTPVQFHKG SVRLSISKDTADAPGTPPDQPKGLTFFEEVY GTTKPSLFGFN RAT FSVVLSQEVAALFEA ALQAGISPIGVYD LEFLGLRP AFNVRITA EYKRIYDHLEIEFGARGQIY AVALALDI DLA FQKL RDDGSIK VEVLSFTDDANLRKQADD A FNWFKTE LLKDFKSSLEPPSF MKQTNTTDLVGRLQSIFQGLNSAQTSPLN PVRGEPTKEPLTPA APPKKQEDGMK STADMNRAATQSGSESSGGGSGADRGISPFQIGFTLKYYRQEELKTRTFEFSEQAAVAR EAAPQGLFTTMVQGLDLSRAIQHVNLDSDFFKRLITTVSASDEFTIAGISTLGVNLEYPGTRKPGEDPLFVDGFVYKSDDLKPRTFTWLNDRKNTYRYQMDIHFTP DSPWVGKEGSVTS DWIITRSRQLTLDPMNEISLFDVQLTLGNMISGQINQVEVELRYQDSANDFNTQKTFLLKPGDPVTHWKLRLMDSEQKTYQYRITYFLQEGVRVQTDWV SSEDPTLVVAEPFKGTLNIRMVPLLDP TTLLEADVELMYHEEDTGYTRRVEKVFSPSDLKGQQISIPTLAENPTSYNTINIIRTDGSTYTLPP TTATTPVLV VSDGAGVTHRILVKLPSKDLSSFGLAALKV D LVGP GDDPDTASVLFPSQTDDKMPALVQPGDGGTFTYSYKVTGYTTQGLPIEGDSGTSSGPTLIVKIPTR

### Methods of producing a CRISPR/Cas system

#### Nucleic Acids and Methods of Introducing a Nucleic Acid in a Cell

Also provided herein are nucleic acids encoding any of the CRISPR-associated proteins or CRISPR-associated arrays as described herein.

Any of the isolated nucleic acids described herein can be introduced into any cell, e.g., a mammalian cell. Non-limiting examples of a mammalian cell include: a human cell, a rodent cell (e.g., a rat cell or a mouse cell), a rabbit cell, a dog cell, a cat cell, a porcine cell, or a non-human primate cell.

Methods of culturing cells are well known in the art. Cells can be maintained *in vitro* under conditions that favor cell proliferation, cell growth, and/or cell differentiation. For example, cells can be cultured by contacting a cell (e.g., any of the cells described herein) with a cell culture medium that includes supplemental growth factors to support cell viability and cell growth.

Methods of introducing nucleic acids (e.g., any of the exemplary nucleic acids described herein) and/or gene delivery vectors (e.g., any of the exemplary gene delivery vectors described herein (e.g., an AAV vector)) into cells (e.g., mammalian cells) are known in the art. Non-limiting examples of methods that can be used to introduce a nucleic acid (e.g., any of the exemplary nucleic acids described herein) and/or a gene delivery vector (e.g., any of the exemplary gene delivery vectors described herein (e.g., an AAV vector)) include: electroporation, lipofection, transfection, microinjection, calcium phosphate transfection, dendrimer-based transfection, anionic polymer transfection, cationic polymer transfection, transfection using highly branched organic compounds, cell-squeezing, sonoporation, optical transfection, magnetofection, particle-based transfection (e.g., nanoparticle transfection), transfection using liposomes (e.g., cationic liposomes), and viral transduction (e.g., lentiviral transduction, adenoviral transduction).

In some embodiments of any of the methods described herein, the method further includes formulating the CRISPR-associated protein, CRISPR-associated array, and/or guide RNA into a composition (e.g., a pharmaceutical composition).

Also provided herein are methods and compositions for specificity of transduction and/or infection, e.g., using any of the AAV capsid proteins or AAV virus serotypes. In some embodiments of any of the methods described herein, specificity of gene expression is determined, e.g., using any of the tissue-specific promoters and/or enhancers described herein.

#### Promoters

In some embodiments, the gene delivery vector (e.g., any of the exemplary gene delivery vectors described herein) can include a promoter sequence. In some embodiments of any of the gene delivery vectors described herein, the promoter sequence is a tissue-specific promoter. In some embodiments, the promoter is an H1 promoter. In some embodiments, a promoter is a ubiquitous promoter. Non-limiting examples of ubiquitous promoters include CAG, EF1 $\alpha$ , UBC, SV40, CMV, or PGK.

Enhancers

In some embodiments, the gene delivery vector (e.g., any of the exemplary gene delivery vectors described herein) can include an enhancer sequence. In some embodiments, an enhancer sequence is a CMV enhancer, a CAG enhancer, or a cHS4 enhancer.

5

Poly(A) Signal

In some embodiments, the gene delivery vector (e.g., any of the exemplary gene delivery vectors described herein) can include a polyadenylation (poly(A)) signal sequence. Poly(A) tails are added to most nascent eukaryotic messenger RNAs (mRNAs) at their 3' end during a complex process that includes cleavage of the primary transcript and a coupled polyadenylation reaction driven by the poly(A) signal sequence. In some embodiments of any of the gene delivery vectors described herein, the gene delivery vector can include a poly(A) signal sequence at the 3' end of the isolated nucleic acid encoding a fusion protein (e.g., any of the fusion proteins described herein).

15 The term “polyadenylation” refers to the covalent linkage of a polyadenylyl moiety, or its modified variant, to the 3' end of an mRNA molecule. A poly(A) tail is a long sequence of adenine nucleotides (e.g., 40, 50, 100, 200, 500, 1000) added to the pre-mRNA by a polyadenylate polymerase.

20 The term “poly(A) signal sequence” or “poly(A) signal” is a sequence that triggers the endonuclease cleavage of a mRNA and the addition of a sequence of adenosine to the 3' end of the cleaved mRNA. Non-limiting examples of poly(A) signals include: bovine growth hormone (bGH) poly(A) signal, human growth hormone (hGH) poly(A) signal. In some embodiments of any of the AAV vectors described herein, the AAV vector can include a poly(A) signal sequence that includes the sequence AATAAA or variations thereof.

25 Additional examples of poly(A) signal sequences are known in the art.

Internal Ribosome Entry Site (IRES) and 2A-Self-Cleaving Peptide

In some embodiments, the gene delivery vector (e.g., any of the exemplary gene delivery vectors described herein) can include an internal ribosome entry site (IRES) sequence. An IRES sequence is used to produce more than one polypeptide from a single gene transcript, and forms a complex secondary structure that allows translation initiation to occur from any position with an mRNA immediately downstream from where the IRES is located. Non-limiting examples of IRES sequences include those from, e.g., hepatitis C virus (HCV), poliovirus (PV), hepatitis A virus (HAV), foot and mouth disease virus (FMDV).

In some embodiments, the gene delivery vector (e.g., any of the exemplary gene delivery vectors described herein) can include a sequence encoding a “self-cleaving” 2A peptide (e.g., T2A, P2A, E2A, or F2A). A self-cleaving 2A-peptide is used to produce more than one polypeptide from a single gene transcript by inducing ribosomal skipping during 5 translation.

In some embodiments, the nucleic acid sequences are operably linked to a promoter or are operably linked to other nucleic acid sequences using a self-cleaving 2A peptide or an IRES sequence.

10 Compositions and Kits

Also provided herein are compositions (e.g., pharmaceutical compositions) that include any of the delivery systems, CRISPR-associated proteins, CRISPR-associated arrays, and/or guide RNAs described herein. Any of the pharmaceutical compositions can include any of the delivery systems, CRISPR-associated proteins, CRISPR-associated arrays, and/or 15 guide RNAs described herein and one or more (e.g., 1, 2, 3, 4, or 5) pharmaceutically or physiologically acceptable carriers, diluents, or excipients. In some embodiments, any of the pharmaceutical compositions described herein can include one or more buffers (e.g., a neutral-buffered saline, a phosphate-buffered saline (PBS)), one or more carbohydrates (e.g., glucose, mannose, sucrose, dextran, or mannitol), one or more proteins, polypeptides, or 20 amino acids (e.g., glycine), one or more antioxidants, one or more chelating agents (e.g., glutathione or EDTA), one or more preservatives, and/or a pharmaceutically acceptable carrier (e.g., PBS, saline, or bacteriostatic water).

In some embodiments, any of the pharmaceutical compositions described herein can further include one or more (e.g., 1, 2, 3, 4, or 5) agents that promote the entry of any of the 25 gene delivery vectors described herein into a cell (e.g., a mammalian cell) (e.g., a liposome or cationic lipid).

The pharmaceutical compositions provided herein can be, e.g., formulated to be compatible with their intended route of administration. In some embodiments, the compositions are formulated for subcutaneous, intramuscular, intravenous, or intrahepatic 30 administration. In some examples, the compositions include a therapeutically effective amount of any of the gene delivery vectors described herein.

Also provided are kits that include any of the compositions (e.g., pharmaceutical compositions), isolated nucleic acids, gene delivery vectors, or fusion proteins described herein. In some embodiments, a kit can include a solid composition (e.g., a lyophilized

composition including any of the gene delivery vectors described herein) and a liquid for solubilizing the lyophilized composition.

In some embodiments, a kit can include a pre-loaded syringe including any of the pharmaceutical compositions described herein.

5 In some embodiments, the kit includes a vial including any of the pharmaceutical compositions described herein (e.g., formulated as an aqueous pharmaceutical composition).

In some embodiments, the kit can include instructions for performing any of the methods described herein.

10 Cells

Also provided herein is a mammalian cell (e.g., a peripheral mammalian cell, a mammalian neural cell, e.g., a human neural cell) that includes any of the gene delivery vectors, fusion proteins, or isolated nucleic acids described herein. Also provided is a mammalian cell (e.g., a mammalian neural cell, e.g. a human neural cell) that is transduced

15 with any of the gene delivery vectors described herein, edited using lentiviral or CRISPR technologies, or otherwise engineered or modified to express any of the fusion proteins described herein. Skilled practitioners will appreciate that the gene delivery vectors

described herein can be introduced into any mammalian cell (e.g., any neural cell), that a variety of technologies can be utilized for modifying the genome of mammalian cells, and

20 that such modified human cells that secrete fusion proteins can be utilized as cell therapies. Non-limiting examples of gene delivery vectors and methods for introducing gene delivery vectors into mammalian cells (e.g., any neural cell, e.g., a human neural cell) are described herein.

In some embodiments, the mammalian cell is a human cell, a rodent cell (e.g., a rat cell or a mouse cell), a rabbit cell, a dog cell, a cat cell, a porcine cell, or a non-human primate cell. In some embodiments, the mammalian cell is present in a subject (e.g., a human subject). In some embodiments, the mammalian cell is an autologous cell obtained from a subject (e.g., a human subject) and cultured *ex vivo*. In some embodiments, the mammalian cell is *in vitro*.

30

**Methods of identifying CRISPR-associated proteins**

Provided herein are methods of identifying a Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR)-associated protein including (a) obtaining a plurality of genomic sequences, wherein a genomic sequence of the plurality of genomic sequences

comprises a CRISPR-associated array; (b) determining a subset of the plurality of genomic sequences comprising a plurality of coding sequences within a 20 kilobase (kb) sequence flanking region either at the 3' or 5' end of the CRISPR-associated array; and (c) analyzing a coding sequence of the plurality of coding sequences and thereby identifying the CRISPR-associated protein based on the coding sequence.

In some embodiments, the obtaining step comprises identifying, within the plurality of genomic sequences, a genomic sequence comprising a CRISPR-associated array.

Also provided herein are methods of identifying a CRISPR-associated proteins including (a) obtaining a plurality of genomic sequences; (b) selecting, within the plurality of genomic sequences, a genomic sequence comprising a CRISPR-associated array; (c) determining a subset of the plurality of genomic sequences comprising a plurality of coding sequences within a 20 kilobase (kb) sequence flanking region either at the 3' or 5' end of the CRISPR-associated array; and (d) analyzing a coding sequence of the plurality of coding sequences and thereby identifying the CRISPR-associated protein based on the coding sequence.

In some embodiments, the plurality of genomic sequences comprise one or more of genomes, wherein the one or more of genomes are selected from: a prokaryotic genome and metagenome. In some embodiments, the selecting step comprises using an algorithm selected from the group consisting of PILER-CR, and CRISPR Recognition Tool (CRT), and combinations thereof.

In some embodiments, the determining step includes filtering the genomic sequences according to the location of the genomic sequence relative to the 20 kb sequence flanking region. In some embodiments, the filtering can include selecting a genomic sequence that is located within the 20 kb flanking region. In some embodiments, the determining step also includes filtering the genomic sequences according to the size of the genomic sequence. In some embodiments, the filtering can include selecting a genomic sequence that is longer than 500 amino acids. In some embodiments, the determining step comprises using an algorithm selected from the group consisting of MetaGeneMark, and Prodigal, and combinations thereof.

As used herein, the term "analyzing" can refer to a process that includes filtering of a plurality of coding sequences based on the size of each coding sequence. In some embodiments, the filtering comprises selecting a coding sequence that comprises more than 500 amino acids (e.g., 550 amino acids, 600 amino acids, 650 amino acids, 700 amino acids, 750 amino acids, or 800 amino acids). In some embodiments, the filtering comprises

selecting a coding sequence that comprises more than 800 amino acids (e.g., 850 amino acids, 900 amino acids, 950 amino acids, 1000 amino acids, 1100 amino acids, 1200 amino acids, 1300 amino acids, 1400 amino acids, or 1500 amino acids).

In some embodiments, the analyzing step further comprises classifying the CRISPR-associated arrays. In some embodiments, the classifying of the CRISPR-associated arrays comprises selecting a CRISPR-associated array comprising three or more coding sequences (e.g., 4 or more, 5 or more, 6 or more, 7 or more, 8 or more, 9 or more, 10 or more, 15 or more, 20 or more, 25 or more, 30 or more, 35 or more, 40 or more, 45 or more, or 50 or more coding sequences) present in the 20 kb flanking regions. In some embodiments, the classifying further comprises determining a relative position of the coding sequence in the 20 kb flanking region relative to the CRISPR-associated array. In some embodiments, the classifying comprises calculating the coding sequence position within the 20 kb flanking region adjacent to the CRISPR-associated array, wherein the coding sequence could be classified based on the position relative to the CRISPR-associated array.

In some embodiments, the analyzing of the coding sequences comprises removing known CRISPR-associated proteins from the identified CRISPR-associated proteins. In some embodiments, the analyzing of the coding sequence comprises using one or more algorithms selected from HHMSCAN and RPS-BLAST. In some embodiments, the analyzing of the coding sequence further comprises determining the presence of a structural domain. In some embodiments, the analyzing of the coding sequence further comprises determining the presence of a functional domain. In some embodiments, the functional domain comprises a functional domain selected from a DNA binding domain, a RNA binding domain, a nuclease, a helicase, a restriction domain, and/or a structural maintenance of chromosomes (SMC) domain. In some embodiments, the analyzing of the coding sequence further comprises determining whether the coding sequence starts with a Methionine

Also provided herein are computer implemented methods including (a) obtaining a plurality of genomic sequences; (b) selecting, within the plurality of genomic sequences, a genomic sequence comprising a CRISPR-associated array; (c) determining a subset of the plurality of genomic sequences comprising a plurality of coding sequences within a 20 kilobase (kb) sequence flanking region either at the 3' or 5' end of the CRISPR-associated array; and (d) analyzing a coding sequence of the plurality of coding sequences and thereby identifying a CRISPR-associated protein based on the coding sequence.

