**ABSTRACT**

Embodiments disclosed herein provide methods for identifying new CRISPR loci and effectors, as well as different CRISPR loci combinations found in various organisms. Class-II CRISPR systems contain single-gene effectors that have been engineered for transformative biological discovery and biomedical applications. Discovery of additional single-gene or multi-component CRISPR effectors may enhance existing CRISPR applications, such as precision genome engineering. Comprehensive characterization of CRISPR-loci may identify novel functional roles of CRISPR loci enabling new tools for biomedicine and biological discovery. CRISPR loci have enormous feature complexity, but classification of CRISPR loci has been focused on a small fraction of highly abundant features. Increased genome sequencing has enhanced the sampling of this feature complexity.

![CRISPR locus diagram](image)

**KEY TERMS**

**Direct Repeat (DR):** Short genomic repeat elements flanking spacer elements

**Spacer:** Heterogeneous sequences that specify nucleic acid targets for CRISPR effectors

**crRNA:** RNA product consisting of DR and spacer

**CRISPR effector:** Cas protein that manipulates nucleic acid sequences in complex with crRNA

---

**METHODS FOR IDENTIFYING NOVEL GENE EDITING ELEMENTS**

 apps:  THE BROAD INSTITUTE, INC., Cambridge, MA (US);
 MASSACHUSETTS INSTITUTE OF TECHNOLOGY, Cambridge, MA (US)

 inventors: Feng ZHANG, Cambridge, MA (US);
 David Arthur SCOTT, Cambridge, MA (US)

 assignees: THE BROAD INSTITUTE, INC., Cambridge, MA (US);
 MASSACHUSETTS INSTITUTE OF TECHNOLOGY, Cambridge, MA (US)

 appl. no.: 15/679,619

 filed: Aug. 17, 2017

 related U.S. application data

 provisional application no. 62/376,383, filed on Aug. 17, 2016.
CRISPR locus
- Set of protein and nucleic acid components comprising RNA-guided effector complexes
- (and) machinery for CRISPR array adaptation

KEY TERMS
Direct Repeat (DR): Short genomic repeat elements flanking spacer elements
Spacer: Heterogeneous sequences that specify nucleic acid targets for CRISPR effectors
crRNA: RNA product consisting of DR and spacer
CRISPR effector: Cas protein that manipulates nucleic acid sequences in complex with crRNA

FIG. 1
300

CRISPR objects for all genomes sampled

305

Individual CRISPR objects

Group objects by organism

310

Calculate euclidean distance between CRISPR objects/groups of CRISPR objects

315

Distance matrix for all CRISPR objects/groups of CRISPR objects

320

Hierarchical clustering

Classification of individual CRISPR objects

Classification of CRISPR-containing organisms

FIG. 3
FIG. 4
CRISPR objects for all genomes sampled

Individual CRISPR objects

Group objects by organism

Deep learning

Classification of individual CRISPR objects

Classification of CRISPR-containing organisms

FIG. 8
METHODS FOR IDENTIFYING NOVEL GENE EDITING ELEMENTS

CROSS REFERENCE TO RELATED APPLICATIONS

[0001] This application claims the benefit of U.S. Provisional Application No. 62/376,383, filed Aug. 17, 2016, which is incorporated herein by reference in its entirety.

STATEMENT AS TO FEDERALLY SPONSORED RESEARCH

[0002] This invention was made with government support under grant numbers MH100706 and MH110049 awarded by the National Institutes of Health. The government has certain rights in the invention.

TECHNICAL FIELD

[0003] The embodiments disclosed herein are directed to methods for identifying novel Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) elements and/or loci.

BACKGROUND

[0004] The subject matter disclosed herein is generally directed to CRISPR, which is currently best understood as a programmable bacterial immune system that acquires phage genomic sequences upon infection and then uses these sequences to target the phage for destruction upon subsequent infection. Bacteriophages are viral elements that infect bacteria and hijack bacterial machinery to replicate, quickly leading to bacterial lysis and release of replicated phage virions to propagate the infection. Hence, bacterial populations that do not possess immunity to a specific phage species are rapidly depleted upon infection, providing a strong positive selective pressure for the acquisition of phage immunity in bacteria. Possessing the ability to defend against phage infection and adapt to infection by novel phages, CRISPR immune systems are widespread and highly conserved in bacteria. Initial investigations into the diversity of these systems has uncovered multiple programmable nucleases that are capable of mitigating phage infection in bacteria and have enabled engineered genome editing in mammalian cells.

[0005] CRISPR systems are located in bacterial genomes by searching for a set of regularly interspaced short repetitive genomic elements, termed the CRISPR array. The short genomic repeat elements of the CRISPR array are termed direct repeats and flank intervening heterogeneous sequences termed spacers, which specify the nucleic acid targets to be manipulated by the CRISPR machinery. The genetic locus surrounding the array (CRISPR locus) contains a set of CRISPR-associated (Cas) genes typically located within 10 kb of the array. Minimally, Cas genes contained in the CRISPR locus encode one or more CRISPR effector proteins responsible for phage interference, such as the well known Cas9-nuclease. Cumulatively, the array is transcribed and processed into CRISPR RNAs (crRNAs) composed of individual spacers and flanking direct repeat sequence, which complex with CRISPR effector proteins and guide them to nucleic acid targets for phage interference. CRISPR loci may also encode a set of Cas proteins responsible for CRISPR array adaptation in response to novel phage challenge.

[0006] A minimal ensemble of two proteins, Cas1 and Cas2, facilitate adaptation by acquiring spacer sequences from novel phage genomes and inserting them into the CRISPR array flanked by direct repeats. In contrast to the identification of diverse CRISPR effectors capable of mediating phage interference, Cas1 and Cas2 remain the only bacterial proteins known to drive CRISPR array adaptation. Hence, classification of CRISPR loci has been primarily guided by the diversity of constituent CRISPR effector genes in an individual locus. Currently the diversity of Cas proteins in CRISPR loci has been divided into type I-V, which are further subdivided into class 1 if interference is mediated by greater than 1 effector, or class 2 if a single effector protein is responsible for interference. Class 1 CRISPR loci (types: I, III, and IV), are observed much more frequently in bacterial and archael genomes than class 2 loci (types: II, V).

SUMMARY OF THE INVENTION

[0007] In one aspect, the invention provides a method for identifying novel CRISPR effector elements, comprising: classifying, by a processor, a CRISPR locus using an unsupervised machine learning applied to all or a subset of CRISPR locus elements as an initial set of inputs; identifying, using the processor, putative novel effector elements in the CRISPR locus, and optionally screening each identified putative novel effector element for one or more biological functions. In one embodiment, the unsupervised machine learning comprises hierarchical clustering. In another embodiment, hierarchical clustering comprises: generating, by the processor and prior to the classifying step, a preliminary set of CRISPR locus classes by separating a set of known CRISPR loci based, at least in part, on a sequence similarity and/or domain similarity between one or more protein elements of the CRISPR loci; generating, by the processor, a distance matrix data structure based, at least in part, on cumulative similarities between constituent proteins in each CRISPR locus class; and wherein the putative CRISPR loci are classified by applying the hierarchical clustering model to the distance matrix. In another embodiment, the CRISPR loci are separated based on a domain similarity. In another embodiment, the domain similarity is determined using a hidden markov model. In another embodiment, the cumulative similarities between constituent proteins is determined, at least in part, by calculating a Euclidean distance between each CRISPR locus. In another embodiment, the unsupervised clustering model is an unsupervised neural network model. In another embodiment, the unsupervised neural network model is trained using a set of known CRISPR loci that include spacer and repeat elements of the known CRISPR loci.

[0008] In another aspect, the invention provides an isolated CRISPR effector, or a polynucleic acid sequence encoding said effector, identified by the method according to any one of the preceding claims. In one embodiment, the invention provides for use of such a CRISPR effector for genome or transcriptome editing or modification in eukaryotic cells.

[0009] These and other aspects, objects, features, and advantages of the example embodiments will become apparent to those having ordinary skill in the art upon consideration of the following detailed description of illustrated example embodiments.
BRIEF DESCRIPTION OF THE DRAWINGS

[0010] FIG. 1 is a schematic of an example anatomy of a CRISPR locus.

[0011] FIG. 2 is a schematic showing an example CRISPR classification scheme.

[0012] FIG. 3 is a block diagram showing a hierarchical clustering approach in accordance with certain example embodiments.

[0013] FIG. 4 is a schematic showing calculation of a distance matrix in accordance with certain example embodiments.

[0014] FIG. 5 shows example results showing clustering around known Cas genes.

[0015] FIG. 6 shows another set of example results showing conserved bacterial loci that cluster around minimal CRISPR arrays.

[0016] FIG. 7 shows another set of example results showing clustering around a novel integrase element.

[0017] FIG. 8 is a schematic showing an alternative unsupervised machine learning approach in accordance with certain example embodiments.

[0018] FIG. 9 is a block diagram depicting a computing machine and a module, in accordance with certain example embodiments.

DETAILED DESCRIPTION OF THE EXAMPLE EMBODIMENTS

General Definitions


[0020] As used herein, the singular forms “a”, “an”, and “the” include both singular and plural referents unless the context clearly dictates otherwise.

[0021] The term “optional” or “optionally” means that the subsequent described event, circumstance or substituent may or may not occur, and that the description includes instances where the event or circumstance occurs and instances where it does not.

[0022] The recitation of numerical ranges by endpoints includes all numbers and fractions subsumed within the respective ranges, as well as the recited endpoints.

[0023] The terms “about” or “approximately” as used herein when referring to a measurable value such as a parameter, an amount, a temporal duration, and the like, are meant to encompass variations of and from the specified value, such as variations of +/-10% or less, +/-15% or less, +/-1% or less, and +/-0.1% or less of and from the specified value, insofar such variations are appropriate to perform in the disclosed invention. It is to be understood that the value to which the modifier “about” or “approximately” refers is itself also specifically, and preferably, disclosed.

[0024] Reference throughout this specification to “one embodiment,” “an embodiment,” “an example embodiment,” means that a particular feature, structure or characteristic described in connection with the embodiment is included in at least one embodiment of the present invention. Thus, appearances of the phrases “in one embodiment,” “in an embodiment,” or “an example embodiment” in various places throughout this specification are not necessarily all referring to the same embodiment, but may. Furthermore, the particular features, structures or characteristics may be combined in any suitable manner, as would be apparent to a person skilled in the art from this disclosure, in one or more embodiments. Furthermore, while some embodiments described herein include some but not other features included in other embodiments, combinations of features of different embodiments are meant to be within the scope of the invention. For example, in the appended claims, any of the claimed embodiments can be used in any combination.

[0025] All publications, published patent documents, and patent applications cited herein are hereby incorporated by reference to the same extent as though each individual publication, published patent document, or patent application was specifically and individually indicated as being incorporated by reference.

Overview

[0026] Embodiments disclosed herein provide methods for identifying new CRISPR loci and effectors, as well as different CRISPR loci combinations found in various organisms. Class-II CRISPR systems contain single-gene effectors that have been engineered for transformative biological discovery and biomedical applications. Discovery of additional single-gene or multi-component CRISPR effectors may enhance existing CRISPR applications, such as precision genome engineering. Comprehensive characterization of CRISPR-loci may identify novel functional roles of CRISPR loci enabling new tools for biomedicine and biological discovery. CRISPR loci have enormous feature complexity, but classification of CRISPR loci has been focused on a small fraction of highly abundant features. Increased genome sequencing has enhanced the sampling of this feature complexity. Presently, approximately 1,000 to 2,000 prokaryotic genomes are added per week, and on track to reach approximately one million genomes per week by 2019-2020 (Land et al., Funct Integr Genomics, 2015; NCBI). Novel effectors and functional roles of CRISPR may have previously eluded discovery because of bias in hypothesis-driven CRISPR classification. The present method uses
an unsupervised learning method to increase feature sampling and to enable comprehensive classification of CRISPR loci.

0027] FIG. 1 provides a general anatomy of a CRISPR locus. FIG. 2 provides an overview of current CRISPR classification.