#### Methods of treatment

Also provided herein are methods for treating a condition or disease in a subject in need thereof, the method including administering to the subject any of the systems described herein, wherein the spacer sequence is substantially complementary to a target nucleic acid associated with the condition or disease; wherein the CRISPR-associated protein associates 5 with the RNA guide to form a complex; wherein the complex binds to the target nucleic acid sequence; and wherein upon binding of the complex to the target nucleic acid sequence the CRISPR- associated protein cleaves the target nucleic acid, thereby treating the condition or disease in the subject.

In some embodiments of these methods, the method can result in at least a 2.0-fold 10 (e.g., at least a 2.5-fold, at least a 3.0-fold, at least a 3.5-fold, at least a 4.0-fold, at least a 4.5-fold, at least a 5.0-fold, at least a 6.0-fold, at least a 7.0-fold, at least a 8.0-fold, at least a 9.0-fold, at least a 10-fold, at least a 15-fold, at least a 20-fold, at least a 30-fold, at least a 40-fold, at least a 50-fold, at least a 60-fold, at least a 80-fold, at least a 100-fold, at least a 120-fold, or at least a 150-fold) decrease in the level of one or more symptoms associated with the 15 condition or disease as compared to the level of the one or more symptoms associated with the condition in the subject prior to the administering. In some examples of these methods, the method can result from about a 2-fold to about a 150-fold, about a 2-fold to about a 100-fold, about a 2-fold to about a 50-fold, about a 2-fold to about a 25-fold, about a 2-fold to about a 10-fold, about a 2-fold to about a 5-fold, about a 5-fold to about a 150-fold, about a 5-fold to about a 100-fold, about a 5-fold to about a 50-fold, about a 5-fold to about a 25-fold, about a 5-fold to about a 10-fold, about a 10-fold to about a 150-fold, a 10-fold to about a 100-fold, about a 10-fold to about a 50-fold, about a 10-fold to about a 25-fold, about a 25-fold to about a 150-fold, about a 25-fold to about a 100-fold, or about a 25-fold to about a 50-fold, decrease in the level of one or more symptoms associated with the condition or disease 20 as compared to the level of the one or more symptoms associated with the condition in the subject prior to the administering.

25

In some embodiments, the condition or disease can include conditions such as cancers, neurodegeneration, cutaneous conditions, endocrine conditions, intestinal diseases, infectious conditions, neurological conditions, liver diseases, heart disorders, or autoimmune 30 diseases. In some embodiments, the condition or disease can be a cancer. In some embodiments, the cancer is selected from a bladder cancer, breast cancer, cervical cancer, colon cancer, endometrial cancer, esophageal cancer, fallopian tube cancer, gall bladder cancer, gastrointestinal cancer, head and neck cancer, hematological cancer, Hodgkin lymphoma, laryngeal cancer, liver cancer, lung cancer, lymphoma, melanoma, mesothelioma,

ovarian cancer, primary peritoneal cancer, salivary gland cancer, sarcoma, stomach cancer, thyroid cancer, pancreatic cancer, renal cell carcinoma, glioblastoma and prostate cancer. In some embodiments, the cancer can be a B-cell acute lymphoblastic leukemia, lung cancer, esophageal cancer, multiple myeloma, or cervical cancer.

- 5       In some embodiments, the condition or disease can be a neurodegenerative disease. In some embodiments, the neurodegenerative disease can be Alzheimer's disease, Huntington's disease, Duchenne muscular dystrophy (DMD), frontotemporal dementia, ryanodine receptor type I (RYR1)-related myopathies, cystic fibrosis, or autosomal recessive juvenile parkinsonism.
- 10      In some embodiments, the condition or disease can be a blood disease or a hemoglobinopathies. In some embodiments, the blood disease can be sickle cell anemia or beta thalassemia. In some embodiments, the condition or disease can be an eye disease. In some embodiments, the eye disease can be retinitis pigmentosa, leber congenital amaurosis, specific retinal dystrophy, or autosomal dominant cone-rod dystrophy. In some embodiments, 15 the condition or disease can be human immunodeficiency virus (HIV), diabetes, autism spectrum disorder, genetic liver disease, or congenital genetic lung disease.

## EXAMPLES

### Methods

20      *Identification/Prediction of candidate CRISPR associated proteins*

An exemplary method of identifying candidate CRISPR-association proteins is as described as shown in **Figure 1**. In order to identify new candidate CRISPR associated proteins 179,804 prokaryotic genomes and 3,396 metagenomes deposited in Genbank from June 1st 2016 - April 21, 2020 were downloaded and analyzed (**Figure 2**). PILER-CR (see, 25 e.g., Edgar et al., *BMC Bioinformatics*, 8, 18 (2007)) and CRT (CRISPR Recognition Tool) (see, e.g., Bland, C. et al., *BMC Bioinformatics*, 8, 209 (2007)) were used to identify CRISPR arrays (or "arrays") (**Figure 2**). Arrays located on sequence contigs shorter than 3 kilobases (kb) were filtered out and 20 kb flanking sequences on both sides of the arrays were extracted. As shown in **Figure 3**, protein sequences were predicted from the 20 kb flanking 30 sequences using MetaGeneMark (see, e.g., Zhu, et al., *Nucleic Acids Research*, 38 e132–e132 (2010); hereinafter "Zhu") and Prodigal (see, e.g., Hyatt et al., *BMC Bioinformatics*, 11, 119 (2010); hereinafter "Hyatt"). Proteins predicted from the two software were merged and sequences shorter than 500 amino acids were filtered out. Subsequently, protein sequences

were clustered using MMseqs2 (see, e.g., Steinegger, *Nat. Biotechnol.* 35, 1026–1028 (2017)) with a sequence identity threshold of 90%. Clusters with less than 3 members were filtered out because they may represent very rare or mis-predicted sequences. For each cluster, the position of each gene (coding sequence) relative to the array was calculated. Ranks were

5 assigned for each cluster, with rank 1 indicating the gene immediately adjacent to the array, rank 2 indicating the second gene adjacent to the array, rank 3 indicating the third gene adjacent to the array, rank 4 indicating the fourth gene adjacent to the array, rank 5 indicating the fifth gene adjacent to the array, rank 6 indicating the sixth gene adjacent to the array, and so forth. Clusters with a median rank above 7 were subsequently filtered out since known  
10 effectors are usually located in proximity to the array (**Figure 3**). This analysis produced 10,913 candidate clusters. **Figures 6A and 6B** shows further annotation and filtering done on the 10,913 candidate clusters. **Figures 7 and 8** shows a summary of the method as described herein.

15 *Annotation/classification of predicted CRISPR associated proteins*

In order to annotate and classify the 10,913 cluster sequences adjacent to the CRISPR arrays, from each cluster a representative sequence was searched against the prokaryotic subset of the non-redundant protein database (bacteria+archaea) using blastp in order to 20 annotate protein sequences and identify known CRISPR genes. Protein sequences matching known CRISPR genes with e-value cutoff of 1e-10 and query coverage of 50% were considered orthologous to known CRISPR genes. Furthermore protein sequences were 25 searched with HMMSCAN against known CRISPR-related profiles from (see, e.g., Burstein, D. et al., *Nature* 542, 237–241 (2017); hereinafter “Burstein”) and with RPS-BLAST against a collection of CRISPR profiles. These protein clusters represent orthologs and are considered known CRISPR associated proteins and thus filtered out or separated for further analysis.  
From the total 10,913 clusters, 3465 clusters were considered known CRISPR and 7,642 novel potential CRISPR associated candidates (**Figure 6A**).

To further annotate the remaining 7,642 protein clusters, for each candidate protein, functional domains were predicted by running RPS-BLAST on CDD database and 30 HMMSCAN against Pfam and associated GO (Gene Ontology) terms were added using Pfam2Go mapping software. Protein clusters were subsequently grouped in subsets based on the presence/absence of characterized and putative domains.

## Results

*Bioinformatic search for novel CRISPR associated proteins*

To identify novel CRISPR associated proteins, 179,804 prokaryotic genomes and 3,396 metagenomes deposited to Genbank from June 1st 2016 - April 21, 2020 were downloaded and analyzed. Using PILER-CR and CRT (CRISPR Recognition Tool), 230,443 CRISPR arrays were identified with 187,324 derived from prokaryote genomes, and 43,119 from metagenomes. Given that most CRISPR class 2 effectors (i.e. single effector proteins like Cas9's, Cas12's, Cas13's) are located in close proximity to their arrays (Makarova, et al., *Nat. Rev. Microbiology*, 18: 67–83 (2020); hereinafter “Makarova”), the search for novel CRISPR associated proteins was limited to a 20 kb window flanking the arrays. Putative protein sequences within the flanking sequences were predicted using MetaGeneMark (Zhu) and Prodigal (Hyatt), filtering out sequences shorter than 500 amino acids as novel class 2 effectors are generally large multidomain proteins (Makarova). **Figures 4A-4B** show the Cas9 size distribution by member and cluster count. This prediction resulted in 829,464 total protein sequences located adjacent to the CRISPR arrays. Given that many of these are likely to be orthologous, protein sequences were clustered using MMseqs2 (Mirdita et al., *Bioinformatics*, 35: 2856–2858 (2019)) with sequence identity threshold set at 90% resulting in 171,774 unique clusters. Clusters with fewer than 3 members (very rare sequences or possible mis-predictions) were filtered out leaving 25,623 clusters. The number of sequences associated with each cluster ranged from 3 to 18,997 (**Figure 3**). These 25,623 clusters were further analyzed to determine the position of each gene (coding sequence) relative to the array was calculated and assigned a rank within the cassette of genes based on the relative position to the array. As described above, rank 1 means that the gene is immediately adjacent to the array and rank 2 indicating the second gene adjacent to the array, and so forth. Known effectors are usually located close to the array. For instance, Cas9-type effectors are usually ranked 3-4, while Cas13-type effectors –are typically ranked 1-2, and Cas12-type effectors are more broadly distributed, but still close to the array (**Figures 5A-5C**). Filtering out all clusters with median rank above 7 reduced the cluster number to 10,913 (**Figure 3**).

To annotate protein sequences and identify known CRISPR proteins, representative sequences for the 10,913 clusters were searched against the prokaryotic subset of the non-redundant protein database (bacteria+archaea) using blastp. Protein sequences matching known CRISPR genes with e-value cutoff of 1e-10 and query coverage of 50% were considered orthologous to known CRISPR genes. Additionally, protein sequences were searched with HMMSCAN against known CRISPR-related profiles (Burstein) and with RPS-BLAST against collection of CRISPR profiles. Hits for both of these searches mostly

overlapped blastp-identified CRISPR sequences, with a few exceptions, which were also added to the CRISPR cluster ortholog set. Together, from the 10,913 clusters, 3465 clusters were considered orthologs to known CRISPR proteins leaving 7,642 potential cluster candidates to be further characterized. Given that many of the 10,913 clusters were generated with a stringent 90% identity using MMseqs2, these clusters were similar and therefore additional filtering was performed. To further reduce the number of sequences, 10,913 clusters can be further clustered with MMseqs2 using default settings, which requires the sequences to overlap by at least 80% (query coverage 0.8). MMseqs2 with default settings generated 4,205 “superclusters”. The supercluster classification reduced the number of known CRISPR-associated clusters to 343 and the number of unknown CRISPR superclusters to 3862. To narrow down the two lists (clusters and superclusters), proteins were further analyzed and protein domains were predicted by running RPS-BLAST on the CDD database and HMMSCAN against Pfam (**Figures 6A-6B**). Associated GO terms were added using Pfam2Go mapping.

For the 3465 clusters consisting of 51,094 orthologs of known CRIPSR proteins, and 343 superclusters consisting of 2614 clusters we found numerous class I systems which have effector modules composed of multiple Cas proteins (e.g. Cas1- 4, 5-8, 10-11), and numerous class II systems which encompass a single multidomain crRNA-binding protein (e.g., Cas9, Cas12, Cas13 etc.).

#### *Predictions of TracR-RNAs*

To annotate known candidates, the arrays were classified into class 1, 2, or unclassified based on the identified CRISPR-related proteins associated with each array. For each array with flanking regions length of at least 3 kb, all those CRISPR-related proteins were collected and if they consistently fell into class 1 or 2 that array was classified as such. If an array had no identifiable CRISPR proteins that could distinguish the class, like arrays flanked by Cas1/Cas2/Cas4 only or no Cas proteins, they were marked as unclassified. If an array had proteins from both classes, it was marked ambiguous. That is because if a cluster was classified as 2, that meant that the array already had an effector protein such as Cas9/Cas12/Cas13 since those are the only proteins that can distinguish class 2 reliably. Those arrays were unlikely to have yet another effector. If the array was classified as 1, which is the majority of classified arrays, naturally, it also could have been discarded since class 2 effector were of primary importance. As such, the aim was to narrow down the candidate CRISPR-associated proteins by further considering only unclassified or ambiguous arrays.

*Choosing the top 50*

Further filtering of the candidate clusters produced a list of 50 candidate proteins to be used for functional assay. Candidates were divided in four main categories: proteins with no

5 blast hits, proteins with no predicted domains and blast hits against hypothetical and unknown proteins, proteins with predicted domains and blast hits against hypothetical and unknown proteins only and proteins with predicted domains and blast hits against characterized

proteins. For each category protein shorter than 800 amino acids (aa) and proteins not starting with methionine (Met) were filtered out. The first category included 25 candidates, 6 are

10 associated with classified arrays and thus not considered for further analysis. Since the majority of the proteins were filtered out because they had predicted domains with a structural potential function or were low complexity proteins including many SR repeats, the protein length threshold for this category was changed to 650 aa and four potential candidates were selected for functional analysis. The second category of proteins with no predicted domains

15 and blast hits against hypothetical and unknown proteins contained 347 candidates of which 120 are associated with an already classified array and thus filtered out. From the remaining 227 proteins, 175 proteins were excluded for being shorter than 800 aa and 14 candidates were excluded for not starting with Met. In addition, proteins with high presence of low complexity/repeats regions were selected out and selected 15 candidates for further analysis.

20 The third category included 1644 proteins with predicted domains and blast hits against hypothetical and unknown proteins of which only 552 candidates were longer of 800 aa. Exclusion of 152 proteins as already associated with classified arrays and proteins not starting with Met left 322 candidate proteins. From this shorter list, 15 were selected based on putative function of the hypothetical domains. Proteins with DNA/RNA binding domains,

25 nucleases, helicases, restriction and SMC domains were included in the final list for further functional analysis. The most abundant category is represented by proteins with predicted domains and blast hits against characterized proteins with 5329 candidates of which 1442 were above 800 aa. After filtering out proteins associated with classified arrays and proteins not starting with Met, the candidate number decreased to 758. SEQ ID NOs: 1-50 represent

30 proteins with DNA/RNA binding domains, nucleases, helicases, restriction and SMC domains that were selected for further analysis. The CRISPR arrays and spacer sequences corresponding to the CRISPR-associated proteins of SEQ ID NOs: 1-50 are listed in **Tables 1-5**.