Example Processes

0028] FIG. 3 is a block flow diagram depicting a method 300 to identify novel CRISPR elements. Method 300 begins at block 305 where CRISPR objects for all genomes are sampled to determine an initial degree of similarity. In certain example embodiments, a local alignment algorithm may be used to determine an initial degree of similarity. An example local alignment algorithm is Blast+. Objects can be clustered based on this initial similarity score. For example, the alignment algorithm may be iterated over the population of CRISPR objects to break into a set of initial classes. Alternatively, the initial degree of similarity may be determined based on domain prediction. For example, hidden markov models trained on known domains may be used to capture relevant features and predict whether a given protein contains a given domain. The initial clustering can be done on the basis of similarity between protein domains. A CRISPR object may be defined as any one of the cas genes, leaders, repeats, spacers, or combinations thereof. Example objects that may serve as a basis for the initial clustering analysis are shown in FIGS. 1-2. As shown in FIG. 3, grouping based on similarity may be done based on individual CRISPR objects or objects may be grouped by organism. In certain example embodiments, both groupings may be performed, including in parallel. The method then proceeds to block 310.

0029] At block 310, the Euclidean distance between CRISPR objects/groups is determined. The CRISPR objects provided as input are compared to other known CRISPR loci. Turning to FIG. 4, an example distance matrix calculation is provided. Two loci are represented, each comprised of a set of CRISPR objects. The “CDS” refers to an object cluster generated as described above. So, for example, locus 1 has a protein from cluster “14,” cluster “10,” and so on. An example distance matrix is represented at the bottom of FIG. 4. Each column in the matrix represents an object with the value “0” indicating the absence of a particular object in from a given cluster and the value of “1” representing the presence of an object from a given cluster. As can be seen in FIG. 4, both Locus 1 and Locus 2 comprise CDS 14, CDS 10, CDS 1, CDS 4, etc. This is captured in the data matrix where values of “1” in the same column represent the presence of that object in the loci. However, each locus further comprises different objects not found in the other loci, such as CDS 253, CDS 941, and CDS 153 in loci 1, and CDS 189, CDS 2011, CDS 21, etc. in loci 2. Thus, columns related to those objects will show a “1” in the loci containing that object and a “0” in the loci not containing the object. The Euclidean distance is then determined based at least in part on the derived distance matrix. The method then proceeds to block 315.

0030] At block 315, hierarchical clustering is then performed based, at least in part, on the calculated Euclidean distance to generate a final clustering used to classify the individual CRISPR objects or group of CRISPR objects. In certain example embodiments, the method may generate a dendrogram showing the distance between different objects or groups of objects.

0031] The results of the hierarchical clustering may then be used to identify clustering around features of interest. FIG. 5 shows example results showing clustering around known Cas genes. FIG. 6 show another set of example results showing conserved bacterial loci that cluster around minimal CRISPR arrays. FIG. 7 shows another set of example results showing clustering around a novel integrase element. In this way, new CRISPR elements may be identified, including loci that may contained non-canonical objects and associated functions not typically found in known loci.

0032] In certain example embodiments, identified CRISPR elements are further assessed by assaying for one or more biological functions. For example, assays may be run on the integrase identified in FIG. 6 to verify the predicted integrase function.

0033] In certain other example embodiments, the method may comprise the use of a neural network. To determine underlying structural features of CRISPR loci the data sets of known CRISPR objects, groups of CRISPR objects, or entire loci may be used. Alternatively, the clustering of loci identified using the above hierarchical clustering method may be used to provide an initial input set for the neural network. See FIG. 8. Additional sets of CRISPR objects and/or groups of objects may then be applied as inputs. The neural network may be used to classify said object or group of objects based on the similarity of CRISPR objects the neural network has previously processed as inputs. In certain example embodiments, the neural network is a self organizing map (SOM). When a pattern is presented to a SOM, a neuron with closest weight vector is considered a winner and its weights are adapted to the pattern, as well as the weights of the nearest neighbor nodes. In other example embodiments, the neural network may be based on an adaptive resonance theory. The neural network is composed of two layers, a comparison field and recognition field. The recognition field determines the best matching neuron based on the vector transferred from the comparison field.

0034] Certain embodiments described herein may be performed on a programmable computer programmed to execute the steps described in detail above.

CRISPR-Cas Systems

0035] CRISPR-Cas system activity, such as CRISPR-Cas system design may involve target disruption, such as target mutation, such as leading to gene knockout. CRISPR-Cas system activity, such as CRISPR-Cas system design may involve replacement of particular target sites, such as leading to target correction. CRISPR-Cas system design may involve removal of particular target sites, such as leading to target deletion. CRISPR-Cas system activity may involve modulation of target site functionality, such as target site activity or accessibility, leading for instance to (transcriptional and/or epigenetic) gene or genomic region activation or gene or genomic region silencing. The skilled person will understand that modulation of target site functionality may involve CRISPR effector mutation (such as for instance generation of a catalytically inactive CRISPR effector) and/or functionalization (such as for instance fusion of
the CRISPR effector with a heterologous functional domain, such as a transcriptional activator or repressor), as described herein elsewhere.

[0036] The present invention may be further illustrated and extended based on aspects of CRISPR-Cas development and use as set forth in the following articles and particularly as relates to delivery of a CRISPR protein complex and uses of an RNA guided endonuclease in cells and organisms:


Cpf1 is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System, Zetsche et al., Cell 163, 759-71 (Sep. 25, 2015).


each of which is incorporated herein by reference, may be considered in the practice of the instant invention, and discussed briefly below:

Cong et al. engineered type II CRISPR-Cas systems for use in eukaryotic cells based on both Streptococcus thermophilus Cas9 and also Streptococcus pyogenes Cas9 and demonstrated that Cas9 nucleases can be directed by short RNAs to induce precise cleavage of DNA in human and mouse cells. Their study further showed that Cas9 as converted into a nicking enzyme can be used to facilitate homology-directed repair in eukaryotic cells with minimal mutagenic activity. Additionally, their study demonstrated that multiple guide sequences can be encoded into a single CRISPR array to enable simultaneous editing of several at endogenous genomic loci sites within the mammalian genome, demonstrating easy programmability and wide applicability of the RNA-guided nuclease technology. This ability to use RNA to program sequence specific DNA cleavage in cells defined a new class of genome engineering tools. These studies further showed that other CRISPR loci are likely to be transplantable into mammalian cells and can also mediate mammalian genome cleavage. Importantly, it can be envisaged that several aspects of the CRISPR-Cas system can be further improved to increase its efficiency and versatility.

Jiang et al. used the clustered, regularly interspaced, short palindromic repeats (CRISPR)-associated Cas9 endonuclease complexed with dual-RNAs to introduce precise mutations in the genomes of Streptococcus pneumoniae and Escherichia coli. The approach relied on dual-RNA:Cas9-directed cleavage at the targeted genomic site to kill unmutated cells and circumvent the need for selectable markers or counterselections. The study reported reprogramming dual-RNA:Cas9 specificity by changing the sequence of short CRISPR RNA (crRNA) to make single- and multinucleotide changes carried on editing templates. The study showed that simultaneous use of two crRNAs enabled multiplex mutagenesis. Furthermore, when the approach was used in combination with recombineering, in S. pneumoniae, nearly 100% of cells that were recovered using the described approach contained the desired mutation, and in E. coli, 65% that were recovered contained the mutation.

Wang et al. (2013) used the CRISPR-Cas system for the one-step generation of mice carrying mutations in multiple genes which were traditionally generated in multiple steps by sequential recombination in embryonic stem cells and/or time-consuming intercrossing of mice with a single mutation. The CRISPR-Cas system will greatly accelerate the in vivo study of functionally redundant genes and of epistatic gene interactions.

Konermann et al. (2013) addressed the need in the art for versatile and robust technologies that enable optical and chemical modulation of DNA-binding domains based CRISPR Cas9 enzyme and also Transcriptional Activator Like Effectors

Ran et al. (2013-A) described an approach that combined a Cas9 nicks active mutant with paired guide RNAs to introduce targeted double-strand breaks. This addresses the issue of the Cas9 nuclease from the microbial CRISPR-Cas system being targeted to specific genomic loci by a guide sequence, which can tolerate certain mismatches to the DNA target and thereby promote undesired off-target mutagenesis. Because individual nicks in the genome are repaired with high fidelity, simultaneous nicking via appropriately offset guide RNAs is required for double-stranded breaks and extends the number of specifically recognized bases for target cleavage. The authors demonstrated that using paired nicking can reduce off-target activity by 50- to 1,500-fold in cell lines and to facilitate gene knockout in mouse zygotes without sacrificing on-target cleavage efficiency. This versatile strategy enables a wide variety of genome engineering applications that require high specificity.

Hsu et al. (2013) characterized SpCas9 targeting specificity in human cells to inform the selection of target sites and avoid off-target effects. The study evaluated >700 guide RNA variants and SpCas9-induced indel mutation levels at >100 predicted genomic off-target loci in 293T and 293FT cells. The authors that SpCas9 tolerates mismatches between guide RNA and target DNA at different positions in a sequence-dependent manner, sensitive to the number, position and distribution of mismatches. The authors further showed that SpCas9-mediated cleavage is unaffected by DNA methylation and that the dosage of SpCas9 and gRNA can be titrated to minimize off-target modification. Additionally, to facilitate mammalian genome engineering applications, the authors reported providing a web-based software tool to guide the selection and validation of target sequences as well as off-target analyses.

Ran et al. (2013-B) described a set of tools for Cas9-mediated genome editing via non-homologous end joining (NHEJ) or homology-directed repair (HDR) in mammalian cells, as well as generation of modified cell lines for downstream functional studies. To minimize off-target cleavage, the authors further described a double-nicking strategy using the Cas9 nickase mutant with paired guide RNAs. The protocol provided by the authors experimentally derived guidelines for the selection of target sites, evaluation of cleavage efficiency and analysis of off-target activity. The studies showed that beginning with target design, gene modifications can be achieved within as little as 1-2 weeks, and modified clonal cell lines can be derived within 2-3 weeks.

Shalem et al. described a new way to interrogate gene function on a genome-wide scale. Their studies showed that delivery of a genome-scale CRISPR-Cas9
knockout (GeCKO) library targeted 18,080 genes with 64,751 unique guide sequences enabled both negative and positive selection screening in human cells. First, the authors showed use of the GeCKO library to identify genes essential for cell viability in cancer and pluripotent stem cells. Next, in a melanoma model, the authors screened for genes whose loss is involved in resistance to vemurafenib, a therapeutic that inhibits mutant protein kinase BRAF. Their studies showed that the highest-ranking candidates included previously validated genes NF1 and MED12 as well as novel hits NF2, CUL3, TADA2B, and TADA1. The authors observed a high level of consistency between independent guide RNAs targeting the same gene and a high rate of hit confirmation, and thus demonstrated the promise of genome-scale screening with Cas9.

[0074] Nishimasu et al. reported the crystal structure of Streptococcus pyogenes Cas9 in complex with sgRNA and its target DNA at 2.5 Å resolution. The structure revealed a bilobed architecture composed of target recognition and nuclease lobes, accommodating the sgRNA:DNA heteroduplex in a positively charged groove at their interface. Whereas the recognition lobe is essential for binding sgRNA and DNA, the nuclease lobe contains the HNH and RuvC nuclease domains, which are properly positioned for cleavage of the complementary and non-complementary strands of the target DNA, respectively. The nuclease lobe also contains a carboxyl-terminal domain responsible for the interaction with the protospacer adjacent motif (PAM). This high-resolution structure and accompanying functional analyses have revealed the molecular mechanism of RNA-guided DNA targeting by Cas9, thus paving the way for the rational design of new, versatile genome-editing technologies.