[Table 2] CRISPR arrays and spacer sequences for candidate CRISPR-associated proteins

Table 2	Protein ID	other CAS protein in	array name	Domain (y or n)	class type	Notes	repeats	spacer sequence (each row denotes a new spacer)	Corresponding SEQ ID NO:	CRISPR - associated protein
	gene_5155-455 GeneM-ark.lmmnl1-389_aa+1 3650 178 19	cas1-cas2-cas4	piller_crt array_VBTK0	cas9	class 2	Cas9-Streptococcus thermophilus	GTTTGTAGAGCT GTGTTGTTTG AATGGTTCCAA AAC (SEQ ID NO: 51)	AGAATACAACATTGTCCTTAATAGGAGACAC (SEQ ID NO: 101) GAATCATGATTGTTATCTGGGCTICA (SEQ ID NO: 102) AAAGAAAATTAAAAAAACCTTAGCGAACACT (SEQ ID NO: 103) TTTCGATAAGACTTCCTCAAAACAAACAT (SEQ ID NO: 104)	SEQ ID NO: 1	

(SEQ ID NO: 117) TACAGCTCTGGTTTCGTTATCCCTATGT (SEQ ID NO: 118) CGCTAGGGTCTCTGGTACGCTGAGGTCTC (SEQ ID NO: 119) CCTGACGCATATGAAATCCTAACGGTCAG (SEQ ID NO: 120) AAAATCAATTAATACATGTGTGAAACAAG (SEQ ID NO: 121) AAGGCATGGACGACAATAATAATTGAAAG (SEQ ID NO: 122) GAACAAAGAAACTTATGAAAGTCGAAAAACGA (SEQ ID NO: 123) TTCGATAAGACTTCCTCAAAACAAACAT (SEQ ID NO: 124)					
gene_3815 793 GeneM ark.hmm 1 090_aa+1 4361 17633	cas1/ca s4 - cas2	piler_crt array -PVTZ0 100000 2.1_339 025- 339866: 40841	cas12b cas12	cas12b - Laceyella sediminis  (SEQ ID NO: 52)	CTTAAAGTGTAT TAGATGAATTA AAATGTGATTAG CAC  (SEQ ID NO: 125) GTCGAATTCCTATTTGGGCTTGAAAGGTTCAGCATTC AAATG  (SEQ ID NO: 126) GCCGAAGATACTGGTGAGAAGTTTCAGCATTC AAATG  (SEQ ID NO: 127) TTAACCTTATTTGATGTTATTTTAACCTTATTTGGAG  (SEQ ID NO: 128) GGAAATCCCTTGATTTCTGGAAATATTC CACCTTTAAAGAACATATACAAACGATCTCGAAGCGG  (SEQ ID NO: 129) GCTAACACAAATCAACACGATTC CACCAAACAATGGTTTTC  (SEQ ID NO: 130) CCATTGATACAGGCAATCTCATGTC GATTGTTGCT  (SEQ ID NO: 131) GGGAGATAAGGTAAAACA TAGACTCCAATAGTGCT  (SEQ ID NO: 132) TGAGTACATCGGGGGATAAAAAGCCGATAGGAATC  (SEQ ID NO: 133) TTAACCTGCCCAATTTCCATTTCAGCTAACGATC  (SEQ ID NO: 134) TTAACCTGCCCAATTTCCATTTCAGCTAACGATC  (SEQ ID NO: 135)

gene_2964 877 GeneM ark.hmm 1 305_aa+ 1 5109 19026	cas1/ca s4 - cas2	piler_crt array_- NALN0 100001 2.1_702 24- 71132:4 0908	cas12a	class 2	cas12 a- Firmicutes bacterium	SEQ ID NO: 3
gene_4147 644 GeneM ark.hmm 1 412_aa+ 2 0684 24922	cas1/4	piler_crt array_- ORUJ0 100000 6.1_107 860- 108175: 40315	cas13a	class 2	cas13a	GTAAAGTAAC TAAATAATTIC TACTGTGTGTA GAT (SEQ ID NO: 138)
meta_gene 174274 GeneMark.hmm 921_aa  -66 2831	no	piler_ar ray_OD FV0100 4017.1_- 2979- 3331:35 77	cas13d	class 2	CasRx (From metagenome s)	ATGGCTGTCTGTTAAGGTGTCTCTG (SEQ ID NO: 136) TTAATTITATTGTGCTGTTAGT (SEQ ID NO: 137) ATTTCACCGTACAGGAGAACACGAT (SEQ ID NO: 138) ATCGACAGGGATAACACAGGCATAGCT (SEQ ID NO: 139) CTATACGCCAGAGGGTAGGCCTGGAA (SEQ ID NO: 140) AAGTAATTGAAAAATAATCATATAGTAAT (SEQ ID NO: 141) CAAATAATCGATAAGGCTCCAGAAGAA (SEQ ID NO: 142) CTATTGGGATACTCTCAIT'AAAAGT (SEQ ID NO: 143) CAAATCTTAICTTTATCTCTTGAG (SEQ ID NO: 144) TACTATGCCGAATAATTTAAAGCTGT (SEQ ID NO: 145) AAAATAATGAAGCTCCCTTACAATTTC (SEQ ID NO: 146) ATAACAAACCGCCCTGTTAGTACTAGG (SEQ ID NO: 147) ATAIACATIAAATATGGCTGGGATACA (SEQ ID NO: 148)
						TTTGAGGTGCCCTTGTAAACCTTGAATCCTAAATTCTA (SEQ ID NO: 149) GTTTGGTACGGTTTTATTTCTTATAGTTTATATATATG (SEQ ID NO: 150) GTCATATTACAAATGCTTCATACTGCTGTCA (SEQ ID NO: 151) AAGCCAACCTAAATCAACACCATCATCACAAAC (SEQ ID NO: 152)

gene_4200_106 GeneM ark.hmm 5 68_aa+ 66 46 8352	no (additi onal cas13d )	crt_arra y_QTX T01000 036.1_6 154- 6455:16 264	cas13d	class 2	*	CTACTACACTG GTGCGAATTG CACTAGTCTAA AACT (SEQ ID NO: 56)	SEQ ID NO: 6
meta_gene_524079 GeneM ark.hmm 5 68_aa+ 66 46 8352	no	crt_arra y_WNG K01002 380.1_7 01- 1392:21 392	n	unclass ified	**	Not included	SEQ ID NO: 7
meta_gene_524079 GeneM ark.hmm 5 68_aa+ 66 46 8352	no	meta_cr t_array_ WNGG G01011662 .1	n	unclass ified	GTCGCTAATGG AGCGGCTCTC GGTTGAGATT (SEQ ID NO: 57)	GAAACTTGAGCTTCCATGAAACCGAATAAGTACTTA (SEQ ID NO: 161) GAAACATTCACCCAAACCCCTCGATATCAAAAGCCATAATCAT (SEQ ID NO: 162) GAAACCCGTTAGCTTGATAACGAGAACGCCCTCGGCCTTA (SEQ ID NO: 163)	SEQ ID NO: 8
meta_gene_336895 GeneM ark.hmm 5 68_aa+ 66 46 8352	no	meta_cr t_array_ WNGG G01011662 .1	n	unclass ified	GTCGCTAATGG AGCGGCTCTC GGTTGAGATT (SEQ ID NO: 57)	GAAACTTGAGCTTCCATGAAACCGAATAAGTACTTA (SEQ ID NO: 161) GAAACATTCACCCAAACCCCTCGATATCAAAAGCCATAATCAT (SEQ ID NO: 162) GAAACCCGTTAGCTTGATAACGAGAACGCCCTCGGCCTTA (SEQ ID NO: 163)	SEQ ID NO: 8
meta_gene_336895 GeneM ark.hmm 5 68_aa+ 66 46 8352	no	piler_cr t_array_ OEIL01 000106. 1_2920 9- 29855:4 0646	n	unclass ified	ATAAAGAATTAA ACATAAGTTG TTTAAAT (SEQ ID NO: 58)	ACTCCAAACATAACCTCTTAAGTACTTAAATACTTCTT (SEQ ID NO: 164) TCTCTTGTCAATTCTCTAAATTATAATTTCCT (SEQ ID NO: 165) AAAAGTGGATTATCTCCACTGGAAGTGGTACTCAA (SEQ ID NO: 166) GGTGTCTCTTTGTATTGATTCTTCTTATTATT (SEQ ID NO: 167) AAAAGAAGAATTACATTAAATTAAAGA (SEQ ID NO: 168) ACTGTAACCTCGATTTTAAAAAATATTTTACTTC (SEQ ID NO: 169) AAAATGAGATAATTATAACGAATTATT (SEQ ID NO: 170) ATTCCAGTTAAATACTCTTCTTATTGGGACACC (SEQ ID NO: 171) AGAGGAAATTGGAAATA (SEQ ID NO: 172)	SEQ ID NO: 9

meta_gene	no	crt_aray_OEE	n	SEQ ID NO: 10
-321445GeneMark.hmm[675_aa]	863.17	ara	Not included	
-	543-	y_OEE		
5020 7047	7748.15	O01000		
	683	863.17		

\* short version casRx ([Ruminococcus sp.])

\*\*\* crispr software failed to recognize array and spacer only repeats not spacer

[Table 3] CRISPR arrays and spacer sequences for candidate CRISPR-associated proteins

gene_38_20393	cas2-cas3-CaM	no	piler_cr	y	uncla	(Actino	GTCGCC	TGTTGAACGACCCCTGAGGCCACGCAGCTGCAG	SEQ ID
GeneM							CGCACGCC	(SEQ ID NO: 173)	NO: 11
ark.hm							CGGGATG	ATCGACGCCAGCGACATCGGCTGGTCCAGGC	
m 1351							TTCGG	(SEQ ID NO: 174)	
aa +23								GTGAAACATCGCGGGGATCACGATCAAGCGGA	
-286 273								(SEQ ID NO: 175)	
41								TGGCTGAGCGGCCACCGTCAAGGGCGGGCTCC	
								(SEQ ID NO: 176)	
								GTTACGAGGGTGGGGGGGGCTTGAGCAG	
								(SEQ ID NO: 177)	
								TCCAGGGACATTAACGCCCGTGGGCCGATC	
								(SEQ ID NO: 178)	
								TCATGGGGCAAGCCAAGAAAAGGGCGATTAA	
								(SEQ ID NO: 179)	
								TACCTGGGGGGCGCGGGGGCGAGCTGAGAA	
								(SEQ ID NO: 180)	
								CCACGGGGGACCCATCGGAAGGGGCCCTCG	
								(SEQ ID NO: 181)	
								CGGCCAGCTAGCCCCGGTGCCGCTGGTCTCC	
								(SEQ ID NO: 182)	
								TGCTCACCGCCCTACCGCGATGGATCCTGAACGC	
								(SEQ ID NO: 183)	
								AAGCCGGCGGAAGGTGCAGGGATGGC	
								(SEQ ID NO: 184)	
								ACTGCAAGCGACTCATCGACGAACAGGCAGGT	
								(SEQ ID NO: 185)	
								CGGTTCCTCGTTCATCGTCGGCTCTCTCTG	
								(SEQ ID NO: 186)	
								GGCGCACCGGATGCCAACGCTACCGACGA	
								(SEQ ID NO: 187)	
								GATTGTGTTAGGGGGGGACCTACAGAACCC	
								(SEQ ID NO: 188)	
								GTGTCTCCTACTGGTCCGGTGGGGAAAGAGCG	
								(SEQ ID NO: 189)	
								CTGGAGGTACATCGCCGCCAGGTGCCGAGTT	
								(SEQ ID NO: 190)	
								CGGACCAGGCTGGCCAGGGGCCAGGGAGAC	
								(SEQ ID NO: 191)	
								GAGTGTAGCTTCGATCTCGCCAGCACGTT	
								(SEQ ID NO: 192)	
								CTGTTCTGAGGCTCGAGCTGGGTGACC	
								(SEQ ID NO: 193)	
								AAGGCCGGCTTCAGGCTACGGCGGTACCT	
								(SEQ ID NO: 194)	
								ATGATGGAGCTGGTGGCCCAGCTCCCCCG	
								(SEQ ID NO: 195)	
								CACGCCCTCTGATCCCGACACCAAGGAGAGAC	
								(SEQ ID NO: 196)	
								TCATGGATGTCGTCCTGGGTGGGGCCGCT	



gene_77 1418 GeneMark. hmm 14 52_aal-[271]7 069	no	no	piler_cr t_array _CABJ CG010 000021 .1_238 1- 2613:2 2613	y	uncl assifie d	GTCAGCC TTATGGAG GCGTGTGG ATTGAAAT (SEQ ID NO: 61)	AACCCGATGGGAAGGTCCTGCCGCTCTGGCTGC (SEQ ID NO:211) TTCCTGCGGTTCTGGGGAGACCAAGATCAAGTCGT (SEQ ID NO:212) GTAAGCTGTCAGGAGATAATGGTGCAGTGTTTCGG (SEQ ID NO:213) CGACAGCTGCGCCGGCAAGTCAAAGGGGGCAACGGGACT (SEQ ID NO:214)	SEQ ID NO: 13
gene_14 33645  GeneM ark.hm m 1422 -aa +54 -89 9757	no	no	piler_cr t_array _DCO L01000 139.1 -10233- 10618: 10617	y _topois meras	uncl assifie d	GTCAAGT GAGATCAG CCGTTCAG GCTGTTGA AAC (SEQ ID NO: 62)	GCTATAGTGTCCGGTTCCGTTTCCGATT (SEQ ID NO:215) AACCATGCTAACCGCACAGGGGATAATAATTTG (SEQ ID NO:216) CTTGTGGTGCCTAACGCTCAACTACTGCGCTGC (SEQ ID NO:217) ACCACCGCGCTTGAAACGGGGAAAATTCGTTGGCTAT (SEQ ID NO:218) CACCATAACGGTGCCAGAATCCGTAAGGACACTGG (SEQ ID NO:219)	SEQ ID NO: 14
gene_44 26209  GeneM ark.hm m 1255 -aa +28 -994 327 61	WYL		piler_ar ray_RQ NV010 00008. 1_1590 -35- 15916 :40131 61	y	uncl assifie d	TAACTAAG TGGAAAC T	CAAGTGCTCATGGTTAATGAAGGCAGGAGATTGG (SEQ ID NO:220)	SEQ ID NO: 15
gene_54 11831  GeneM ark.hm m 1213 -aa +12 -801 164 42				y	uncl assifie d	GACTAAAT CCAAGTAG ATTGGAAT TTAAC (SEQ ID NO: 64)	GCACTGATTCATATTGAACTCTAATT (SEQ ID NO:221) TGAAAAAACTTCCAAACACGCTGACAAGGGAAACTA (SEQ ID NO:222) ATCGAAAAATTITACGTTAAGAGAGCTTCTGAAAGA (SEQ ID NO:223) AACTCAGGAAAATCAACGTCAGGAACACTAAACGGAAA (SEQ ID NO:224) GCAACTCCTAACATGCCCTAATTACACCGA (SEQ ID NO:225) TCGGCAGTGGAGCGCCITAAGGAAGGGGAAAT	SEQ ID NO: 16

				(SEQ ID NO: 226) AATGTAAGCCCTAACATCTCCATGATGCCATACCTCA (SEQ ID NO: 227) TTTTATCGATTCTCATCACAAATTGAGCAACATCTT (SEQ ID NO: 228)	SEQ ID NO: 17
gene_94 1761 GeneMark. hmml1 23_aa-  22964  26335	y	unclassified	ATTAAAT ACATCCTA TGTTATGGT TCAATCA (SEQ ID NO: 65)	TGGCCTAGCATGGCAGCTAGGAAAAATAAACCT (SEQ ID NO: 229) CCTACAGATGTGCAAATGGTCTAAATAAAATA (SEQ ID NO: 230)	SEQ ID NO: 18
gene_15 46948  GeneM ark.hm ml949_ aa-  10158  13007	y	unclassified	GTAGCATT CACCCCCA AGGTGGG TGCCCCGT GAAAC (SEQ ID NO: 66)	CTCC CCTGTTGCGTTCATGCCCTGGGGAGTT (SEQ ID NO: 231) GAAACTGCTATCGCTATTCGCTCGTTTTGTCATACGCTTA (SEQ ID NO: 232) CTCC CCTGTTGCGTTCATGCCCTGGGGAGTT (SEQ ID NO: 233)	SEQ ID NO: 19
meta_gene_154 50 Gene Mark.h mm 803 aa 14 847 172 58	y	unclassified	GTTTCAGA GCAGATGC TGGCTTGA GTTAAGAT GTAAC (SEQ ID NO: 67)	CGTCAATTTCGGCGTGAAGAACATGGGGATATAGGC (SEQ ID NO: 234) CGCGACGGCCAGAACATACGCTCCAGTGCTTCGTTG (SEQ ID NO: 235) GCGAGGGCCAGAACGGCCAGAAAAACGAGACTGCC (SEQ ID NO: 236) CCGGCGGCCACACGCTGGGGATTCTCTACCA (SEQ ID NO: 237) ACAAAGAGCTGGCTACGAGAACGGGATTGAATGCGT (SEQ ID NO: 238) AGTACGACCCGACGCTTGGAAACAAATAACCCCG (SEQ ID NO: 239) TGAAGGGCTGTCGGCCCTGGCCCCATCCCCATGCA	SEQ ID NO: 19

Table 4 | CRISPR arrays and spacer sequences for candidate CRISPR-associated proteins

gene_30 7407 GeneMark. hmm 16 97_aa+  14906 1 9999	Hipotetic al	uncl assifie d	GGGAA CACCCC CGCAGG CGCGGG GACCAC (SEQ ID NO: 69)	CCGACCCCTGACCAACGGGGCCGGGCAGC (SEQ ID NO: 247) GACGAGGACCGGTATCCCCGGCTGGCTGGGAGT (SEQ ID NO: 248) AACGGGTCGATCACGGATGTGGGACCCGGCC (SEQ ID NO: 249) GGGGTCCAAGGTGGGGCAAGGTGCTAATGC (SEQ ID NO: 250) TATGGCGACATGTCCTGCGTCTGGCGGCCGA (SEQ ID NO: 251) CCGCACTCGACTACCCGACCGAGTGGGCCCA (SEQ ID NO: 252) GAGGCCCTICGGGAGTGGCCCTCAGGCCAC (SEQ ID NO: 253) CAGGCCGGGGAGGGAGGGAGGGAGGGCGGGCGC (SEQ ID NO: 254) GCCGCACTCCAGGGCCGGCCGACGGCGGATG (SEQ ID NO: 255) CAGGACACCCACTCGTCCTGGGGCTTCC (SEQ ID NO: 256) CAGGCCGGGACAGGGGGCGGGCCGGGGCGCGCG (SEQ ID NO: 257) GGAGCACGCCCGATGACCAACCCGACGACCA (SEQ ID NO: 258) CCACCCCTCCACCGTGGCGCACCGGACAGGCC (SEQ ID NO: 259) GTCATCGTGCCTCTGCCCTCTGAGGGCCCTCGC (SEQ ID NO: 260) GAGGTGGTCGCCCTCCGGGCCAGCTGCC (SEQ ID NO: 261) TGGGAAGCTGATGGGGTCCCGATGCTGCCG (SEQ ID NO: 262)	SEQ ID NO: 21
gene_14 32510  GeneM ark.hm m 1564 aa+127 392 320 86	Hipotetic al	uncl assifie d	CATAAG TCCTTT GTGGAT GAGCTG TGGAGG GACGCA CTGGCA GT (SEQ ID NO: 70)	TATTCACTTTTGATGATCTGCGGAGAGATGTTCTGGGGT (SEQ ID NO: 263) TATTGTGGCAGACTGCGAATGTTTGGAGGGGGAGGGGT (SEQ ID NO: 264) CTATGTGAATGGCAACAAGTATCTGGTGAGGGACGGCAGAC (SEQ ID NO: 265) ACAACGAGGAACCTGATCGTGGAGG (SEQ ID NO: 266) AAAATGAGAAGCTTGATCGTGAAGG (SEQ ID NO: 267) ACAATGTGCCAAATAAAATAACTGACGAGTGTCTGGCAAAT (SEQ ID NO: 268) GTTTCTGTAGTAGGTCTCTATGACGAAATAATGGTTGGTGGAG (SEQ ID NO: 269) ATCTCGTAATCTAAAGCAAGACAGATCATGTGGAGTGTCTGGTAGAG (SEQ ID NO: 270) TTCTCTGTAGTGGGGCCCTTATTGTGACGAAATGTTCTGGCTAGAG	SEQ ID NO: 22

			(SEQ ID NO: 271) TGTATGGAGGAGCATGGGG (SEQ ID NO: 272)	
gene_55 70191  Genem ark.hm m 1502 _aa -  1126 5 634	Hipotet ical	uncla ssifie d	AGCTCG TGCACC GTCAGC CGATAG AGCACCC AGGTCT TCCGGC CGA  (SEQ ID NO: 273) GCGGCCCTGTCACGGGATATCCAGTTGGGGTTGGG (SEQ ID NO:274) TCGGTTATTTCGCAAGTCCGGCGGGCGGGCTCCCTGCACTGAA (SEQ ID NO: 275) AACATGCTTGAACCGTCTGGCATAGACCGCTACAGGGTCACC (SEQ ID NO: 276) ACCCTAAACCACTAGCGCACCTGGACGTCGTAGTGGATGC (SEQ ID NO: 277)	SEQ ID NO: 23
gene_24 35065  Genem ark.hm m 1265 _aa - 13 005 168 02	Hipotet ical	uncla ssifie d	TCTTTG ACCGGC AGGTCA CATCGG ACGGGG CACAAAC C  (SEQ ID NO: 71)	SEQ ID NO: 24
meta_ge ne_343 942 Gen eMark.h mm 122 0_aa -  15010  18672	Hipotet ical	uncla ssifie d	Not included	SEQ ID NO: 25

gene_14_56430 GeneMark_hmm.m1196_aa[+19_091 226_81	Hypothetical	unclassified	GATTAA CGGC GGACA AATTAA AAGAC GGCTCC GCGGAC CTCAAA GACGG GACG	GATCTTCTTCGGCGTTCAACGCTCAAGGACGGCTCT (SEQ ID NO: 279) ACGCTTGCATCTGGCGCATCACAGTAAAGGGCGGTCC (SEQ ID NO: 280)	SEQ ID NO: 26
gene_31_7827 GeneMark_hmm.m10_89_aa-[7063]1_0332	Hypothetical	unclassified	CGATAA GCATGT GAGTGA GACATC CCGAAT A	CCTTCAGCAAACAGAATCATCTAAAAGTCGC (SEQ ID NO: 281) CCTCAATTACCACTATAACCGTACAAAAATTAA (SEQ ID NO: 282) CTCCATCTCTAAACAATTATTATTATA (SEQ ID NO: 283) CCGTGGCATTACCAACTCGTACAGACTCTGAG (SEQ ID NO: 284) CGTTTCACTCGTTACAGACAATCTGTCGATTGCT (SEQ ID NO: 285) ATGGCCGTGGCTTACAAGATCTGCCGTGGC (SEQ ID NO: 286) TAAACTGGCACAAAAATGTAGTTATGTATTGTA (SEQ ID NO: 287) TACAAACGCCAACATGGACACACACATAGTG (SEQ ID NO: 288) ACCTGACCACAAATCAAGAGTTATTGAGCTTG (SEQ ID NO: 289) GGTCATGAATGGGATCGCAGTCCCTCAACCGC (SEQ ID NO: 290) TCGAATCCCCCAGCCGACACTCAGCA (SEQ ID NO: 291)	SEQ ID NO: 27
gene_44_21494 GeneMark_hmm.m1044_aa[+24_202 273_36	Hypothetical	unclassified	GTTTAG	AATTAAATACTTGTICAACCATGTCAAAACCGAACITTCGTTGCT (SEQ ID NO: 292) AGGGTAGCTTCCCTCGATAGCAAAAGTTCCGA (SEQ ID NO: 293) TTAATGTCGCTAAATGGGCTCTGGCCCTGTA (SEQ ID NO: 294)	SEQ ID NO: 28

gene_30 1145 GeneM ark.hmm m 1037 aa +19 -556 226 69	Hypothetical unclassified	AACCTA CCGTCT GGGTT GCAGCG AAC	Not included	SEQ ID NO: 29
gene_25 90511 GeneM ark.hmm m 979 -aa-  30548  33487	Hypothetical unclassified	CCGTCA AACAGC AGTTTA ATAATG CGTGGAA AA	GGAAACAAATCTTGCAAAGGCTGTGAAAGTTGG (SEQ ID NO: 295) TTCACAGGTAACATACTCCACCCACCA (SEQ ID NO: 296)	SEQ ID NO: 30
meta_gene_463 174 GeneMark.hmm 896 aa +10 631 133 21	Hypothetical unclassified	ATGGAC ATCCAA CAATAA AACCAC AAGCCA TTATA	GGGTGATAACCCCTCAAATTITGTCAGCTTGAAGAGCTGG (SEQ ID NO: 297) TGAATGCTTAAAGGCTGCCATAATGAGGTATTACATA (SEQ ID NO: 298) TATAATCTGGACATACACTTGAAGATTAGCCATGCA (SEQ ID NO: 299) TAGGTGTAAGCATTTGGCGTCCCTCACGGAAAACAGCGC (SEQ ID NO: 300) GTAGCAGTCAAATTCCTTTAGGGGTTCAAGATAAG (SEQ ID NO: 301) CCITGATGAGATTCACGTGAAAACCCCAGCCGATCTGCA (SEQ ID NO: 302) AATATAAGACATTCTCGTGTATAACGTCTTATGGCGTTATC (SEQ ID NO: 303) AGGCCTCGAATATAAAACTTCTCGTGTATAACGTCTTACG (SEQ ID NO: 304)	SEQ ID NO: 31
gene_77 3846 GeneMark.hmm 88 7_aa +3 216 587 9	Hypothetical unclassified	TCACTT GTGCTG TGTCTGG TCATGC GGCACCC GC	GAACAATAATTCACCTTCTATAGTTTCCATT (SEQ ID NO: 305) TGATTTCAGCCATTCTTGTATAAGCAAAATAGAA (SEQ ID NO: 306) AAAGAAGTACGAAAATCTGTATGAAATTAAATT (SEQ ID NO: 307) AAACTAGCAGATGTCTTGGTGTAAACTACTGAT (SEQ ID NO: 308) ATTITTCGCTGATAATAAGTGAAGTGAAGTGA (SEQ ID NO: 309) AGGTCAAAGGGATTATGAGAGGAAAAGGCAATAT	SEQ ID NO: 32

				(SEQ ID NO: 310) ATTGCTTAACATCTTACCAACGTTCTGCTCGTT (SEQ ID NO: 311) TTCATAACTAAAATTTCGGGTATTCCATCAA (SEQ ID NO: 312) GGAGATAGTAAGGAAGTTGCACAGGCATTAGAA (SEQ ID NO: 313)
gene_11 88229  GeneM ark.hm m 840_ aa+ 130_ 70 1559 2	Hypothe tical	uncla ssifie d		TGAATGCCAGCGCTGCCGGGATGCCACC (SEQ ID NO: 314) TCGATAACGCCGGTAAATACGTGTCAACTAA (SEQ ID NO: 315) GCGCTTCCCATCGCACAGCGCACGGCGCTTCC (SEQ ID NO: 316) GTGACACGCTGTGACAACCCACCTTCCAGC (SEQ ID NO: 317) CAGCACAAATAAA1CCCCCTTGACAGCCCCCTCG (SEQ ID NO: 318) TTTGGGGTATACGACGCCGACGGCGAAA (SEQ ID NO: 319) GGTGATTTATTCAAAAAAAGAGAGGGTGA (SEQ ID NO: 320) CGCGACCGGCCATCAATTITGTCTCGTTGC (SEQ ID NO: 321) GGTTCGGGGGTTCGTGGTGGAGTGCAACCGC (SEQ ID NO: 322) TTATCGGAGAGCAGCAAGAGTTGTGATGAT (SEQ ID NO: 323) ATTTCGGCTGGGCTCTGCTCTCAAGTGGAA (SEQ ID NO: 324) GCCGCTACGGCAATTAAAAGGTTTTCACCA (SEQ ID NO: 325) AGCCCCAATTTTTAGTGTGACGCAAAGCTCG (SEQ ID NO: 326) GCCTTAAACCGTTACGATCCCCGGCCGGTGTG (SEQ ID NO: 327) TTGAAAATATTGTGTGCTGC GTGTTTGTGTG (SEQ ID NO: 328)

gene_80 0233 GeneMark. hmm 83 8_aa-[ 23798] 26314		Hipothetical	unclassified	UPI000C9AE9FB	GTTTCA ATCCAC GCACTC GTGAGA GTGCGA C	CCCCATTCGCCTGAAGCACGGGCCCTACCATCTC (SEQ ID NO: 329) GGCATCAAGGCTTCGGTGGTCCCTCGGTGGAA (SEQ ID NO: 330) GAGGCTGGGGACAACCTCCGAGTTTGGGCCCA (SEQ ID NO: 331) TCTAACCTGCTGGCAATCAAAGACGCCCTGGCG (SEQ ID NO: 332) GCACGATCTGGAGAATGGGATAGCGAAAAGAA (SEQ ID NO: 333) GGGTGAAACATCCGGGATTATTCGCTTATGGACG (SEQ ID NO: 334) TGACGCCAAGGGGCCGGCGCAAGTGCACAAATTAGTG (SEQ ID NO: 335) AGAAAAGGGAAATGGTTAGCCCAGAAAGATGTT (SEQ ID NO: 336) TGTGATTTCACAAGGGCGAAGGTAGCCGGATTCC (SEQ ID NO: 337) CTGGCAAAACGGCCAGGTGGCCAGGGCGGGACG (SEQ ID NO: 338)	SEQ ID NO: 34
--	--	--------------	--------------	---------------	--------------------------------------	--	---------------