[0075] Wu et al. mapped genome-wide binding sites of a catalytically inactive Cas9 (dCas9) from Streptococcus pyogenes loaded with single guide RNAs (sgRNAs) in mouse embryonic stem cells (mESCs). The authors showed that each of the four sgRNAs tested targets dCas9 to between tens and thousands of genomic sites, frequently characterized by a 5-nucleotide seed region in the sgRNA and an NGG protospacer adjacent motif (PAM). Chromatin inaccessibility decreases dCas9 binding to other sites with matching seed sequences; thus 70% of off-target sites are associated with genes. The authors showed that targeted sequencing of 295 dCas9 binding sites in mESCs transfected with catalytically active Cas9 identified only one site mutated above background levels. The authors proposed a two-state model for Cas9 binding and cleavage, in which a seed match triggers binding but extensive pairing with target DNA is required for cleavage.

[0076] Platt et al. established a Cre-dependent Cas9 knockin mouse. The authors demonstrated in-vivo as well as ex vivo genome editing using adeno-associated virus (AAV)-, lentivirus-, or particle-mediated delivery of guide RNA in neurons, immune cells, and endothelial cells.

[0077] Hsu et al. (2014) is a review article that discusses generally CRISPR-Cas9 history from yogurt to genome editing, including genetic screening of cells.

[0078] Wang et al. (2014) relates to a pooled, loss-of-function genetic screening approach suitable for both positive and negative selection that uses a genome-scale lentiviral single guide RNA (sgRNA) library.

[0079] Doench et al. created a pool of sgRNAs, tiling across all possible target sites of a panel of six endogenous mouse and three endogenous human genes and quantitatively assessed their ability to produce null alleles of their target gene by antibody staining and flow cytometry. The authors showed that optimization of the PAM improved activity and also provided an on-line tool for designing sgRNAs.

[0080] Swiech et al. demonstrate that AAV-mediated SpCas9 genome editing can enable reverse genetic studies of gene function in the brain.

[0081] Konermann et al. (2015) discusses the ability to attach multiple effector domains, e.g., transcriptional activator, functional and epigenomic regulators at appropriate positions on the guide such as stem or tetraloop with and without linkers.

[0082] Zetsche et al. demonstrates that the Cas9 enzyme can be split into two and hence the assembly of Cas9 for activation can be controlled.

[0083] Chen et al. relates to multiplexing by demonstrating that a genome-wide in vivo CRISPR-Cas9 screen in mice reveals genes regulating lung metastasis.

[0084] Ran et al. (2015) relates to SaCas9 and its ability to edit genomes and demonstrates that one cannot extrapolate from biochemical assays.

[0085] Shalem et al. (2015) described ways in which catalytically inactive Cas9 (dCas9) fusions are used to synthetically repress (CRISPRi) or activate (CRISPRa) expression, showing advances using Cas9 for genome-scale screens, including arrayed and pooled screens, knockack approaches that inactivate genomic loci and strategies that modulate transcriptional activity.

[0086] Xu et al. (2015) assessed the DNA sequence features that contribute to single guide RNA (sgRNA) efficiency in CRISPR-based screens. The authors explored efficiency of CRISPR/Cas9 knockouts and nucleotide preference at the cleavage site. The authors also found that the sequence preference for CRISPRi/a is substantially different from that for CRISPR/Cas9 knockout.

[0087] Parnas et al. (2015) introduced genome-wide pooled CRISPR-Cas9 libraries into dendritic cells (DCs) to identify genes that control the induction of tumor necrosis factor (TNF) by bacterial lipopolysaccharide (LPS). Known regulators of TLR4 signaling and previously unknown candidates were identified and classified into three functional modules with distinct effects on the canonical responses to LPS.

[0088] Ramanan et al. (2015) demonstrated cleavage of viral episomal DNA (ecccDNA) in infected cells. The HBV genome exists in the nuclei of infected hepatocytes as a 3.2 kb double-stranded episomal DNA species called covalently closed circular DNA (ecccDNA), which is a key component in the HBV life cycle whose replication is not inhibited by current therapies. The authors showed that sgRNAs specifically targeting highly conserved regions of HBV robustly suppresses viral replication and depleted ecccDNA.
[0089] Nishimasu et al. (2015) reported the crystal structures of SaCas9 in complex with a single guide RNA (sgRNA) and its double-stranded DNA targets, containing the 5′-TTGAAT-3′ PAM and the 5′-TGGGGT-3′ PAM. A structural comparison of SaCas9 with SpCas9 highlighted both structural conservation and divergence, explaining their distinct PAM specificities and orthologous sgRNA recognition.

[0090] Camver et al. (2015) demonstrated a CRISPR-Cas9-based functional investigation of non-coding genomic elements. The authors developed pooled CRISPR-Cas9 guide RNA libraries to perform in situ saturating mutagenesis of the human and mouse BCL11A enhancers which revealed critical features of the enhancers.

[0091] Zetsche et al. (2015) reported characterization of Cpf1, a class 2 CRISPR nuclease from Francisella novicida U112 having features distinct from Cas9. Cpf1 is a single RNA-guided endonuclease lacking tracrRNA, utilizes a Trich protospacer-adjacent motif, and cleaves DNA via a staggered DNA double-stranded break.

[0092] Shmakov et al. (2015) reported three distinct Class 2 CRISPR-Cas systems. Two system CRISPR enzymes (C2c1 and C2c3) contain RuvC-like endonuclease domains distantly related to Cpf1. Unlike Cpf1, C2c1 depends on both crRNA and tracrRNA for DNA cleavage. The third enzyme (C2c2) contains two predicted HEPN RNase domains and is tracrRNA independent.

[0093] Slaymaker et al. (2016) reported the use of structure-guided protein engineering to improve the specificity of Streptococcus pyogenes Cas9 (SpCas9). The authors developed "enhanced specificity" SpCas9 (eSpCas9) variants which maintained robust on-target cleavage with reduced off-target effects.

[0094] The methods and tools provided herein are exemplified for Cas9, a type II nuclease that requires a tracrRNA. Orthologs of Cas9 have been identified in different bacterial species as described previously (e.g. WO2014093712). Further type II nucleases with similar properties can be identified using methods described in the art (Shmakov et al. 2015, 60,835-397; Abudayeh et al. 2016, Science, 5; 336,1629). In particular embodiments, such methods for identifying novel CRISPR effector proteins may comprise the steps of selecting sequences from the database encoding a seed which identifies the presence of a CRISPR Cas locus, identifying loci located within 10 kb of the seed comprising Open Reading Frames (ORFs) in the selected sequences, selecting therefrom loci comprising ORFs of which only a single ORF encodes a novel CRISPR effector having greater than 700 amino acids and no more than 90% homology to a known CRISPR effector. In particular embodiments, the seed is a protein that is common to the CRISPR-Cas system, such as Cas1. In further embodiments, the CRISPR array is used as a seed to identify new effector proteins.

[0095] The effectiveness of the present invention has been demonstrated. Preassembled recombinant CRISPR-Cas9 complexes comprising Cas9 and crRNA may be transfected, for example by electroporation, resulting in high mutation rates and absence of detectable off-target mutations. Hur, J. K. et al. Targeted mutagenesis in mice by electroporation of Cpf1 ribonucleoprotein, Nat Biotechnol. 2016 Jun. 6; doi: 10.1038/nbt.3596. [Epub ahead of print]. Genome-wide analyses shows that Cpf1 is highly specific. By one measure, in vitro cleavage sites determined for SpCas9 in human HEK293T cells were significantly fewer than for SpCas9. Kim, D. et al., Genome-wide analysis reveals specificities of Cpf1 endonucleases in human cells, Nat Biotechnol. 2016 Jun. 6. doi: 10.1038/nbt.3609. [Epub ahead of print]. An efficient multiplexed system employing Cas9 has been demonstrated in Drosophila employing gRNAs processed from an array containing inventing tRNAs. Port, F. et al. Expansion of the CRISPR toolbox in an animal with tRNA-RNA-linked Cas9 and Cpf1 gRNAs. doi: http://dx.doi.org/10.1101/046417.


[0100] In addition, mention is made of PCT application PCT/US14/70057, Attorney Reference 47627.99.2060 and BI-2013/107 entitled “DELIVERY, USE AND THERAPEUTIC APPLICATIONS OF THE CRISPR-CAS SYSTEMS AND COMPOSITIONS FOR TARGETING DISORDERS AND DISEASES USING PARTICLE DELIVERY COMPONENTS” (claiming priority from one or more of all of US provisional patent applications: 62/054, 490, filed Sep. 24, 2014; 62/010,441, filed Jun. 10, 2014; and 61/915,118, 61/915,215 and 61/915,148, each filed on Dec. 12, 2013) (“the Particle Delivery PCT”), incorporated herein by reference, and of PCT application PCT/US14/70127, Attorney Reference 47627.99.2091 and BI-2013/101 entitled “DELIVERY, USE AND THERAPEUTIC APPLICATIONS OF THE CRISPR-CAS SYSTEMS AND COMPOSITIONS FOR GENOME EDITING” (claiming priority from one or more of all of US provisional patent applications: 61/915,176; 61/915,192; 61/915,215; 61/915,107; 61/915,145; 61/915,148; and 61/915,153 each filed Dec. 12, 2013) (“the Eye PCT”), incorporated herein by reference, with respect to a method of preparing an sgRNA-and-Cas9 protein containing particle comprising adming a mixture comprising an sgRNA and Cas9 protein (and optionally HDR template) with a mixture comprising or consisting essentially of or consisting of surfactant, phospholipid, biodegradable polymer, lipoprotein and alcohol; and particles from such a process. For example, wherein Cas9 protein and sgRNA were mixed together at a suitable, e.g., 3:1 to 1:3 or 2:1 to 1:2 or 1:1 molar ratio, at a suitable temperature, e.g., 15-30 C, e.g., 20-25 C, e.g., room temperature, for a suitable time, e.g., 15-45, such as 30 minutes, advantageously in sterile, nuclelease free buffer, e.g., 1×PBS. Separately, particle components such as or comprising: a surfactant, e.g., cationic lipid, e.g., 1,2-dioleoyl-3-trimethylammonium-propane (DOTAP); phospholipid, e.g., dimyristoylphosphatidylcholine (DMPC); biodegradable polymer, such as an ethylene-glycol polymer or PEG, and a lipoprotein, such as a low-density lipoprotein, e.g., cholesterol were dissolved in an alcohol, advantageously a C1-6 alkyl alcohol, such as methanol, ethanol, isopropanol, e.g., 100% ethanol. The two solutions were mixed together to form particles containing the Cas9-sgRNA complexes. Accordingly, sgRNA may be pre-complexed with the Cas9 protein, before formulating the entire complex in a particle. Formulations may be made with a different molar ratio of different components known to promote delivery of nucleic acids into cells (e.g., 1,2-dioleoyl-3-trimethylammonium-propane (DOTAP), 1,2-ditridecanoyl-sn-glycero-3-phosphocholine (DMPC), polyethylene glycol (PEG), and cholesterol) For example DOTAP; DMPC: PEG: Cholesterol Molar Ratios may be DOTAP 100, DMPC 0, PEG 0, Cholesterol 0; or DOTAP 90, DMPC 0, PEG 10, Cholesterol 0; or DOTAP 90, DMPC 0, PEG 5, Cholesterol 5. DOTAP 100, DMPC 0, PEG 0, Cholesterol 0. That application accordingly comprehends adming sgRNA, Cas9 protein and components that form a particle; as well as particles from such adming. Aspects of the instant invention can involve particles; for example, particles using a process analogous to that of the Particle Delivery PCT or that of the Eye PCT, e.g., by adming a mixture comprising sgRNA and/or Cas9 as in the instant invention and components that form a particle, e.g., as in the Particle Delivery PCT or in the Eye PCT, to form a particle and particles from such adming (or, of course, other particles involving sgRNA and/or Cas9 as in the instant invention).

[0101] The subject invention may be used as part of a research program wherein there is transmission of results or data. A computer system (or digital device) may be used to receive, transmit, display and/or store results, analyze the data and/or results, and/or produce a report of the results and/or data and/or analysis. A computer system may be understood as a logical apparatus that can read instructions from media (e.g. software) and/or network port (e.g. from the internet), which can optionally be connected to a server having fixed media. A computer system may comprise one or more of a CPU, disk drives, input devices such as keyboard and/or mouse, and a display (e.g. a monitor). Data communication, such as transmission of instructions or reports, can be achieved through a communication medium to a server at a local or a remote location. The communication medium can include any means of transmitting and/or receiving data. For example, the communication medium can be a network connection, a wireless connection, or an internet connection. Such a connection can provide for communication over the World Wide Web. It is envisioned that data relating to the present invention can be transmitted over such networks or connections (or any other suitable means for transmitting information, including but not limited to mailing a physical report, such as a print-out) for reception and/or for review by a receiver. The receiver can be but is not limited to an individual, or electronic system (e.g. one or more computers, and/or one or more servers). In some embodiments, the computer system comprises one or more processors. Processors may be associated with one or more controllers, calculation units, and/or other units of a computer system, or implanted in firmware as desired. If implemented in software, the routines may be stored in any computer readable memory such as in RAM, ROM, flash memory, a magnetic disk, a laser disk, or other suitable
storage medium. Likewise, this software may be delivered to a computing device via any known delivery method including, for example, over a communication channel such as a telephone line, the internet, a wireless connection, etc., or via a transportable medium, such as a computer readable disk, flash drive, etc. The various steps may be implemented as various blocks, operations, tools, modules and techniques which, in turn, may be implemented in hardware, firmware, software, or any combination of hardware, firmware, and/or software. When implemented in hardware, some or all of the blocks, operations, techniques, etc. may be implemented in, for example, a custom integrated circuit (IC), an application specific integrated circuit (ASIC), a field programmable logic array (FPGA), a programmable logic array (PLA), etc. A client-server, relational database architecture can be used in embodiments of the invention. A client-server architecture is a network architecture in which each computer or process on the network is either a client or a server. Server computers are typically powerful computers dedicated to managing disk drives (file servers), printers (print servers), or network traffic (network servers). Client computers include PCs (personal computers) or workstations on which users run applications, as well as example output devices as disclosed herein. Client computers rely on server computers for resources, such as files, devices, and even processing power. In some embodiments of the invention, the server computer handles all of the database functionality. The client computer can have software that handles all the front-end data management and can also receive data input from users. A machine readable medium comprising computer-executable code may take many forms, including but not limited to, a tangible storage medium, a carrier wave medium or physical transmission medium. Non-volatile storage media include, for example, optical or magnetic disks, such as any of the storage devices in any computer(s) or the like, such as may be used to implement the databases, etc. shown in the drawings. Volatile storage media include dynamic memory, such as main memory of such a computer platform. Tangible transmission media include coaxial cables; copper wire and fiber optics, including the wires that comprise a bus within a computer system. Carrier-wave transmission media may take the form of electric or electromagnetic signals, or acoustic or light waves such as those generated during radio frequency (RF) and infrared (IR) data communications. Common forms of computer-readable media therefore include for example: a floppy disk, a flexible disk, hard disk, magnetic tape, any other magnetic medium, a CD-ROM, DVD or DVD-ROM, any other optical medium, punch cards paper tape, any other physical storage medium with patterns of holes, a RAM, a ROM, a PROM and EPROM, a FLASH-EPROM, any other memory chip or cartridge, a carrier wave transporting data or instructions, cables or links transporting such a carrier wave, or any other medium from which a computer may read programming code and/or data. Many of these forms of computer readable media may be involved in carrying one or more sequences of one or more instructions to a processor for execution. Accordingly, the invention comprehends performing any method herein-discussed and storing and/or transmitting data and/or results thereof and/or analysis thereof, as well as products from performing any method herein-discussed, including intermediates.

[0102] Throughout this disclosure there has been mention of CRISPR or CRISPR-Cas complexes or systems. CRISPR systems or complexes can target nucleic acid molecules, e.g., CRISPR-Cas9 complexes can target and cleave or nick or simply sit upon a target DNA molecule (depending if the Cas9 has mutations that render it a nickase or “dead”). Such systems or complexes are amenable for achieving tissuespecific and temporally controlled targeted deletion of candidate disease genes. Examples include but are not limited to genes involved in cholesterol and fatty acid metabolism, amyloid diseases, dominant negative diseases, latent viral infections, among other disorders. Accordingly, target sequences for such systems or complexes can be in candidate disease genes, e.g.:  