[Table 5] CRISPR arrays and spacer sequences for candidate CRISPR-associated proteins

Protein ID	other CAS protein	tracr RNA	array name	Domai n (y or n)	class type	Notes	repeats	spacer sequence (each row denotes a new spacer)	CRISPR-associated protein	Corresponding SEQ ID NO:
gene_55 43656  GeneMark.hm m 1679 aa-[ 20468] 25507		n		unclassified	7	CGGGTCCC CGGGCTG CGGGGTG GTCCC	(SEQ ID NO: 339) AGTTGCTGGAGCCCGATGAACATGCCGC (SEQ ID NO: 340) CATGACGGGGTTCGGTCCGGACGATCATGACGG (SEQ ID NO: 341) GGGTGGCCCTCGCTTCGTTGGGACCATAC (SEQ ID NO: 342) CGTGCCTGGGTCAAGCTCGCTCGGTGACCCAG (SEQ ID NO: 343) TTCATCGGGGGCGGCGATCCGGACGGAGCAT (SEQ ID NO: 344)	CCGGCCGGATCTGGAAACGGCCGGCCAGCA (SEQ ID NO: 339) AGTTGCTGGAGCCCGATGAACATGCCGC (SEQ ID NO: 340) CATGACGGGGTTCGGTCCGGACGATCATGACGG (SEQ ID NO: 341) GGGTGGCCCTCGCTTCGTTGGGACCATAC (SEQ ID NO: 342) CGTGCCTGGGTCAAGCTCGCTCGGTGACCCAG (SEQ ID NO: 343) TTCATCGGGGGCGGCGATCCGGACGGAGCAT (SEQ ID NO: 344)	SEQ ID NO: 36	

gene_39 43627  GeneM ark.hm m 1660 aa - [25075] 30057	n unclas sified	4 GTGGTCCC CGGGCTG CGGGGTG TTCCC (SEQ ID NO: 82)	CCGAGCCGACGTGGGGATGCTCCGGCAG (SEQ ID NO: 345) CCGGGTCCGTGACAAGCCAGCCGACGAGCAGG (SEQ ID NO: 346) GCGGAGCAGTGGGGCTGGGGCATGATCAT (SEQ ID NO: 347)	SEQ ID NO: 37
gene_50 85315  GeneM ark.hm m 1043 aa + 31 940 350 71	n unclas sified	4 CTCCGAGA CCATCCTCC ACTAAAAAC AAGGAATTAA AGAC (SEQ ID NO: 83)	GATTCACATTGGTCTTCCACATAAGCCTGTG (SEQ ID NO: 348) GTTCGATTGGAACTCGATAACCGGATTTCCTCTGC (SEQ ID NO: 349) CCCCCTCTATAATTACTATAAGATTGGATGGGGGAT (SEQ ID NO: 350)	SEQ ID NO: 38
gene_40 28206  GeneM ark.hm m 986 aa + 150 28 1798 8	n unclas sified	3 reverse GGTACAGA CGAACCT TGTGGGAT TGAAGC (SEQ ID NO: 84)	TAACATGAGTGACTATGGCGCTGACTTTCTGACGG (SEQ ID NO: 351) CTCGAAGGGCGGCCGATGACGACGGCAAGGGCG (SEQ ID NO: 352)	SEQ ID NO: 39
gene_19 61732  GeneM ark.hm m 838 aa -  18364 352	n unclas sified	4 GTCACCGA CCACGATC CACCAAGAA CAAGGATT GAAAC (SEQ ID NO: 85)	CTGATCGCGTAGGTGAGCAGCTTCAGGTATCCTCG (SEQ ID NO: 353) CGGAGTTCAAATGGGGGGCTTGAACCTTCCAC (SEQ ID NO: 354) CAATTCTGTTGCCAAATCCGGCGAACACTGTACCAAAC (SEQ ID NO: 355)	SEQ ID NO: 41

gene_27 55817  GeneM ark.hm ml 816_ aa + 114_ 62 1391_ 2	n 31443  GeneM ark.hm ml 802_ aa + 174_ 89 1989_ 7	n unclas sified 4	GTACGACCGGGAAATTGACAGCTGAGGCACGGCA (SEQ ID NO: 356) GTGTTCTCCCTGGGGAGAGCACCATAAGCAGTTCG (SEQ ID NO: 357) TCCAGAATTAAATGCCACGCATCAACCTACGATA (SEQ ID NO: 358)	SEQ ID NO: 42
gene_28 31443  GeneM ark.hm ml 802_ aa + 174_ 89 1989_ 7	n unclas sified 8	reverse GTGCGCTCCT TGTACCGG AGCGTGA TTGAAAC (SEQ ID NO: 87)	AATAAAAGATAATCCGCAAATCTGTGGCCCTTAAG (SEQ ID NO: 359) GGTACTGGTGGAGGGTTATACTAGGAAGCGCAA (SEQ ID NO: 360) CGTTCGGATCGATGGTAAAGACCTGAGTTGGCC (SEQ ID NO: 361) TAAGGGAGGTAACGGACTAATGCCCTTICATCGACA (SEQ ID NO: 362) TAGATCCAAAATTACACGACACGATTTCGACA (SEQ ID NO: 363) GACTGTACAAGGAATTAGGTAAIGCTTTGAAG (SEQ ID NO: 364) TATATTACCCATAATCAAGAAGCTAAAGCTGCC (SEQ ID NO: 365)	SEQ ID NO: 43
meta_ge ne_118 560 Gen eMark_h mm 195_ 8_aai+6_ 937 128_ 13	n unclas sified 4	GTTGGTCCC CGCGCGTG CGGGGTG TTCCC (SEQ ID NO: 88)	CCGAGCCGACGTCGGGGATGCTCCGGCAG (SEQ ID NO: 366) CCGGGTGTCGACAAGCCAGCCGACGAGCAGG (SEQ ID NO: 367) GCGGAGCAGTCGGGGCTCGGGGCCATGATCAT (SEQ ID NO: 368)	SEQ ID NO: 44
meta_ge ne_324 030 Gen eMark_h mm 126_ 4_aai- [24458_ 28252	n unclas sified 3	GTTTTGA ACCATTCT GTITAGCA AAGG (SEQ ID NO: 89)	GGTACCAAAAGGCAGTTATGATACTAGGCCATGGCTGAAACAA (SEQ ID NO: 369) GGTACCAAAAGGAGTAGCTATAAATTAAAGCGAAATCGATAGA (SEQ ID NO: 370)	SEQ ID NO: 45

meta_genome_295 919 GeneMark.hmm 112 9_aa+[1 8998 22 387	n unclassified	4 TTAGAAAAAAGAAAAAA (SEQ ID NO: 371) AATAAAATTCAAGAAGATTAAAGAGAAAGG (SEQ ID NO: 372) CAACAAGAATTAAAAATGCTACTAAAGATCTAGGAGAT (SEQ ID NO: 373)	SEQ ID NO: 46
meta_genome_237 613 GeneMark.hmm 908 aa-[25932 28658	n unclassified	4 GTTGTGATT TGCTTAAA AATACTAA TCTTGTGG TAGCAAACA ACAAACCT (SEQ ID NO: 91)	SEQ ID NO: 47
meta_genome_350 66 GeneMark.hmm 890 aa+[10 428 131 00	n unclassified	4 GGAACACC TGGTACAC CTGGTGG (SEQ ID NO: 92) Not included crispr software failed to recognize array and spacer	SEQ ID NO: 48
meta_genome_524 019 GeneMark.hmm 872 aa-[88341 1452	n unclassified	3 CACTTGCA GTCCCCCTAA ATCGGGG TGAGACCA TTGCAAC (SEQ ID NO: 93)	SEQ ID NO: 49

meta_genome_523	n	unclassified	3 CACTTGCA GTCCCCTA AATCGGGG TGAGACCA TTGCAAC	TTCAAGTATTGGCACATGCTGGGGAAAGAGCGTG (SEQ ID NO: 379) CGCGCTGCTTCACGGGGAGATGGCCCTCGC (SEQ ID NO: 380)	SEQ ID NO: 50
517 GeneMark.hmm 809					
aa-[142]13					
850					

### **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 modifications are within the scope of the following claims.

**WHAT IS CLAIMED IS:**

1. A method of identifying a Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR)-associated protein comprising:
  - (a) obtaining a plurality of genomic sequences, wherein a genomic sequence of the plurality of genomic sequences comprises a CRISPR-associated array;
  - (b) determining a subset of the plurality of genomic sequences comprising a plurality of coding sequences within a 20 kilobase (kb) sequence flanking region either at the 3' or 5' end of the CRISPR-associated array; and
  - (c) analyzing a coding sequence of the plurality of coding sequences and thereby identifying the CRISPR-associated protein based on the coding sequence.
2. The method of claim 1, wherein the obtaining step comprises selecting, within the plurality of genomic sequences, a genomic sequence comprising a CRISPR-associated array.
3. A method of identifying a CRISPR-associated protein comprising:
  - (a) obtaining a plurality of genomic sequences;
  - (b) selecting, within the plurality of genomic sequences, a genomic sequence comprising a CRISPR-associated array;
  - (c) determining a subset of the plurality of genomic sequences comprising a plurality of coding sequences within a 20 kilobase (kb) sequence flanking region either at the 3' or 5' end of the CRISPR-associated array; and
  - (d) analyzing a coding sequence of the plurality of coding sequences and thereby identifying the CRISPR-associated protein based on the coding sequence.
4. The method of any one of the preceding claims, wherein the plurality of genomic sequences comprise one or more of genomes, wherein the one or more of genomes are selected from: a prokaryotic genome and metagenome.
5. The method of any one of claims 2-4, wherein the selecting step comprises using an algorithm selected from the group consisting of PILER-CR, CRISPR Recognition Tool (CRT), and combinations thereof.

6. The method of any one of the preceding claims, wherein the determining step comprises using an algorithm selected from the group consisting of MetaGeneMark, Prodigal, and combinations thereof.

5

7. The method of any one of the preceding claims, wherein the analyzing step comprises filtering the coding sequence that comprises more than 500 amino acids.

8. The method of any one of the preceding claims, wherein the analyzing step comprises 10 filtering a coding sequence that comprises more than 800 amino acids.

9. The method of any one of the preceding claims, wherein the analyzing step further comprises classifying the CRISPR-associated array based on having three or more coding sequences present in the 20 kb flanking region.

15

10. The method of any one of the preceding claims, wherein the analyzing step further comprises determining a relative position of the coding sequence in the 20 kb flanking region relative to the CRISPR-associated array.

20 11. The method of any one of the preceding claims, wherein the analyzing of the coding sequence further comprises removing known CRISPR-associated proteins from the identified CRISPR-associated proteins.

25 12. The method of any one of the preceding claims, wherein the analyzing of the coding sequence comprises using an algorithm selected from the group consisting of HHMSCAN and RPS-BLAST.

13. The method of any one of the preceding claims, wherein the analyzing of the coding sequence further comprises determining the presence of a structural domain.

30

14. The method of any one of the preceding claims, wherein the analyzing of the coding sequence comprises determining the presence of a functional domain.

15. The method of claim 14, wherein the functional domain comprises a DNA binding domain, a RNA binding domain, a nuclease, a helicase, a restriction domain, or a structural maintenance of chromosomes (SMC) domain.

5

16. A computer implemented method comprising:

(a) obtaining a plurality of genomic sequences;

(b) selecting, within the plurality of genomic sequences, a genomic sequence comprising a CRISPR-associated array;

10 (c) determining a subset of the plurality of genomic sequences comprising a plurality of coding sequences within a 20 kilobase (kb) sequence flanking region either at the 3' or 5' end of the CRISPR-associated array; and

(d) analyzing a coding sequence of the plurality of coding sequences and thereby identifying a CRISPR-associated protein based on the coding sequence.

15

17. The method of claim 16, wherein the plurality of genomic sequences comprises one or more of genomes, wherein the one or more of genomes are selected from: a prokaryotic genome and metagenome.

20 18. The method of claim 16 or 17, wherein the selecting step comprises using an algorithm selected from the group consisting of PILER-CR, CRISPR Recognition Tool (CRT), and combinations thereof.

25 19. The method of any one of claims 16-18, wherein the determining step comprises using an algorithm selected from the group consisting of MetaGeneMark, Prodigal, and combinations thereof.

20. The method of any one of claims 16-19, wherein the analyzing step comprises filtering the coding sequence that comprises more than 500 amino acids.

30

21. The method of any one of claims 16-20, wherein the analyzing step comprises filtering a coding sequence that comprises more than 800 amino acids.
  22. The method of any one of claims 16-21, wherein the analyzing step further comprises classifying the CRISPR-associated array based on having three or more coding sequences present in the 20 kb flanking region.  
5
  23. The method of any one of claims 16-22, wherein the analyzing step further comprises determining a relative position of the coding sequence in the 20 kb flanking region relative to the CRISPR-associated array.  
10
  24. The method of any one of claims 16-23, wherein the analyzing of the coding sequence further comprises removing known CRISPR-associated proteins from the identified CRISPR-associated proteins.  
15
  25. The method of any one of claims 16-24, wherein the analyzing of the coding sequence comprises using an algorithm selected from the group consisting of HHMSCAN and RPS-BLAST.  
20
  26. The method of any one of claims 16-25, wherein the analyzing of the coding sequence further comprises determining the presence of a structural domain.
  27. The method of any one of claims 16-26, wherein the analyzing of the coding sequence comprises determining the presence of a functional domain.  
25
  28. The method of claim 27, wherein the functional domain comprises a DNA binding domain, a RNA binding domain, a nuclease, a helicase, a restriction domain, or a structural maintenance of chromosomes (SMC) domain.
  - 30
29. A non-naturally occurring CRISPR/Cas system comprising:

- (a) a guide RNA, wherein the guide RNA comprises a repeat sequence and a spacer sequence capable of hybridizing to a target nucleic acid; and
- (b) a CRISPR-associated protein or a nucleic acid encoding the CRISPR-associated protein, wherein the CRISPR-associated protein comprises an amino acid sequence that is at least 80% identical to a sequence selected from SEQ ID NOs: 1-50.

5 30. The system of claim 29, wherein the CRISPR-associated protein is capable of binding to the guide RNA.

10 31. The system of claim 29 or 30, wherein the CRISPR-associated protein comprises an amino acid sequence that is at least 85% identical to a sequence selected from SEQ ID NOs: 1-50.

15 32. The system of any one of claims 29-31, wherein the CRISPR-associated protein comprises an amino acid sequence that is at least 90% identical to a sequence selected from SEQ ID NOs: 1-50.

20 33. The system of any one of claims 29-32, wherein the CRISPR-associated protein comprises an amino acid sequence that is at least 95% identical to a sequence selected from SEQ ID NOs: 1-50.

34. The system of any one of claims 29-33, wherein the CRISPR-associated protein comprises an amino acid sequence selected from SEQ ID NO: 1-50.

25 35. The system of any one of claims 29-34, wherein the target nucleic acid is an RNA or DNA.

36. The system of any one of claims 29-35, wherein the targeting of the target nucleic acid results in a modification of the target nucleic acid.

30 37. The system of claim 36, wherein the modification of the target nucleic acid is a cleavage event.

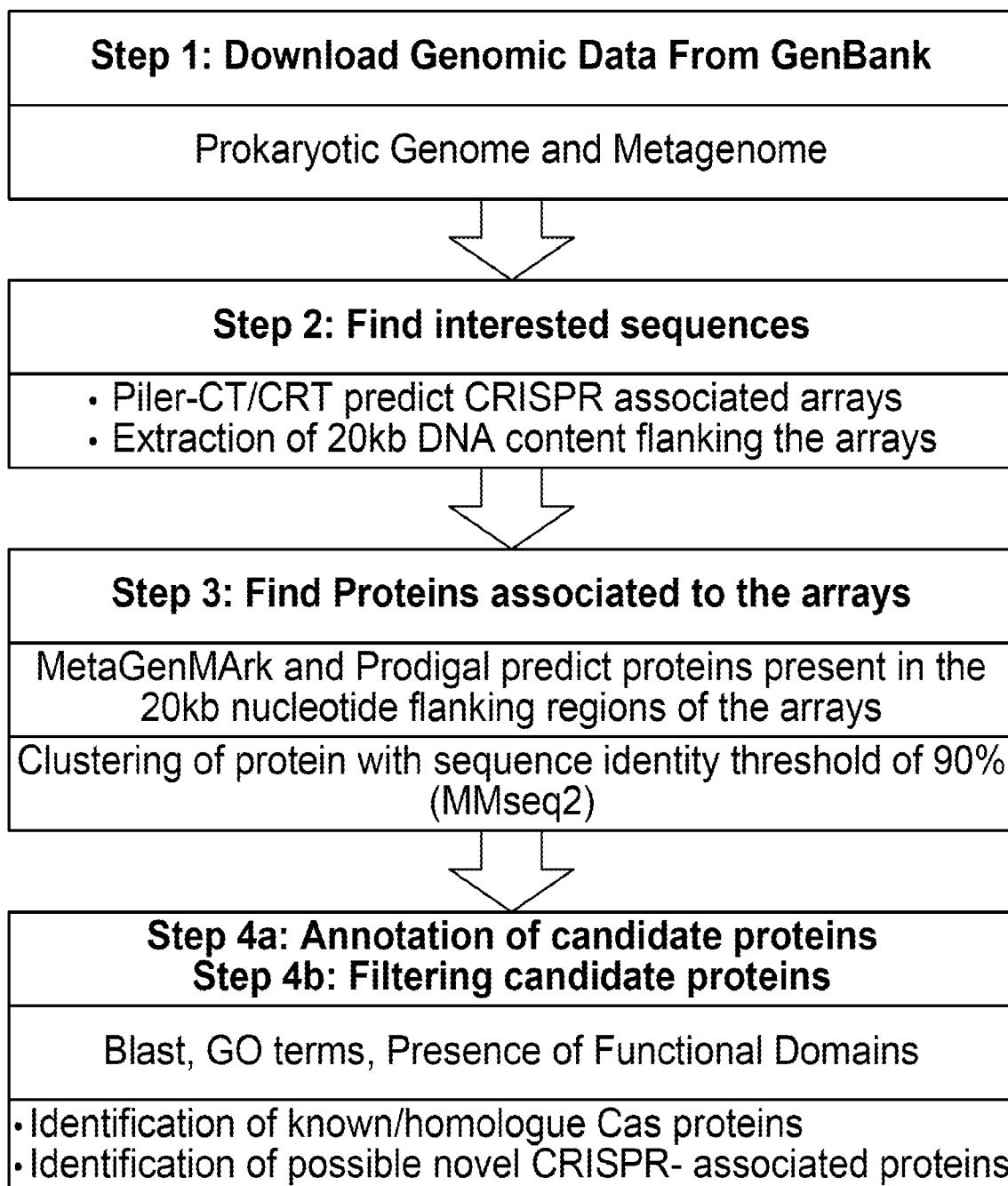
38. The system of any one of claims 29-37, wherein the guide RNA further comprises a trans-activating CRISPR RNA (tracrRNA).

5       39. The system of any one of claims 29-38, wherein the system is present in a delivery system.

10      40. The system of claim 39, wherein the delivery system comprises a delivery vehicle selected from the group consisting of an adeno-associated virus, a nanoparticle, and a liposome.

15      41. A method of treating a condition or disease in a subject in need thereof, the method comprising administering to the subject a system of any one of claims 29-40, wherein the spacer sequence is substantially complementary to a target nucleic acid associated with the condition or disease;

20      wherein the CRISPR-associated protein associates with the guide RNA to form a complex; wherein the complex binds to the target nucleic acid sequence; and wherein upon binding of the complex to the target nucleic acid sequence the CRISPR-associated protein cleaves the target nucleic acid, thereby treating the condition or disease in the subject.

**FIG. 1**

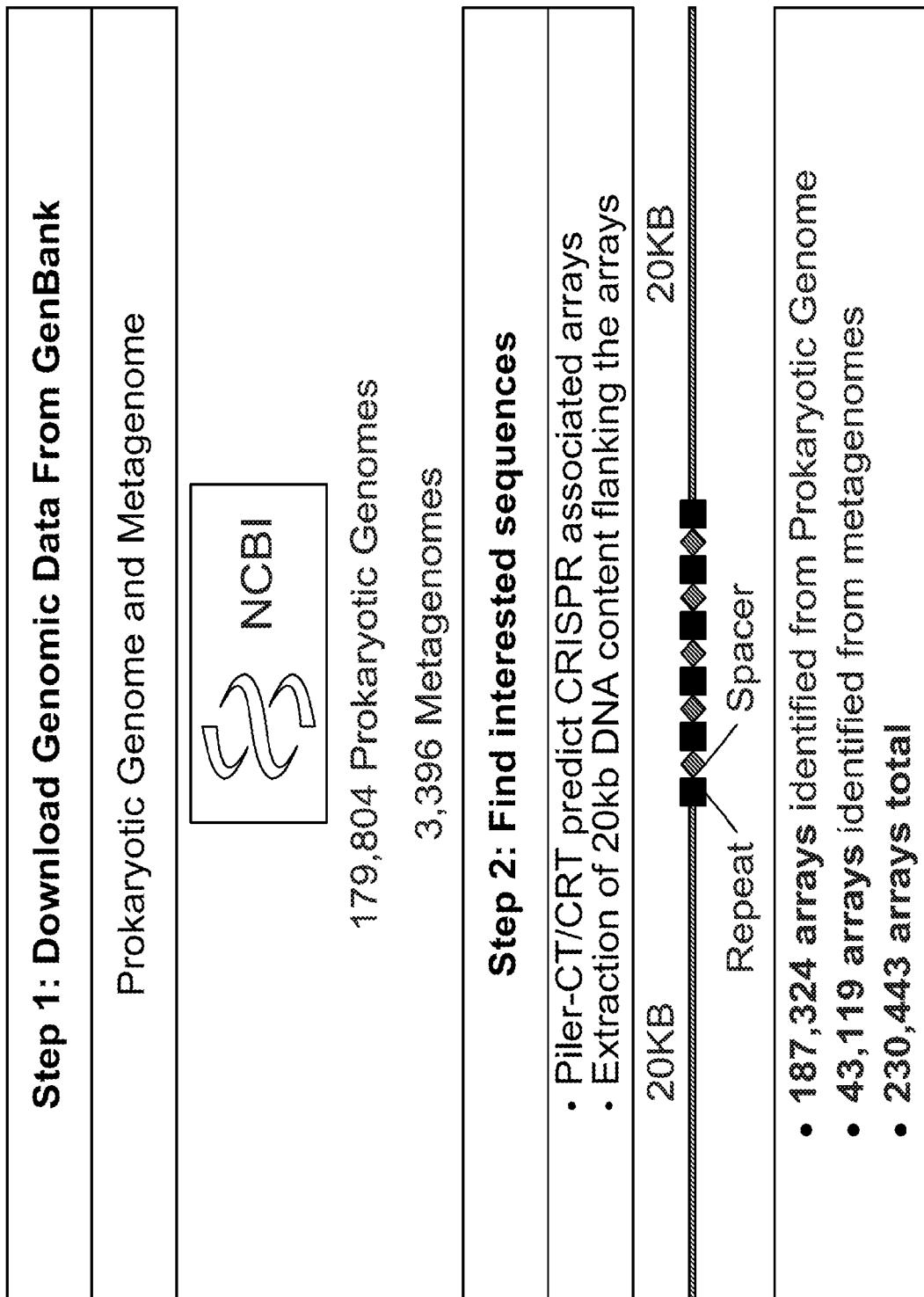
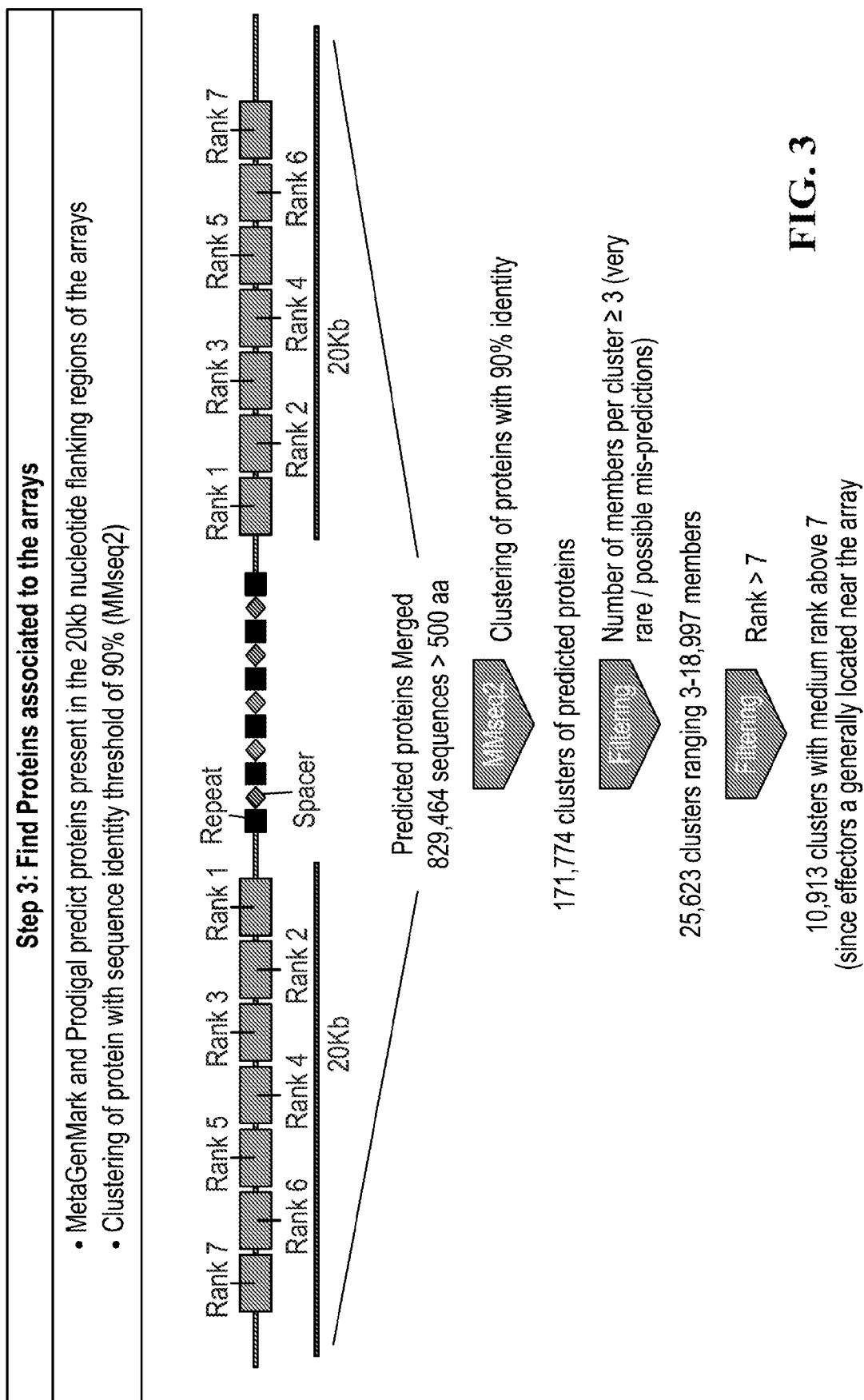
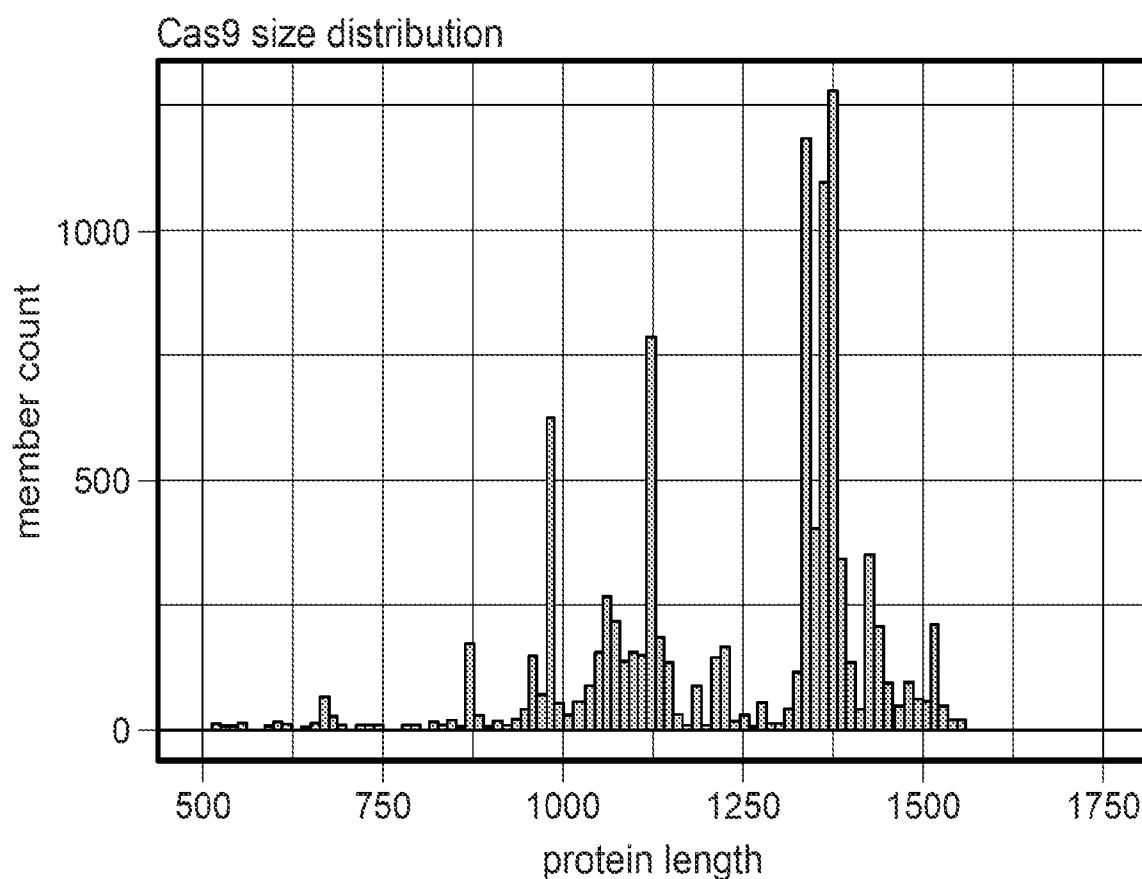
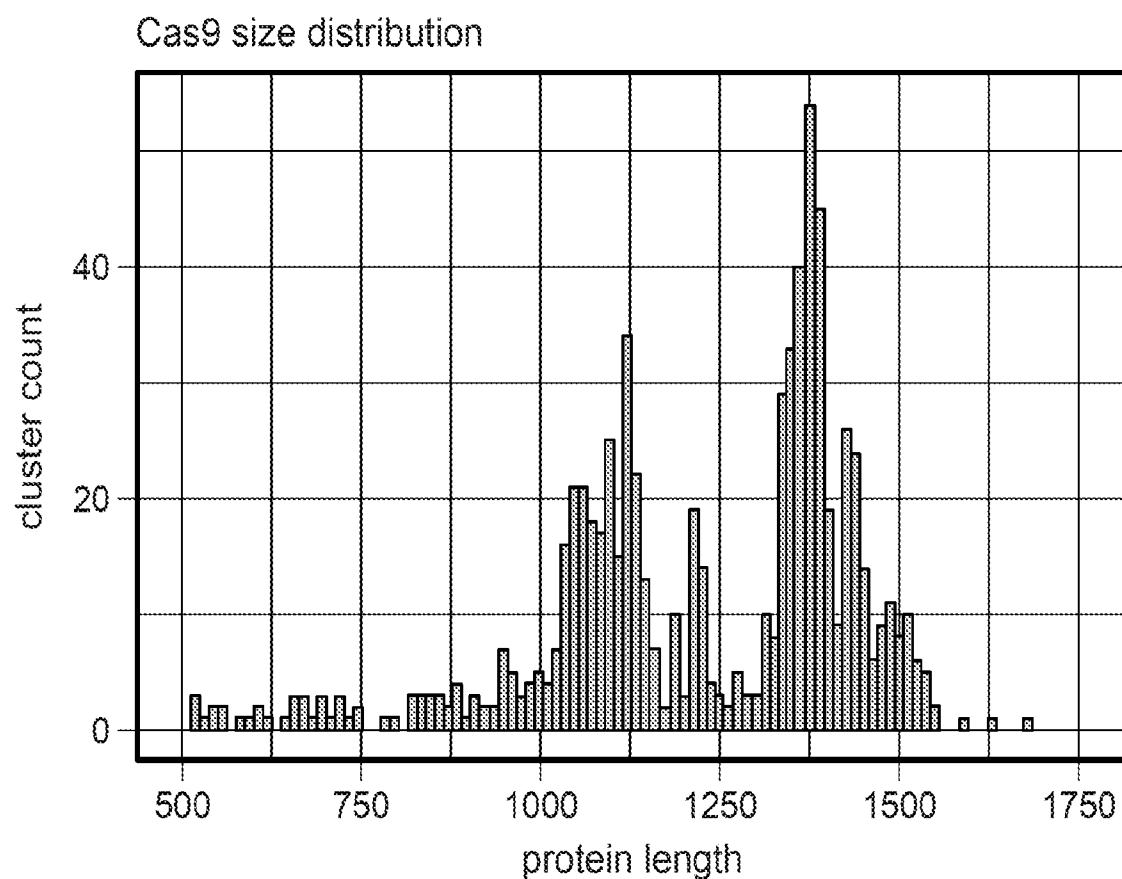