<table>
<thead>
<tr>
<th>Disease</th>
<th>GENE</th>
<th>SPACER</th>
<th>PAM</th>
<th>Mechanism</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hyper-cholesterolemia</td>
<td>SQLE</td>
<td>CGAGGAGACCCC</td>
<td>TGG</td>
<td>Knockout</td>
<td>Potential role of nonstatin cholesterol lowering agents (Trapani et al. JBBM Life, Volume 63, Issue 11, pages 964-971, November 2011)</td>
</tr>
</tbody>
</table>
Thus, the present invention, with regard to CRISPR or CRISPR-Cas complexes contemplates correction of hematopoietic disorders. For example, Severe Combined Immune Deficiency (SCID) results from a defect in lymphocytes T maturation, always associated with a functional defect in lymphocytes B (Cavazzana-Calvo et al., Annu. Rev. Med., 2005, 56, 585-602; Fischer et al., Immunol. Rev., 2005, 203, 98-109). In the case of Adenosine Deaminase (ADA) deficiency, one of the SCID forms, patients can be treated by injection of recombinant Adenosine Deaminase enzyme. Since the ADA gene has been shown to be mutated in SCID patients (Giblett et al., Lancet, 1972, 2, 1067-1069), several other genes involved in SCID have been identified (Cavazzana-Calvo et al., Annu. Rev. Med., 2005, 56, 585-602; Fischer et al., Immunol. Rev., 2005, 203, 98-109). There are four major causes for SCID: (i) the most frequent form of SCID, SCID-X1 (X-linked SCID or X-SCID), is caused by mutation in the IL2RG gene, resulting in the absence of mature T lymphocytes and NK cells. IL2RG encodes the gamma C protein (Noguchi, et al., Cell, 1993, 73, 147-157), a common component of at least five interleukin receptor complexes. These receptors activate several targets through the JAK3 kinase (Machi et al., Nature, 1995, 377, 65-68), which inactivation results in the same syndrome as gamma C inactivation; (ii) mutation in the ADA gene results in a defect in purine metabolism that is lethal for lymphocyte precursors, which in turn results in the quasi absence of B, T and NK cells; (iii) V(DJ) recombination is an essential step in the maturation of immunoglobulins and T lymphocyte receptors (TCRs). Mutations in Recombination Activating Gene 1 and 2 (RAG1 and RAG2) and Artemis, three genes involved in this process, result in the absence of mature T and B lymphocytes; and (iv) Mutations in other genes such as CD45, involved in T cell specific signaling have also been reported, although they represent a minority of cases (Cavazzana-Calvo et al., Annu. Rev. Med. 2005, 56, 585-602; Fischer et al., Immunol. Rev., 2005, 203, 98-109). In aspect of the invention, relating to CRISPR or CRISPR-Cas complexes contemplates system, the invention contemplates that it may be used to correct ocular defects that arise from several genetic mutations further described in Genetic Diseases of the Eye, Second Edition, edited by Elias I. Traboulsi, Oxford University Press, 2012. Non-limiting examples of ocular defects to be corrected include macular degeneration (MD), retinitis pigmentosa (RP). Non-limiting examples of genes and proteins associated with ocular defects include but are not limited to the following proteins: ABCA4 (ATP-binding cassette, sub-family A (ABC1), member 4 ACHM1 achromatopsia (rod monochromacy) 1 ApoE Apolipoprotein E (ApoE) C1QTNF5 (C1R) C1q and tumor necrosis factor related protein 5 (C1QTNF5) C2 Complement component 2 (C2) C3 Complement components (C3) CCL2 Chemokine (C-C motif) Ligand 2 (CCL2) CCR2 Chemokine (C-C motif) receptor 2 (CCR2) CD2 Cluster of Differentiation 36 CFB Complement factor B CFH Complement factor H CFHR1 complement factor H-related 1 CFHR3 complement factor H-related 3 CNGB3 cyclic nucleotide gated channel beta 3 CP ceruloplasmin (CP) CRP C reactive protein (CRP) CST3 cystatin C or cystatin 3 (CST3) CTSB Cathepsin D (CTSD) CXCR1 chemokine (C-X-C motif) receptor 1 ELAVL4 elongation of very long chain fatty acids 4 ERCC6 excision repair cross-complementing rodent repair deficiency, complementation group 6 FBLN5 Fibulin-5 FBLN6 Fibulin 6 FSCN2 fascin (FSCN2) HMCN1 Hemerican 1 HMCN1 hemerican 1 HTRA1 HtrA serine peptidase 1 (HTRA1) HTRA1 HtrA serine peptidase 1 IL-6 Interleukin 6 IL-8 Interleukin 8 LOC387715 Hypothetical protein PLEKHA1 Pleckstrin homology domain-containing family A member 1 (PLEKHA1) PROM1 Prominin 1(PROM1 or CD133) PRPH2 Peripherin-2 RGR retinitis pigmentosa GTPase regulator SERPING1 serpin peptidase inhibitor, clade G, member 1 (C1-inhibitor) TCOF1 Treacle TIMP3 Metalllopeptidase inhibitor 3 (TIMP3) TLR3 Toll-like receptor 3 The present invention, with regard to CRISPR or CRISPR-Cas complexes contemplates also contemplates delivering to the heart. For the heart, a myocardium troponin adena-associated virus (AAVM) is preferred, in particular AAVM41 which showed preferential gene transfer in the heart (see, e.g., Lim-Yang et al., PNAS, Mar. 10, 2009, vol. 106, no. 10). For example, US Patent Publication No. 20110023139, describes use of zinc finger nucleases to genetically modify cells, animals and proteins associated with cardiovascular disease. Cardiovascular diseases generally include high blood pressure, heart attacks, heart failure, and stroke and TIA. By way of example, the chromosomal sequence may comprise, but is not limited to, IL1B (interleukin 1, beta), XDH (xanthine dehydrogenase), TP53 (tumor protein p53), PTGIS (prostaglandin 12 (prostacyclin synthase), MI3 (myoglobin), IL4 (interleukin 4), ANGPT1 (angiopoietin 1), ABCG8 (ATP-binding cassette, sub-family G (WHITE), member 8), C1S (cathespin K), PTGIR (prostaglandin 12 (prostacyclin) receptor (IP)), KCNJ11 (potassium inwardly-rectifying channel, subfamily J, member 11), INS (insulin), CRP (C-reactive protein, pentraxin-related), PDGFBR (platelet-derived growth factor receptor, beta polypeptide), CCNA2 (cyclin A2), PDGFB (platelet-derived growth factor beta polypeptide (simian sarcoma viral (v-sis) oncogene homolog)), KCNJ5 (potassium inwardly-rectifying channel, subfamily J, member 5), KCN3 (potassium intermediate/small conductance calcium-activated channel, subfamily N, member 3), CAPN10 (calpain 10), PTGES (prostaglandin E synthase), ADRA2B (adrenergic, alpha-2B-, receptor), ABCG5 (ATP-binding cassette, sub-family G (WHITE), member 5), PRDX2 (peroxiredoxin 2), CAPN5 (calpain 5), PARP14 (poly (ADP-ribose) polymerase family, member 14), MEX3C (mex-3 homolog C (C. elegans)), ACE angiotensin I converting enzyme (pepallidy-I-depeptide A (1)), TNF (tumor necrosis factor (TNF superfamily), member 2), IL-6 (interleukin 6 (interferon, beta 2)), STN (statin), SERPINE1 (serpin peptidase inhibitor, clade E (nexin, plasminogen activator inhibitor type 1), member 1), ALB (albumin), ADIPQ (adiponectin, C1Q and collagen domain containing), APOB (apolipoprotein B (including Ag(x) antigens), APOE (apolipoprotein E), LEP (leptin), MTHFR (5,10-methylenetetrahydrofolate reductase (NADPH)), APOAI (apolipoprotein A-I), EDN1 (endothelin 1), NPPB (natriuretic peptide precursor B), NOS3 (nitric oxide synthase 3 (endothelial cell)), PPARG (peroxisome proliferator-activated receptor gamma), PLAT (plasminogen activator, tissue), PTGS2 (prostaglandin-endoperoxide synthase 2 (prostaglandin G/H synthase and cyclooxygenase)), CETP (cholesterol ester transfer protein, plasma), AGTRL1 (angiotensin II receptor, type 1), HMGCR (3-hydroxy-3-methyl-glutaryl-Coenzyme A reductase), IGF1 (insulin-like growth factor 1 (somatomedin C)), SELE (selectin E), REN (renin), PPARA (peroxisome proliferator-activated receptor alpha),
PON1 (paraoxonase 1), KNG1 (kininogen 1), CCL2 (chemokine (C-C motif) ligand 2), LPL (lipoprotein lipase), VWF (von Willebrand factor), F2 (coagulation factor II (thrombin)), ICAM1 (intercellular adhesion molecule 1), TGFβ1 (transforming growth factor, beta 1), NPPA (natriuretic peptide precursor A), IL10 (interleukin 10), EPO (erythropoietin), SOD1 (superoxide dismutase 1, soluble), VCAM1 (vascular cell adhesion molecule 1), IFNγ (interferon, gamma), LPA (lipoprotein, Lp(a)), MPO (myeloperoxidase), ESR1 (estrogen receptor 1), MAPK1 (mitogen-activated protein kinase 1), HP (haptoglobin), F3 (coagulation factor III (thromboplastin, tissue factor)), CST3 (cystatin C), COG2 (component of oligomeric golgi complex 2), MMP9 (matrix metalloproteinase 9 (gelatinase B, 92 kDa gelatinase, 92 kDa type IV collagenase)), SERPIN1 (serpin peptidase inhibitor, clade C (antithrombin), member 1), F8 (coagulation factor VIII, procoagulant component), HMOX1 (heme oxygenase (deoxyelinate 1)), APOC3 (apolipoprotein C-III), IL8 (interleukin 8), PROK1 (prokineticin 1), CBS (cystathionine-beta-synthase), NOS2 (nitric oxide synthase 2, inducible), ITLR4 (toll-like receptor 4), SELL (selectin P) (granule membrane protein 140 kDa, antigen CD62), ABCA1 (ATP-binding cassette, sub-family A (ABC), member 1), AGT (angiotensinogen (serpin peptidase inhibitor, clade A, member 8)), LDLR (low density lipoprotein receptor), GPT (glutamic-pyruvate transaminase), VEGFA (vascular endothelial growth factor A), NR3C2 (nuclear receptor subfamily 3, group C, member 2), IL18 (interleukin 18, interleukin-gamma inducing factor), NOS1 (nitric oxide synthase 1 (neuronal)), NR3C1 (nuclear receptor subfamily 3, group C, member 1 (glucocorticoid receptor)), FGB (fibrinogen beta chain), HGF (hepatocyte growth factor (hepatopoietin A, scatter factor), IL1A (interleukin 1, alpha), RETN (resistin), AKT1 (v-akt murine thymoma viral oncogene homolog 1), LPC (lipase, hepatic), HSPD1 (heat shock 60 kDa protein 1 (chaperonin)), MAPK14 (mitogen-activated protein kinase 14), SPP1 (secreted phosphoprotein 1), ITGβ3 (integrin, beta 3 (platelet glycoprotein IIb/IIIa, antigen CD61)), CAT (catalase), UTS2 (urotensin 2), THBD (thrombomodulin), F10 (coagulation factor X), CP (ceruloplasmin (ferroxidase)), TFNRF51B1 (tumor necrosis factor receptor superfamily, member 1B), EDNRA (endothelin receptor type A), EGFR (epidermal growth factor receptor (erythroblastic leukemia viral (v-erb-b) oncogene homolog, avian)), MMP2 (matrix metalloproteinase 2 (gelatinase A, 72 kDa gelatinase, 72 kDa type IV collagenase)), PLG (plasminogen), NPY (neuropeptide Y), RHOD (rat homolog gene family, member D), MAPK8 (mitogen-activated protein kinase 8), MYC (v-myc myelocytomatosis viral oncogene homolog (avian)), FN1 (fibronectin 1), CMA1 (chymase 1, mast cell), PLAU (plasminogen activator, urokinase), GNB3 (guanine nucleotide binding protein (G protein, beta polypeptide 3)), ADRB2 (adrenergic, beta-2, receptor, surface), APOA5 (apolipoprotein A-V), SOD2 (superoxide dismutase 2, mitochondrial), F5 (coagulation factor V (proaccelerin, labile factor)), VDR (vitamin D (1,25-dihydroxyvitamin D3) receptor), ALOX5 (arachidonate 5-lipoxygenase), I.HA-DRB1 (major histocompatibility complex, class II, DR beta 1), PARP1 (poly(ADP-ribose) polymerase 1), CD40LG (CD40 ligand), PON2 (paraoxonase 2), AGER (advanced glycosylation end product-specific receptor), IRS1 (insulin receptor substrate 1), PTGS1 (prostaglandin-endoperoxide synthase 1 (prostaglandin G/H synthase and cyclooxygenase)), ECE1 (endothelin converting enzyme 1), F7 (coagulation factor VII (serum prothrombin conversion accelerator)), URN (interleukin 1 receptor antagonist), EPHX2 (epoxide hydrolase 2, cytoplasmic), IGFBP1 (insulin-like growth factor binding protein 1), MAPK10 (mitogen-activated protein kinase 10), FAS (Fas (TNF receptor superfamily, member 6)), ABCB1 (ATP-binding cassette, subfamily B (MDR/TAP), member 1), JUN (jun oncogene), IGFBP3 (insulin-like growth factor binding protein 3), CD14 (CD14 molecule), PDE5A (phosphodiesterase 5A, cGMP-specific), AGTR2 (angiotensin II receptor, type 2), CD40 (CD40 molecule, TNF receptor superfamily member 5), LCAT (lecithin-cholesterol acyltransferase), CCR5 (chemokine (C-C motif) receptor 5), MMP1 (matrix metalloproteinase 1 (interstitial collagenase)), TIMP1 (TIMP metalloproteinase inhibitor 1), ADM (adrenomedullin), DTY10 (dystonia 10), STAI3 (signal transducer and activator of transcription 3 (acute-phase response factor)), MMP3 (matrix metalloproteinase 3 (stromelysin 1, progelatinase)), ELN (elastin), USF1 (upstream transcription factor 1), CFI (complement factor H), HSPA4 (heat shock 70 kDa protein 4), MMP12 (matrix metalloproteinase 12 (macrophage elastase)), MME (membrane metallo-endopeptidase), F2R (coagulation factor II (thrombin) receptor), SELL (selectin L), CTSB (cathepsin B), ANXA5 (annexin A5), ADR1B (adrenergic, beta-1, receptor), CYBA (cytchrome b-245, alpha polypeptide), FGA (fibrinogen alpha chain), GGT1 (gamma-glutamyltransferase 1), LIPG (lipase, endothelial), HI1A (hypoxxia inducible factor 1, alpha subunit (basic helix-loop-helix transcription factor)), CXCR4 (chemokine (C-X-C motif) receptor 4), PROC (protein C (inactivator of coagulation factors Va and VIIa)), SCAR1B (scavenger receptor class B, member 1), CD79A (CD79a molecule, immunoglobulin-associated alpha), PLTP (phospholipid transfer protein), ADD1 (adducin 1 (alpha)), FGB (fibrinogen gamma chain), SAA1 (serum amyloid A1), KCNH2 (potassium voltage-gated channel, subfamily H ( eag-related), member 2), DPP4 (dipeptidyl-peptidase 4), GrpLP (glucose-6-phosphate dehydrogenase), NPR1 (natriuretic peptide receptor A/guanylate cyclase A (atrionatriuretic peptide receptor A)), VTIN (vitronectin), KIAA0101 (KIAA0101), FOS (FBJ murine osteosarcoma viral oncogene homolog), TLR2 (toll-like receptor 2), PPIG (peptidylprolyl isomerase G (cyclophilin G)), IL1R1 (interleukin 1 receptor, type I), AR (androgen receptor), CYP1A1 (cytchrome P450, family 1, subfamily A, polypeptide 1), SERPIN1 (serpin peptidase inhibitor, clade A (alpha-antitrypsin, antitrypsin), member 1), MTR (5-methyltetrahydrofolatredoxorhodofolate-homocysteine methyltransferase), RBP4 (retinol binding protein 4, plasma), APOA4 (apolipoprotein A-IV), CDKN2A (cyclin-dependent kinase inhibitor 2A (melanoma, p16, inhibits CDK4)), FGFR2 (fibroblast growth factor 2 (basic)), EDNRB (endothelin receptor type B), ITGβ2 (integrin, alpha 2 (CD49B, alpha 2 subunit of VLA-2 receptor)), CABN1 (calcineurin binding protein 1), SHBG (sex hormone-binding globulin), HMG1 (high-mobility group box 1), HSP90B2P (heat shock protein 90 kDa beta (Grp94), member 2 (pseudogene)), CYP3A4 (cytchrome P450, family 3, subfamily A, polypeptide 4), GA1 (gap junction protein, alpha 1, 43 kDa), CAV1 (caveolin 1, caveoleae protein, 22 kDa), ESRR2 (estrogen receptor 2 (ER beta)), LTA (lymphotoxin alpha (TNF superfamily, member 1)), GDF15 (growth differentiation factor 15), BDNF (brain-derived neurotrophic factor), CYP2D6 (cytchrome P450,
family 2, subfamily D, polypeptide 6), NGF (nerve growth factor (beta polypeptide)), SP1 (Sp1 transcription factor), TGFβ1 (TGFB-induced factor homeobox 1), SRC (v-src sarcoma (Schmidt-Ruppin A-2) viral oncogene homolog (avian)), EGF (epidermal growth factor (beta-urogastrone)), PKCγ (phosphoinositide-3-kinase, catalytic, gamma polypeptide), HLA-A (major histocompatibility complex, class I, A), KCNQ1 (potassium voltage-gated channel, KQT-like subfamily, member 1), CNR1 (cannabinoid receptor 1 (brain)), FBXN1 (fibrillin 1), CHKA (choline kinase alpha), BEST1 (bestrophin 1), APP (amyloid beta (A4) precursor protein), CTNNB1 (catenin (cadherin-associated protein), beta 1, 88 kDa), IL2 (interleukin 2), CD36 (CD36 molecule (thrombospondin receptor)), PRKAB1 (protein kinase, AMP-activated, beta 1 non-catalytic subunit), TPO (thyroid peroxidase), ALDH1A1 (aldehyde dehydrogenase 7 family, member A1), CXCRC1 (chemokine (C-X-C motif) receptor 1), TH1 (tyrosine hydroxylase), F9 (coagulation factor IX), GH1 (growth hormone 1), TF (transferrin), HFE (hemochromatosis), IL7A (interleukin 17A), PTE (phosphatase and tensin homolog), GSTM1 (glutathione S-transferase mu 1), DMD (dystrophin), GATA4 (GATA binding protein 4), F13A1 (coagulation factor XIII, A1 polypeptide), TTR (transthyretin), FABP4 (fatty acid binding protein 4, adipocyte), PON3 (paraoxonase 3), APOC1 (apoprotein C-I), INSR (insulin receptor), TNRFRF1B (tumor necrosis factor receptor superfamily, member 1B), HTR2A (5-hydroxytryptamine (serotonin) receptor 2A), CSF3 (colony stimulating factor 3 (granulocyte)), CYP2C9 (cytochrome P450, family 2, subfamily C, polypeptide 9), TTN (thioredoxin), CYP11B2 (cytochrome P450, family 11, subfamily B, polypeptide 2), PTH (parathyroid hormone), CSF2 (colony stimulating factor 2 (granulocyte-macrophage)), KDR (kinase insert domain receptor tyrosine kinase), PLA2G2A (phospholipase A2, group II A (platelets, synovial fluid)), B2M (beta-2-microglobulin), THBS1 (thrombospondin 1), CGC (glucagon), RHOA (ras homolog gene family, member A), ALDH2 (aldehyde dehydrogenase 2 family (mitochondrial)), TCF7L2 (transcription factor 7-like 2 (T-cell specific, HMG-box)), BDKRB2 (bradykinin receptor B2), NFE2L2 (nuclear factor (erythroid-derived 2)-like 2), NOTCH1 (Notch homolog 1, translocation-associated (Drosophila)), UGT1A1 (UDP glucuronosyltransferase 1 family, polypeptide A1), IPN1A (interferon, alpha 1), PPARD (peroxisome proliferator-activated receptor delta), SIRT1 (sirtuin (silent mating type information regulation 2 homolog) 1 (S. cerevisiae)), GNRH1 (gonadotropin-releasing hormone 1 (luteinizing-releasing hormone)), PAPPB (pregnancy-associated plasma protein A, pappalysin 1), AR3 (arrestin 3, retinal (X-arrestin)), NPPC (natriuretic peptide precursor C), AHSF (alpha hemoglobin stabilizing protein), PTK2 (PTK2 protein tyrosine kinase 2), IL13 (interleukin 13), MTROR (mechanistic target of rapamycin (serine/threonine kinase)), ITGB2 (integrin, beta 2 (complement component 3 receptor 3 and 4 subunits)), GSTT1 (glutathione S-transferase theta 1), IL6ST (interleukin 6 signal transducer (gp130, oncostatin M receptor)), CPBP2 (carboxypeptidase B2 (plasma)), CYP1A2 (cytochrome P450, family 1, subfamily A, polypeptide 2), HNF4A (hepatocyte nuclear factor 4, alpha), SLCO6A4 (solute carrier family 6 (neurotransmitter transporter, serotonin), member 4), PLA2G6 (phospholipase A2, group VI (cytosolic, calcium-independent)), TNFRSF11 (tumor necrosis factor (ligand) superfamily, member 11), SLC6A1 (solute carrier family 8 (sodium/calcium exchanger), member 1), F2RL1 (coagulation factor II (thrombin) receptor-like 1), AKR1A1 (aldo-keto reductase family 1, member A1 (aldoxide reductase)), ALDH9A1 (aldehyde dehydrogenase 9 family, member A1), BGLAP (bone gamma-carboxyglutamate (glia) protein), MMP9 (microsomal trypsin-like transfer protein), MTRR (5-methyltetrahydrofolate-homocysteine methyltransferase reductase), SULT1A3 (sulfotransferase family, cytosolic, 1A, phenol-prefering, member 3), RAGE (renal tumor antigen), C4B (complement component 4B (Chido blood group)), P2RY12 (purinergic receptor P2Y, G-protein coupled, 12), RNLS (renilase, FAD-dependent amine oxidase), CREB1 (AMP responsive element binding protein 1), POMC (proopiomelanocortin), RAC1 (ras-related C3 botulinum toxin substrate 1 (rho family, small GTP binding protein Rac1)), LMNA (lamin NC), CD59 (CD59 molecule, complement regulatory protein), SCN5A (sodium channel, voltage-gated, type V, alpha subunit), CYP1B1 (cytochrome P450, family 1, subfamily B, polypeptide 1), MIF (macrophage migration inhibitory factor (glycosylation-inhibiting factor)), MMP13 (matrix metalloproteinase 13 (collagenase 3)), TIMP2 (TIMP metalloproteinase inhibitor 2), CYP19A1 (cytochrome P450, family 19, subfamily A, polypeptide 1), CYP21A2 (cytochrome P450, family 21, subfamily A, polypeptide 2), PTNP22 (protein tyrosine phosphatase, non-receptor type 22 (lymphoid)), MTH14 (myosin, heavy chain 14, non-muscle), MBL2 (mannose-binding lectin (protein C) 2, soluble (opsonic defect)), SELPLG (selectin P ligand), AOC3 (amine oxidase, copper containing 3 (vascular adhesion protein 1)), CTS1 (cathepsin L1), PCNA (proliferating cell nuclear antigen), IGF2 (insulin-like growth factor 2 (somatomedin A)), ITGB1 (integrin, beta 1 (fibronectin receptor, beta polypeptide, antigen CD29 includes MDF2, MSK12)), CST (calpastatin), CXCL12 (chemokine (C-X-C motif) ligand 12 (stromal cell-derived factor 1)), IGHE (immunoglobulin heavy constant epsilon), KCN1 (potassium voltage-gated channel, Isk-related family, member 1), TFRC (transferrin receptor (p90, CD71)), COL1A1 (collagen, type I, alpha 1), COL1A2 (collagen, type I, alpha 2), IL2RB (interleukin 2 receptor, beta), PLAZG10 (phospholipase A2, group X), ANGPT2 (angiopoietin 2), PROCR (protein C receptor, endothelial (EPCR)), NOX4 (NADPH oxidase 4), HAMP (hepcidin antimicrobial peptide), PTTPN11 (protein tyrosine phosphatase, non-receptor type 11), SLCO2A1 (soluble carrier family 2 (facilitated glucose transporter), member 1), IL2RA (interleukin 2 receptor, alpha), CCL5 (chemokine (C-C motif) ligand 5), IRF1 (interferon regulatory factor 1), CFLAR (CASP8 and FADD-like apoptosis regulator), CALCA (calcitonin-related polypeptide alpha), ELF4E (eukaryotic translation initiation factor 4E), GSTP1 (glutathione S-transferase pi 1), JAK2 (Janus kinase 2), CYP3A5 (cytochrome P450, family 3, subfamily A, polypeptide 5), HSPA5 (hepoxan sulfat proteoglycan 2), CCL3 (chemokine (C-C motif) ligand 3), MYD88 (myeloid differentiation primary response gene (88)), VIP (vasoactive intestinal peptide), SOAT1 (sterol O-acetyltransferase 1), ADRBK1 (adrenergic, beta, receptor kinase 1), NR4A1 (nuclear receptor subfamily 4, group A, member 2), MMP8 (matrix metalloproteinase 8 (neutrophil collagenase)), NPR2 (natriuretic peptide receptor B/guanylate cyclase B (atrio natriuretic peptide receptor B)), GCH1 (GTP cyclohydrolase 1), EPRS (glutamy-oxylutyl-RNA synthetase), PPARGC1A (peroxisome proliferator-activated receptor gamma, coactivator 1 alpha), F12 (coagu-
lation factor XII (Hageman factor)), PECAM1 (platelet/endothelial cell adhesion molecule), CCL4 (chemokine (C-C motif) ligand 4), SERPINA3 (serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 3), CASR (calcium-sensing receptor), GJA5 (gap junction protein, alpha 5, 40 kDa), FABP2 (fatty acid binding protein, 2), intestinal), TTF2 (transcription termination factor, RNA polymerase II), PROS1 (protein S (alpha)), CTF1 (cardiothrophin 1), SGCB (sarcoglycan, beta (43 kDa dystrophin-associated glycoprotein)), YME1L1 (YME1-like 1 (S. cerevisiae)), CAMP (cathelicidin antimicrobial peptide), ZC3H12A (zinc finger CCHC-type containing 12A), AKR1B1 (aldo-keto reductase family 1 member B1 (aldose reductase)), DES (desmin), MMP7 (matrix metalloproteinase 7 (matrilysin, uterine)), AHR (aryl hydrocarbon receptor), CSF1 (colony stimulating factor 1 (macrophage)), HDAC9 (histone deacetylase 9), CTGF (connective tissue growth factor), KCNMA1 (potassium large conductance calcium-activated channel, subfamily M, alpha member 1), UGT1A1 (UDP glucuronosyltransferase 1 family, polypeptide A complex locus), PRKCA (protein kinase C, alpha), COMT (catechol-O-methyltransferase), S100B (S100 calcium binding protein B), EGFR (early growth response 1), PRL (prolactin), IL15 (interleukin 15), DRD4 (dopamine receptor D4), CAMK2G (calcium/calmodulin-dependent protein kinase II gamma), SLC22A2 (solute carrier family 22 (organic cation transporter), member 2), CCL11 (chemokine (C-C motif) ligand 11), PIGF (placental growth factor), TPH2 (tryptophan-2,3-dioxygenase, 1P1 (platelet), TACR1 (tachykinin receptor 1), NT5 (neurotensin), HNFA1 (HNFA1 homeobox A), SSTR (somatostatin receptor), KCND1 (potassium voltage-gated channel, Shal-related subfamily, member 1), LOC646627 (phospholipase B inhibitor), TBSX1 (thrombomodulin synthase 1 (platelet), CYP2J2 (cytochrome P450, family 2 subfamily J, polypeptide 2), TBX2A (thrombomodulin 2, alpha), ADH1C (alcohol dehydrogenase 1C (class 1), gamma polypeptide), ALOX12 (arachidonate 12-lipoxygenase), AHSG (alpha-2-HS-glycoprotein), BHM (beta-hemocysteine methyltransferase), GJA4 (gap junction protein, alpha 4, 37 kDa), SLC25A4 (solute carrier family 25 (mitochondrial carrier, adenine nucleotide translocator), member 4), ACLY (ATP citrate lyase), ALOX5AP (arachidonate 5-lipoxygenase activating protein), NUMA1 (nuclear mitotic apparatus protein 1), CYP2B11 (cytochrome P450, family 27, subfamily B, polypeptide 1), CYSLT2 (cysteinyl leukotriene receptor 2), SOD3 (superoxide dismutase 3, extracellular), LTC4S (leukotriene C4 synthase), UCN (urocortin), GHRL (ghrelin/obestatin prepropeptide), APOC2 (apolipoprotein C-II), LECA4 (C-type lectin domain family 4, member A), KBTBD10 (kelch repeat and BTB (POZ) domain containing 10), TNC (tenascin C), TMY3 (thymidylate synthetase), SHC1 (SRC homology 2 domain containing transforming protein 1), LRPI (low density lipoprotein receptor-related protein 1), SOCS3 (suppressor of cytokine signaling 3), ADH1B (alcohol dehydrogenase 1B (class 1)), beta polypeptide), KLK3 (kallikrein-related peptidase 3), HSD11B1 (hydroxysteroid (11-beta) dehydrogenase 1), VKORC1 (vitamin K epoxide reductase complex subunit 1), SERPINB2 (serpin peptidase inhibitor, clade B (ovalbumin), member 2), TNS1 (tensin 1), RNF19A (ring finger protein 19A), EPO (erythropoietin receptor), ITGAM (integrin, alpha M (complement component 3 receptor 3 subunit)), PTX2 (paired-like homeodomain 2), MAPK7 (mitogen-activated protein kinase 7), FCGR3A (Fc fragment of IgG, low affinity IIa, receptor (CD16a)), LEPR (leptin receptor), ENG (endoitin), GPK1 (glutathione peroxidase 1), GOT2 (glutamic-oxaloacetic transaminase 2, mitochondrial (aspartate aminotransferase 2)), IGRH1 (histamine receptor 1H), NR112 (nucleotide receptor subfamily 1, group I, member 2), CRH (corticotropin releasing hormone), HTR1A (5-hydroxytryptamine (serotonin) receptor 1A), VDAC1 (voltage-dependent anion channel 1), HPSE (hepsanase), SFTPD (surfactant protein D), TAP2 (transporter 2, ATP-binding cassette, subfamily B (MDR/TAP)), RNF123 (ring finger protein 123), PTKB2 (PTKB2 protein tyrosine kinase 2 beta), NTRK2 (neurotrophic tyrosine kinase, receptor, type 2), IL6R (interleukin 6 receptor), ACH (acetylcholinesterase (Y1 blood group)), GLPRI (glucagon-like peptide 1 receptor), GHR (growth hormone receptor), GSN (growth hormone), NQO1 (NAD(P)H dehydrogenase, quinone 1), NR5A1 (nuclear receptor subfamily 5, group A, member 1), GJB2 (gap junction protein, beta 2, 26 kDa), SLC9A1 (solute carrier family 9 (sodium/hydrogen exchanger), member 1), MAOA (monoamine oxidase A), PCSK9 (proprotein convertase subtilisin/kexin type 9), FCGR2A (Fc fragment of IgG, low affinity IIa, receptor (CD32)), SERPINF1 (serpin peptidase inhibitor, clade F (alpha-2 antiplasmin, pigment epithelium derived factor), member 1), EED3 (endothelin 3), DHFR (dihydrofolate reductase), GAS6 (growth arrest-specific 6), SMPD1 (sphingomyelin phosphodiesterase 1, acid lysosomal), UCP2 (uncoupling protein 2 (mitochondrial, proton carrier)), TAF1A2 (transcription factor AP-2 alpha (activating enhancer binding protein 2 alpha)), C4BPA (complement component 4 binding protein, alpha), SERPINF2 (serpin peptidase inhibitor, clade F (alpha-2 antiplasmin, pigment epithelium derived factor), member 2), TYMP (thymidine phosphorylase), ALPP (alkaline phosphatase, placental (Regan isozyme)), CCR2 (chemokine (C-C motif) receptor 2), SLCO3A3 (solute carrier family 39 (zinc transporter), member 3), ABCG2 (ATP-binding cassette, sub-family G (WHITE), member 2), ADA (adenosine deaminase), JAK3 (Janus kinase 3), HSPA1A (heat shock 70 kDa protein 1A), EASN (fatty acid synthase), FGF1 (fibroblast growth factor 1 (acidic)), F11 (coagulation factor XI), APT7A (APThase, Cu+1 transporting, alpha polypeptide), CR1 (complement component (3b/4b) receptor 1 (Knops blood group)), GAPF (glial fibrillary acidic protein), ROCK1 (Rho-associated, coiled-coil containing protein kinase 1), MECP2 (methyl CpG binding protein 2 (Rett syndrome)), MYLK (myosin light chain kinase), BCSHE (butyrylcholinesterase), LRP5 (lipase, hormone-sensitive), PRDX5 (peroxiredoxin 5), ADORA1 (adenosine A1 receptor), WRN ( Werner syndrome, RecQ helicase-like), CCRX3 (chemokine (C-X-C motif) receptor 3), CD81 (CD81 molecule), SMAD7 (SMAD family member 7), LAMC2 (laminin, gamma 2), MAP3K5 (mitogen-activated protein kinase kinase 5), CHGA (chromogranin A (parathyroid secretary protein 1)), IAPP (islet amyloid polypeptide), RHO (rhodopsin), ENPP1 (ectonucleotide pyrophosphatase/phosphodiesterase 1), PTHLH (parathyroid hormone-like hormone), NRG1 (neuregulin 1), VEGF (vascular endothelial growth factor C), ENPEP (gutamyl aminopeptidase (aminopeptidase A)), CEBPB (CCAAT/enhancer binding protein (C/EBP), beta), NAGLU (N-acetylglucosaminidase, alpha-), F2RL3 (coagulation factor II (thrombin) receptor-like 3), CXCL1 (chemokine (C-X-C motif) ligand 1), BDKKR31 (bradyki-
nin receptor B1), ADAMTS13 (ADAM metalloprotease with thrombospondin type 1 motif, 13), ELANE (elastase, neutrophil expressed), ENPP2 (ectonucleotide pyrophosphatase/phosphodiesterase 2), CISH (cytokine inducible SH2-containing protein), GAST (gastrin), MYOC (myocilin, trabecular meshwork inducible glucocorticoid response), ATP1A2 (ATPase, Na+/K+ transporting, alpha 2 polypeptide), NF1 (neurofibromin 1), JGB1 (gap junction protein beta 1, 32 kDa), MEF2A (myocyte enhancer factor 2A), VCL (vinculin), BPM2 (bone morphogenetic protein receptor, type II (serine/threonine kinase)), TUBB (tubulin, beta), CDC42 (cell division cycle 42 (GTP binding protein, 25 kDa)), KRT18 (keratin 18), HSF1 (heat shock transcription factor 1), MYB (v-myb myeloblastosis viral oncogene homolog (avian)), PRKAA2 (protein kinase, AMP-activated, alpha 2 catalytic subunit), ROCK2 (Rho-associated, coiled-coil containing protein kinase 2), TTP1 (tissue factor pathway inhibitor (lipoprotein-associated coagulation inhibitor)), PRKG1 (protein kinase, cGMP-dependent, type I), BMP2 (bone morphogenetic protein 2), CTNNOD1 (caldesmolin-associated protein), delta 1), CTH (cystathionine (cystathionine gamma-lyase)), CTSS (cathepsin S), SAV2 (vav 2 guanine nucleotide exchange factor), NPY2R (neuropeptide Y receptor Y2), IGFBP2 (insulin-like growth factor binding protein 2, 36 kDa), CD28 (CD28 molecule), GSTA1 (glutathione S-transferase alpha 1), PPIA (peptidylprolyl isomerase A (cyclophilin A)), APOH (apoprotein H (beta-2-glycoprotein I)), S100A8 (S100 calcium binding protein A8), IL11 (interleukin 11), ALOX15 (arachidonate 15-lipoxygenase), FBLN1 (fibulin 1), NR1H3 (nuclear receptor subfamily 1, group H, member 3), SCD (stearoyl-CoA desaturase (delta-9-desaturase)), GIP (gastric inhibitory polypeptide), CHGB (chromogranin B (secretogranin 1)), PRKCB (protein kinase, beta), SRD5A1 (steroid-5-alpha-reductase, alpha polypeptide 1 (3-oxo-5-alpha-steroid 4-dehydrogenase (3b-hydroxysteroid dehydrogenase 1 alpha)), HSD11B2 (hydroxysteroid (11-beta) dehydrogenase 2), CALCRL (calcitonin receptor-like), GALNT2 (UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 2 (GalNAc-T2)), ANGPTL4 (angiopoietin-like 4), KCN4N (potassium intermediate/small conductance calcium-activated channel, subfamily N, member 4), PIK3C2A (phosphoinositide-3-kinase, class C, alpha polypeptide), HHBE GF (heparin-binding EGF-like growth factor), CYP7A1 (cytochrome P450, family 7, subfamily A, polypeptide 1), HLA-DRB5 (major histocompatibility complex, class II, DR beta 5), DNN3P (BCL2/adenovirus E1B 19 kDa interacting protein 3), GCKR (glucokinase (hexokinase 4) regulator), S100A12 (S100 calcium binding protein A12), PADI4 (peptidyl arginine deiminase, type IV), HSPA14 (heat shock protein 70 kDa protein 14), CXYR1 (chemokine (C-X-C motif) receptor 1), H119 (H119, imported maternally expressed transcript (non-protein coding)), KRTAP19-3 (keratin associated protein 19-3), IDDM2 (insulin-dependent diabetes mellitus 2), RAC2 (ras-related C3 botulinum toxin substrate 2 (rho family, small GTP binding protein Rac2)), RYR1 (ryanodine receptor 1 (skeletal)), CLOCK (clock homolog (mouse)), NGFR (nerve growth factor receptor (TNFR superfamily, member 16)), DJB1 (dopamine beta-hydroxylase (dopamine beta-monooxygenase)), CHRNA4 (cholinergic receptor, nicotinic, alpha 4), CACNA1C (calcium channel, voltage-dependent, L type, alpha 1C subunit), PRKAG2 (protein kinase, AMP-activated, gamma 2 non-catalytic subunit), CHAT (choline acetyltransferase), PTGIS (prostaglandin D2 synthase 21 kDa (brain)), NR1H2 (nuclear receptor subfamily 1, group H, member 2), TEK (TEK tyrosine kinase, endothelial, VEGFB (vascular endothelial growth factor B), MEF2C (myocyte enhancer factor 2C), MAPKAPK2 (mitogen-activated protein kinase-activated protein kinase 2), TNFRSF11A (tumor necrosis factor receptor superfamily, member 11a, NFκB activator), HSPA9 (heat shock 70 kDa protein 9 (mortalin)), CYSLTR1 (cysteiny1 leukotrien receptor 1), MAT1A (methylene adenosyltransferase I, alpha), OPRL1 (opiate receptor-like 1), IMPA1 (inositol (myo)-1(4)-monophosphatase 1), CLCN2 (chloride channel 2), DLD (dihydrodipicarboxamide dihydrogenase), PSMAl6 (prostatese (prosome, macropain) subunit, alpha type, 6), PSMB8 (prostatese (prosome, macropain) subunit, beta type, 8 (large multifunctional peptidase 7)), CHI3L1 (chitinase 3-like 1 (cartilage glycoprotein 39)), ALDH1B1 (aldehyde dehydrogenase 1 family, member B1), PARP2 (poly (ADP-ribose) polymerase 2), STAR (steroidogenic acute regulatory protein), LBP (lipopolysaccharide binding protein), ABCC6 (ATP-binding cassette, sub-family C (CFTR/MRP), member 6), RGS2 (regulator of G-protein signaling 2, 24 kDa), EFNB2 (ephrin-B2), GBJ6 (gap junction protein, beta 6, 30 kDa), APOA2 (apolipoprotein A-II), AMPD1 (adenosine monophosphate deaminase 1), DYSF (dysferlin, limb girdle muscular dystrophy 2B (autosomal recessive)), FDFT1 (farnesyl-diphosphate farnesyltransferase 1), EDN2 (endothelin 2), CCR6 (chemokine (C-C motif) receptor 6), GJB3 (gap junction protein, beta 3, 31 kDa), IL1R1 (interleukin 1 receptor-like 1), ENTPD1 (ectonucleoside triphosphate diphosphohydrolase 1), BB84 (Bardet-Biedl syndrome 4), CERS2 (caldesmin, EGF LG7 seven-pass G-type receptor 2 (flamingo homolog, Drosophila)), F11R (F11 receptor), RAPGEF3 (Rap guanine nucleotide exchange factor (GEF)), HSYL1 (hyaluronoglucosaminidase), ZNF259 (zinc finger protein 259), ATOX1 (ATX1 antioxidant protein 1 homolog (yeast)), ATF6 (activating transcription factor 6), HKH (ketohexokinase (fructokinase)), SAT1 (spermidine/spermine N1-acetyltransferase 1), GGH (gamma-glutamyl hydrolase (conjugase, folypolyglycinnamylglutamyl hydrolase)), TIMP4 (TIMP metalloproteinidase inhibitor 4), SLC4A4 (solute carrier family 4, sodium bicarbonate cotransporter, member 4), PDE2A (phosphodiesterase 2A, cGMP-stimulated), PDE3B (phosphodiesterase 3B, cGMP-inhibited), FADS1 (fatty acid desaturase 1), FADS2 (fatty acid desaturase 2), TMSB4X (thymosin beta 4, X-linked), TNXIP (thioredoxin interacting protein), LIMS1 (LIM and senescent cell antigen-like domains 1), RHOB (ras homolog gene family, member B), LYM6 (lymphocyte antigen 96), FOXO1 (forkhead box O1), PNPLA2 (patatin-like phospholipase domain containing 2), TRH (thyrotropin-releasing hormone), GJC1 (gap junction protein, gamma 1, 45 kDa), SLC17A5 (solute carrier family 17 (anion/sugar transporter), member 5), FTO (fat mass and obesity associated), GJD2 (gap junction protein, delta 2, 36 kDa), PSRC1 (proline-serine-rich coiled-coil 1), CASP12 (caspase 12 (gene/pseudogene)), GPBAR1 (G protein-coupled bile acid receptor 1), PKX (PDGF receptor containing serine/threonine kinase), IL33 (interleukin 33), TRIB1 (trichobial homolog 1 (Drosophila)), PBX4 (pre-B-cell leukemia homeobox 4), NUPR1 (nuclear protein, transcriptional regulator), September 15 (15 kDa selenoprotein), CILP2 (cartilage intermediate layer protein 2), TERC (telomerase RNA component), GGT2 (gamma-glutamyltrans-
ferase 2), MT-CoI (mitochondrially encoded cytochrome c oxidase 1), and UOX (urate oxidase, pseudogene). In an additional embodiment, the chromosomal sequence may further be selected from Pont1 (paroxygenase 1), LDLR (LDL receptor), ApoE (Apolipoprotein E), Apo B-100 (Apolipoprotein B-100), ApoA (Apolipoprotein(a)), ApoA1 (Apolipoprotein A1), CBS (Cystathione B-synthase), Glycophosphatidylinositol-anchored protein (GPI), MTHFR (5,10-methylene tetrahydrofolate reductase (MTHFR), and combinations thereof. In one iteration, the chromosomal sequences and proteins encoded by chromosomal sequences involved in cardiovascular disease may be chosen from CaCN1C, Sod1, Ptn, Ppar(α), Apo E, Leptin, and combinations thereof. The text herein accordingly provides exemplary targets as to CRISPR or CRISPR-Cas systems or complexes.