FIG. 2

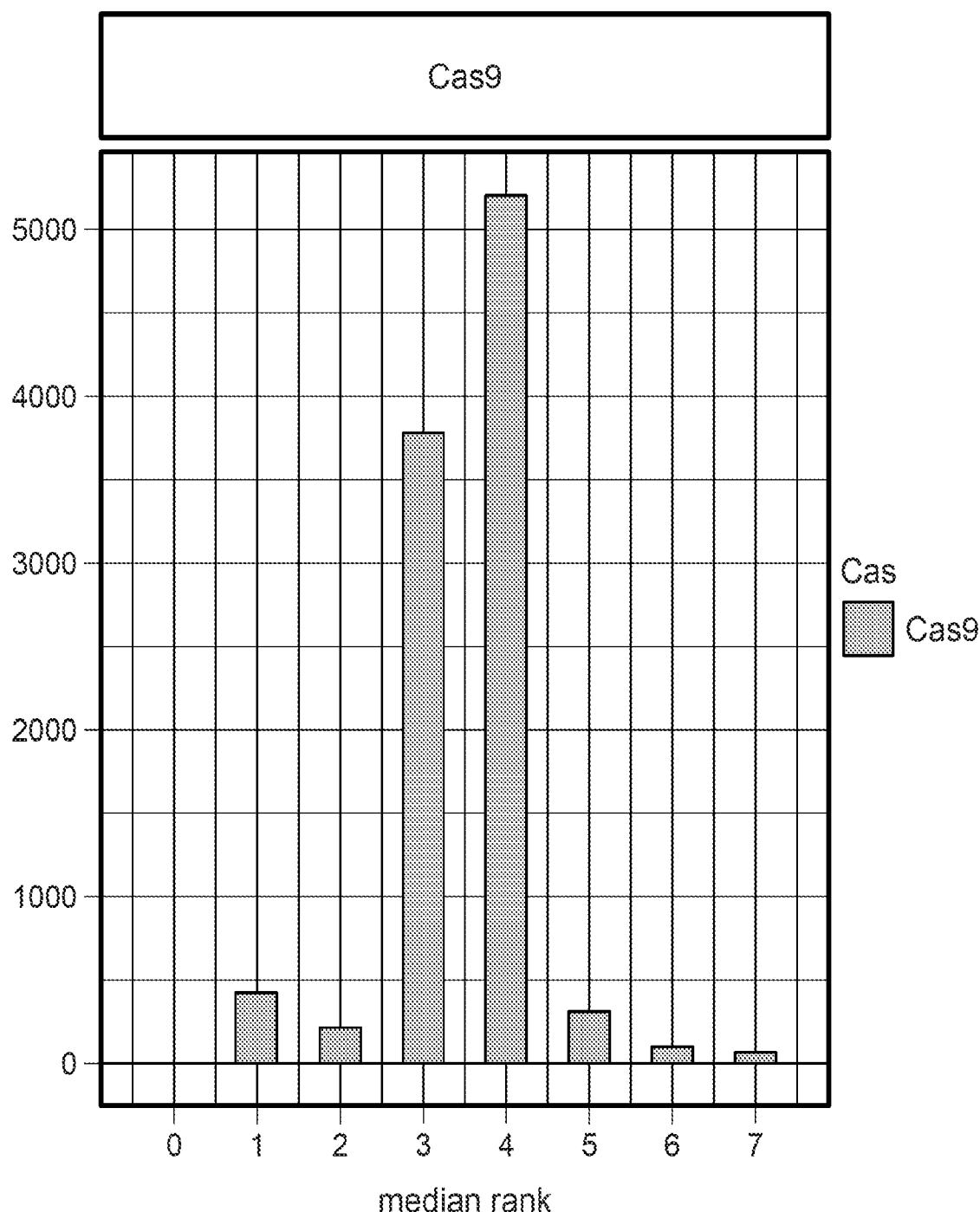
**FIG. 3**

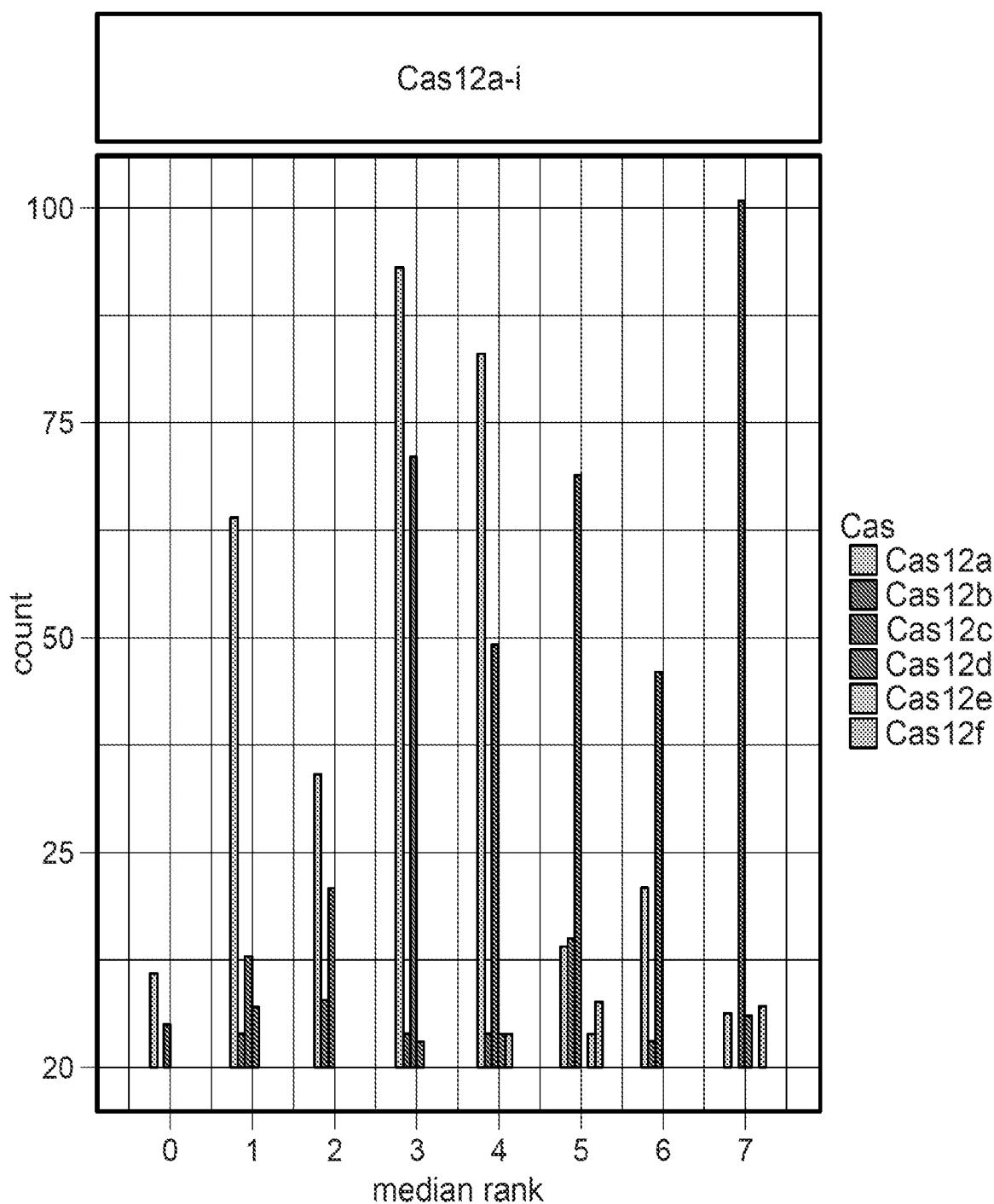


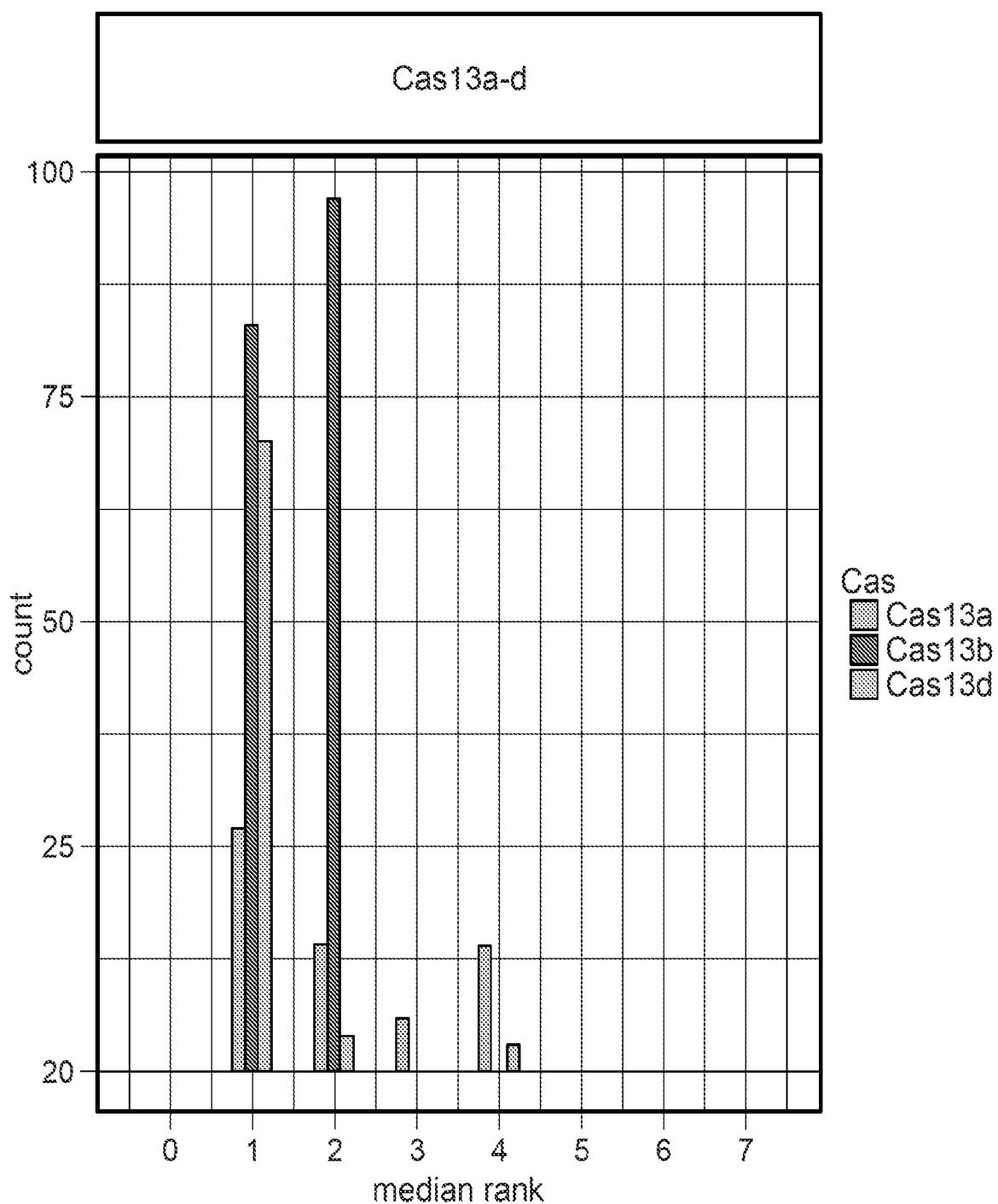
**FIG. 4A**



**FIG. 4B**

**FIG. 5A**

**FIG. 5B**

**FIG. 5C**

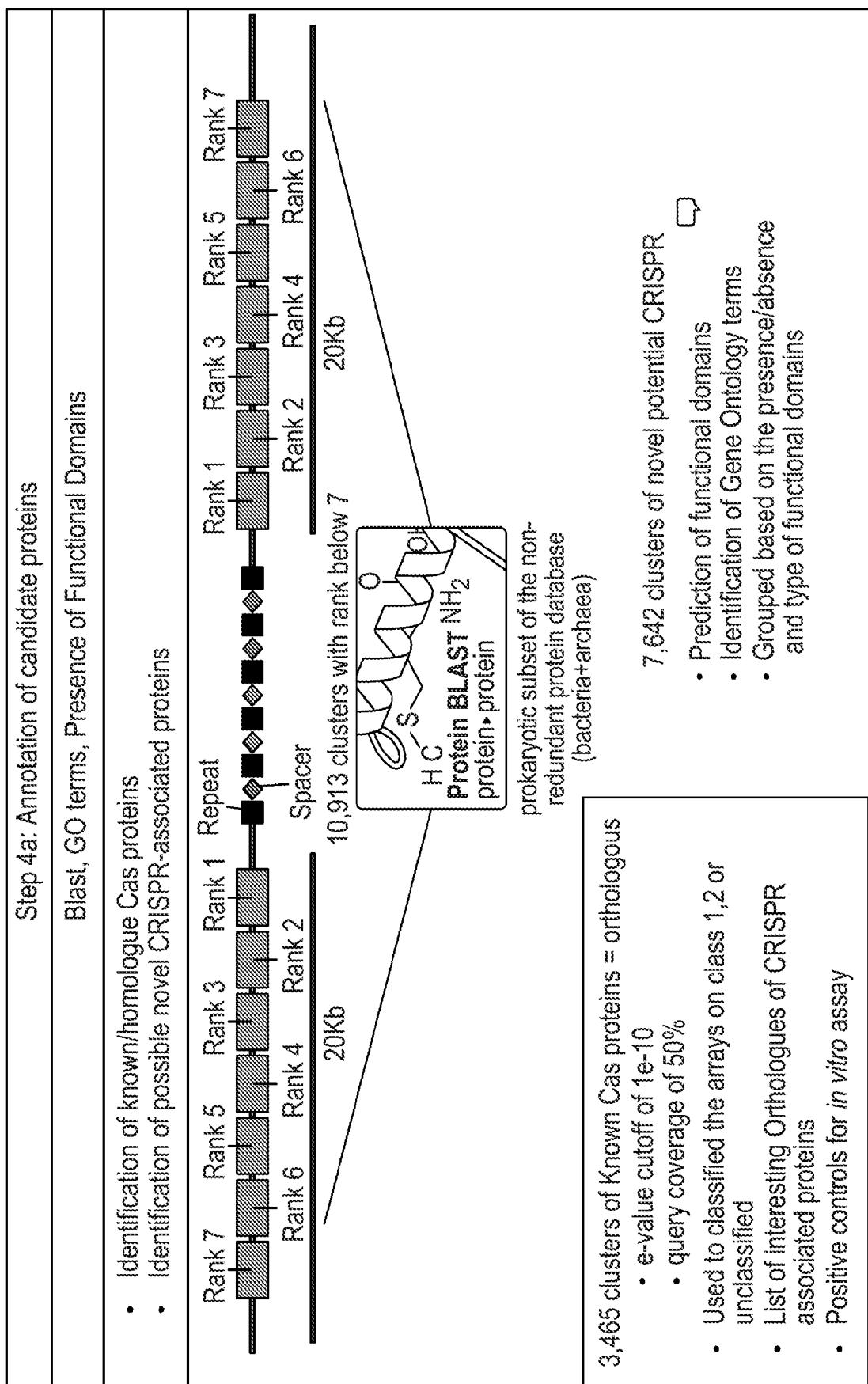


FIG. 6A

Step 4b: Filtering of candidate proteins (7,642 novel potential CRISPR)

Filtering Criteria:



1. Proteins associated to already classified arrays:

- 2417 clusters associated to defined arrays
  - 63 clusters associated with ambiguous arrays
    - Cas9/12/13
    - Cas7/5/SS etc....
  - 5162 clusters associated to unclassified arrays
    - Repeat      Spacer
- 
- 

- Proteins < 800aa and not starting with Met
- Proteins with high number of repeats and low complexity regions

2. Proteins length and complexity

- Proteins > 800aa and starting with Met
- Proteins with structural domains

3. Presence of putative or hypothetical domains

- Transmembrane domains
- Structural function like Collagens or wall proteins
- Peptidase functions
- Toxin functions
- DNA/RNA binding Domains
- Nuclease/Helicase Domains
- Restriction Domains
- SMC Domains

**FIG. 6B**

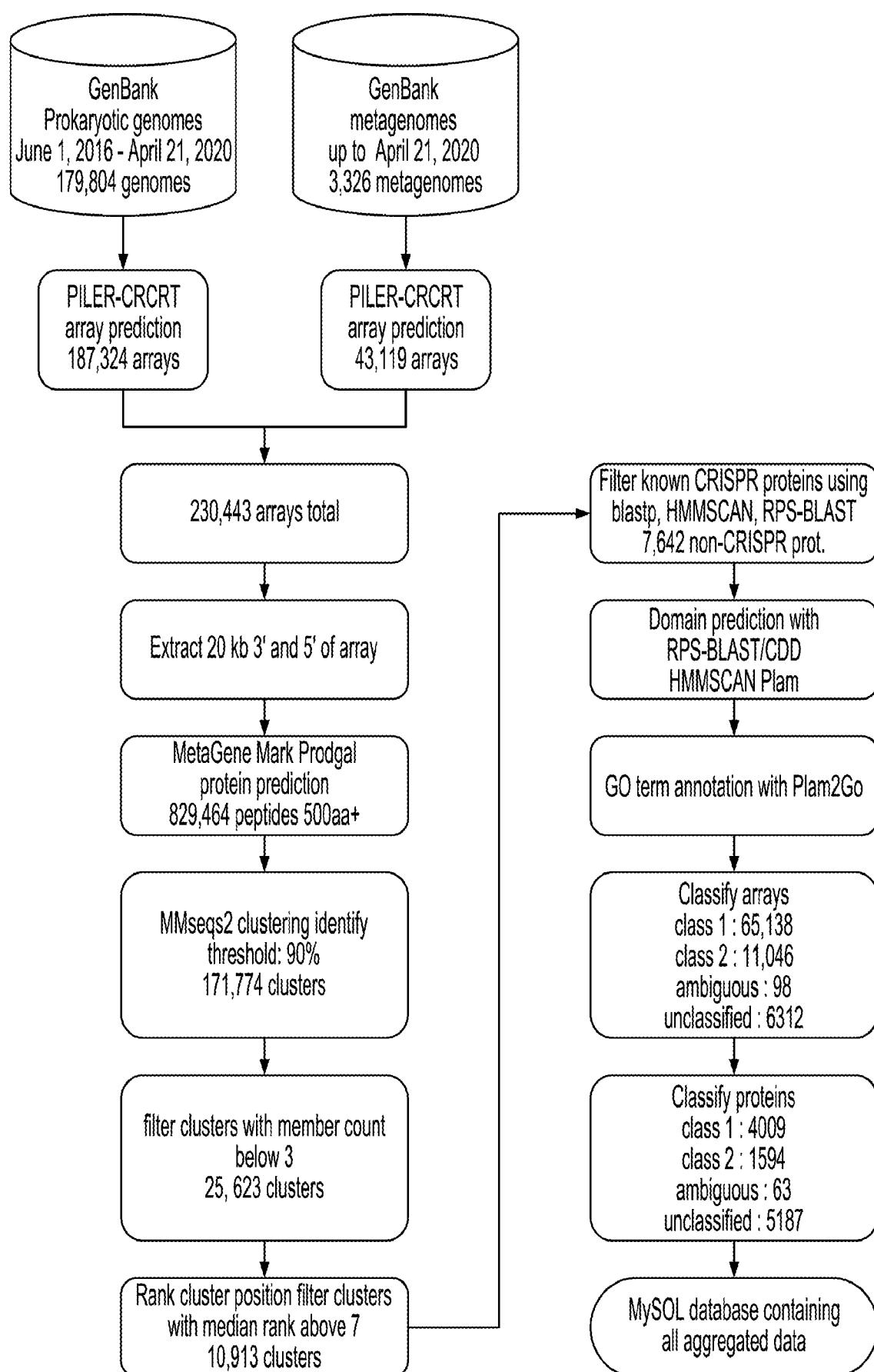
Summary

 Repeat Spacer      No presence of any known Cas

 Repeat Spacer      Cas1/2/4 /  
Potential novel CRISPR effectors

Final list of candidate proteins:

Category	Number of selected candidates	Array Class	Type of proteins
Controls	6	2	Cas9/12/13
No blast-hit	4	undefined	No functional domains identified, and no blast hit
No functional domains	15	undefined	No functional domains identified
With characterized domains	10	undefined	Include DNA/RNA binding, Nuclease/Helicase, Restriction and SMC Domains
With hypothetical domains	15	undefined	Include DNA/RNA binding, Nuclease/Helicase, Restriction and SMC Domains

**FIG. 8**

SUBSTITUTE SHEET (RULE 26)

## INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 21/60547

## A. CLASSIFICATION OF SUBJECT MATTER

IPC - C12N 9/22, C12Q 1/6869, C12Q 1/6888 (2022.01)

CPC - C12N 15/111, C12N 2310/20, C12Q 2521/301, C12N 15/1034, C12N 15/1082

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)

See Search History document

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

See Search History document

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

See Search History document

## C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	LANGE et al. CRISPRmap: an automated classification of repeat conservation in prokaryotic adaptive immune systems. Nucleic Acids Research, 1 September 2013, Vol 41, No 17, pp 8034 -8044, pg 8035, col 1, para 2, col 2, para 2, pg 8037, col 1, para 5, col 2, para 1, pg 8042, col 2, para 3-4, Lange supplemental information, S1 1 Table 1	1-4, 16-18
A	COUVIN et al. CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. Nucleic Acids Research, 2 July 2018, Vol 46, No W1, pp W246-W251, abstract, Fig. 1	1-4, 16-18

 Further documents are listed in the continuation of Box C. See patent family annex.

* Special categories of cited documents:	"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
"A" document defining the general state of the art which is not considered to be of particular relevance	"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
"D" document cited by the applicant in the international application	"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
"E" earlier application or patent but published on or after the international filing date	"&"	document member of the same patent family
"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		

Date of the actual completion of the international search

24 March 2022

Date of mailing of the international search report

APR 07 2022

Name and mailing address of the ISA/US

Mail Stop PCT, Attn: ISA/US, Commissioner for Patents  
P.O. Box 1450, Alexandria, Virginia 22313-1450

Facsimile No. 571-273-8300

Authorized officer

Kari Rodriguez

Telephone No. PCT Helpdesk: 571-272-4300

**INTERNATIONAL SEARCH REPORT**

International application No.

PCT/US 21/60547

**Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)**

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1.  Claims Nos.: because they relate to subject matter not required to be searched by this Authority, namely:
  
2.  Claims Nos.: because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:
  
3.  Claims Nos.: 5-15, 19-28 and 32-41 because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

**Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)**

This International Searching Authority found multiple inventions in this international application, as follows:

----- see extra sheet -----

1.  As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.
2.  As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.
3.  As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:
  
4.  No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.: 1-4 and 16-18

**Remark on Protest**

- The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.
- The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.
- No protest accompanied the payment of additional search fees.

**INTERNATIONAL SEARCH REPORT**

International application No.

PCT/US 21/60547

Continuation of Box No. III, Observations where unity of invention is lacking:

This application contains the following inventions or groups of inventions which are not so linked as to form a single general inventive concept under PCT Rule 13.1. In order for all inventions to be searched, the appropriate additional search fees must be paid.

**Group I:** Claims 1-4 and 16-18, directed to method of identifying a Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR)-associated protein comprising: obtaining genomic (prokaryotic) sequences comprising a CRISPR-associated array, determining, via computer implementation, a subset of said sequences within 20 kb of either side of said CRISPR array, and identifying a CRISPR-associated protein based on the coding sequences found in said genomic sequences.

**Group II+:** Claims 29-31 directed to a non-naturally occurring CRISPR/Cas system composition comprising: a guide RNA comprising a repeat sequence and a spacer sequence capable of hybridizing to a target, and a CRISPR-associated protein or a nucleic acid encoding the CRISPR-associated protein. Group II+ will be searched upon payment of additional fees. The CRISPR/Cas composition may be searched, for example, to the extent that the CRISPR-associated protein encompasses an amino acid sequence that is at least 80% identical to SEQ ID NO: 1. It is believed that claims 29-31, limited to said CRISPR/Cas composition, read on this exemplary invention. Additional CRISPR/Cas compositions will be searched upon the payment of additional fees. Applicants must specify the claims that encompass any additionally elected CRISPR/Cas compositions. Failure to clearly identify how any paid additional invention fees are to be applied to the "+" group(s) will result in only the first claimed invention to be searched. An exemplary election would be a CRISPR/Cas composition comprising a CRISPR-associated protein that is at least 80% identical to SEQ ID NO: 2 (Claims 29-31).

The inventions listed as Groups I, and II+ do not relate to a single general inventive concept under PCT Rule 13.1 because, under PCT Rule 13.2, they lack the same or corresponding special technical features for the following reasons:

**Special Technical Features**

No technical features are shared between the amino acid sequences of Group II+ and, accordingly, these groups lack unity a priori.

Additionally, even if the inventions listed as Group I and Group II+ were considered to share technical features, these shared technical features are previously disclosed by the prior art, as further discussed below.

Group I requires a method of identifying a CRISPR-associated protein, not required by group II+.

Group II+ requires an isolated CRISPR/Cas composition, not required by group I.

**Common Technical Features**

The inventions of Groups I and II+ share the technical feature of a nucleic acid encoding a CRISPR-associated protein (aka 'Cas' protein).

However, this shared technical feature does not represent a contribution over prior art, because the shared technical feature is anticipated by US 2018/0094257 A1 to the Jackson Laboratory (hereinafter 'Jackson labs'). Jackson labs teaches a nucleic acid comprising a CRISPR-associated protein (aka 'Cas' protein) (para [0009] "the invention provides a polynucleotide comprising: ... a DNA-targeting sequence that is complementary to a target polynucleotide sequence", para [0059] "the invention provides a kit comprising: (1) a subject polynucleotide, or a subject vector; (2) a subject second vector encoding the Cas9 protein").

As the technical feature was known in the art at the time of the invention, this cannot be considered a special technical feature that would otherwise unify the inventions.

Groups I and II+ therefore lack unity under PCT Rule 13 because they do not share the same or corresponding special technical feature.