Other Example Embodiments

[0104] FIG. 9 depicts a computing machine 2000 and a module 2050 in accordance with certain example embodiments. The computing machine 2000 may correspond to any of the various computers, servers, mobile devices, embedded systems, or computing systems presented herein. The module 2050 may comprise one or more hardware or software elements configured to facilitate the computing machine 2000 in performing the various methods and processing functions presented herein. The computing machine 2000 may include various internal or attached components such as a processor 2010, system bus 2020, system memory 2030, storage media 2040, input/output interface 2060, and a network interface 2070 for communicating with a network 2080.

[0105] The computing machine 2000 may be implemented as a conventional computer system, an embedded controller, a laptop, a server, a mobile device, a smartphone, a set-top box, a kiosk, a vehicular information system, one more processors associated with a television, a customized machine, any other hardware platform, or any combination or multiplicity thereof. The computing machine 2000 may be a distributed system configured to function using multiple computing machines interconnected via a data network or bus system.

[0106] The processor 2010 may be configured to execute code or instructions to perform the operations and functionality described herein, manage request flow and address mappings, and to perform calculations and generate commands. The processor 2010 may be a general purpose processor, a processor core, a multiprocessor, a reconfigurable processor, a microcontroller, a digital signal processor (“DSP”), an application specific integrated circuit (“ASIC”), a graphics processing unit ("GPU"), a field programmable gate array ("FPGA"), a reprogrammable logic device ("PLD"), a controller, a state machine, a logic device, a discrete hardware component, any other processing unit, or any combination or multiplicity thereof. The processor 2010 may be a single processing unit, a single processing unit, a single processing core, multiple processing cores, special purpose processing cores, co-processors, or any combination thereof. According to certain embodiments, the processor 2010 along with other components of the computing machine 2000 may be a virtualized computing machine executing within one or more other computing machines.

[0107] The system memory 2030 may include non-volatile memories such as read-only memory (“ROM”), programmable read-only memory (“PROM”), erasable programmable read-only memory (“EPROM”), flash memory, or any other device capable of storing program instructions or data with or without applied power. The system memory 2030 may also include volatile memories such as random access memory (“RAM”), static random access memory (“SRAM”), dynamic random access memory (“DRAM”), and synchronous dynamic random access memory (“SDRAM”). Other types of RAM also may be used to implement the system memory 2030. The system memory 2030 may be implemented using a single memory module or multiple memory modules. While the system memory 2030 is depicted as being part of the computing machine 2000, one skilled in the art will recognize that the system memory 2030 may be separate from the computing machine 2000 without departing from the scope of the subject technology. It should also be appreciated that the system memory 2030 may include, or operate in conjunction with, a non-volatile storage device such as the storage media 2040.

[0108] The storage media 2040 may include a hard disk, a floppy disk, a compact disk read only memory (“CD-ROM”), a digital versatile disk (“DVD”), a solid state drive (“SSD”), any magnetic storage device, any optical storage device, any electrical storage device, any semiconductor storage device, any physical-based storage device, any other data storage device, or any combination or multiplicity thereof. The storage media 2040 may store one or more operating systems, application programs, and program modules such as the module 2050, data, or any other information. The storage media 2040 may be part of, or connected to, the computing machine 2000. The storage media 2040 may also be part of one or more other computing machines that are in communication with the computing machine 2000 such as servers, database servers, cloud storage, network attached storage, and so forth.

[0109] The module 2050 may comprise one or more hardware or software elements configured to facilitate the computing machine 2000 with performing the various methods and processing functions presented herein. The module 2050 may include one or more sequences of instructions stored as software or firmware in association with the system memory 2030, the storage media 2040, or both. The storage media 2040 may therefore represent examples of code or a stored executable computer readable media on which instructions or code may be stored for execution by the processor 2010. Machine or computer readable media may generally refer to any medium or media used to provide instructions to the processor 2010. Such machine or computer readable media associated with the module 2050 may comprise a software product. It should be appreciated that a computer software product comprising the module 2050 may also be associated with one or more processes or methods for delivering the module 2050 to the computing machine 2000 via the network 2080, any signal-bearing medium, or any other communication or delivery technology. The module 2050 may also comprise hardware circuits or information for configuring hardware circuits such as an FPGA or other PLD.

[0110] The input/output (“I/O”) interface 2060 may be configured to couple to one or more external devices, to receive data from the one or more external devices, and to send data to the one or more external devices. Such external devices along with the various internal devices may also be known as peripheral devices. The I/O interface 2060 may include both electrical and physical I/O connections and may include coupling the various peripheral devices to the computing machine 2000 or the processor 2010. The I/O interface 2060 may be configured to communicate data, addresses, and control signals between the peripheral devices, the computing machine 2000, or the processor 2010. The I/O interface 2060 may be configured to implement any standard interface, such as small computer system interface (“SCSI”),
serial-attached SCSI ("SAS"), fiber channel, peripheral component interconnect ("PCI"), PCI express (PCIe), serial bus, parallel bus, advanced technology attached ("ATA"), serial ATA ("SATA"), universal serial bus ("USB"), Thunderbolt, FireWire, various video buses, and the like. The I/O interface 2060 may be configured to implement only one interface or bus technology. Alternatively, the I/O interface 2060 may be configured as part of, all of, or to operate in conjunction with, the system bus 2020. The I/O interface 2060 may include one or more buffers for buffering transmissions between one or more external devices, internal devices, the computing machine 2000, or the processor 2010.

[0111] The I/O interface 2060 may couple the computing machine 2000 to various input devices including mice, touch-screens, scanners, biometric readers, electronic digitizers, sensors, receivers, touchpads, trackballs, cameras, microphones, keyboards, any other pointing devices, or any combinations thereof. The I/O interface 2060 may couple the computing machine 2000 to various output devices including video displays, speakers, printers, projectors, tactile feedback devices, automation control, robotic components, actuators, motors, fans, solenoids, valves, pumps, transmitters, signal emitters, lights, and so forth.

[0112] The computing machine 2000 may operate in a networked environment using logical connections through the network interface 2070 to one or more other systems or computing machines across the network 2080. The network 2080 may include wide area networks (WAN), local area networks (LAN), intranets, the Internet, wireless access networks, wired networks, mobile networks, telephone networks, optical networks, or combinations thereof. The network 2080 may be packet switched, circuit switched, of any topology, and may use any communication protocol. Communication links within the network 2080 may involve various digital or an analog communication media such as fiber optic cables, free-space optics, waveguides, electrical conductors, wireless links, antennas, radio-frequency communications, and so forth.

[0113] The processor 2010 may be connected to the other elements of the computing machine 2000 or the various peripherals discussed herein through the system bus 2020. It should be appreciated that the system bus 2020 may be within the processor 2010, outside the processor 2010, or both. According to some embodiments, any of the processor 2010, the other elements of the computing machine 2000, or the various peripherals discussed herein may be integrated into a single device such as a system on chip ("SOC"), system on package ("SOP"), or ASIC device.

[0114] Embodiments may comprise a computer program that embodies the functions described and illustrated herein, wherein the computer program is implemented in a computer system that comprises instructions stored in a machine-readable medium and a processor that executes the instructions. However, it should be apparent that there could be many different ways of implementing embodiments in computer programming, and the embodiments should not be construed as limited to any one set of computer program instructions. Further, a skilled programmer would be able to write such a computer program to implement an embodiment of the disclosed embodiments based on the appended flow charts and associated description in the application text. Therefore, disclosure of a particular set of program code instructions is not considered necessary for an adequate understanding of how to make and use embodiments. Further, those skilled in the art will appreciate that one or more aspects of embodiments described herein may be performed by hardware, software, or a combination thereof, as may be embodied in one or more computing systems. Moreover, any reference to an act being performed by a computer should not be construed as being performed by a single computer as more than one computer may perform the act.

[0115] The example embodiments described herein can be used with computer hardware and software that perform the methods and processing functions described herein. The systems, methods, and procedures described herein can be embodied in a programmable computer, computer-executable software, or digital circuitry. The software can be stored on computer-readable media. For example, computer-readable media can include a floppy disk, RAM, ROM, hard disk, removable media, flash memory, memory stick, optical media, magneto-optical media, CD-ROM, etc. Digital circuitry can include integrated circuits, gate arrays, building block logic, field programmable gate arrays (FPGA), etc.

[0116] The example systems, methods, and acts described in the embodiments presented previously are illustrative, and, in alternative embodiments, certain acts can be performed in a different order, in parallel with one another, omitted entirely, and/or combined between different example embodiments, and/or certain additional acts can be performed, without departing from the scope and spirit of various embodiments. Accordingly, such alternative embodiments are included in the invention claimed herein.

[0117] Although specific embodiments have been described above in detail, the description is merely for purposes of illustration. It should be appreciated, therefore, that many aspects described above are not intended as required or essential elements unless explicitly stated otherwise. Modifications of, and equivalent components or acts corresponding to, the disclosed aspects of the example embodiments, in addition to those described above, can be made by a person of ordinary skill in the art, having the benefit of the present disclosure, without departing from the spirit and scope of embodiments defined in the following claims, the scope of which is to be accorded the broadest interpretation so as to encompass such modifications and equivalent structures.

[0118] Various modifications and variations of the described methods, pharmaceutical compositions, and kits of the invention will be apparent to those skilled in the art without departing from the scope and spirit of the invention. Although the invention has been described in connection with specific embodiments, it will be understood that it is capable of further modifications and that the invention as claimed should not be unduly limited to such specific embodiments. Indeed, various modifications of the described modes for carrying out the invention that are obvious to those skilled in the art are intended to be within the scope of the invention. This application is intended to cover any variations, uses, or adaptations of the invention following, in general, the principles of the invention and including such departures from the present disclosure that within known customary practice within the art to which the invention pertains and may be applied to the essential features herein before set forth
What is claimed is:

1. A method to identify novel CRISPR effector elements, comprising:
   classifying, by a processor, a CRISPR locus using an unsupervised machine learning applied to all or a subset of CRISPR locus elements as an initial set of inputs;
   identifying, using the processor, putative novel effector elements in the CRISPR locus, and optionally screening each identified putative novel effector element for one or more biological functions.

2. The method of claim 1, wherein the unsupervised machine learning comprises hierarchical clustering.

3. The method of claim 2, wherein hierarchical clustering comprises:
   generating, by the processor and prior to the classifying step, a preliminary set of CRISPR locus classes by separating a set of known CRISPR loci based, at least in part, on a sequence similarity and/or domain similarity between one or more protein elements of the CRISPR locus;
   generating, by the processor, a distance matrix data structure based, at least in part, on cumulative similarities between constituent proteins in each CRISPR locus class; and

wherein the putative CRISPR loci are classified by applying the hierarchical clustering model to the distance matrix.

4. The method of claim 3, wherein the CRISPR loci are separated based on a domain similarity.

5. The method of claim 4, wherein the domain similarity is determined using a hidden markov model.

6. The method of claim 3, wherein the cumulative similarities between constitute proteins is determined, at least in part, by calculating a Euclidean distance between each CRISPR locus.

7. The method of claim 1, wherein the unsupervised clustering model is an unsupervised neural network model.

8. The method of claim 7, wherein the unsupervised neural network model is trained using a set of known CRISPR loci that include spacer and repeat elements of the known CRISPR loci.

9. An isolated CRISPR effector, or a polynucleic acid sequence encoding said effector, identified by the method of claim 1.

10. Use of a CRISPR effector according to claim 9 for genome, transcriptome editing or modification in eukaryotic cells.

11. A computer program product, comprising:
   a non-transitory computer-executable storage device having computer readable program instructions embodied
thereon that when executed by a computer cause the computer to identify novel CRISP effector elements, the computer-executable program instruction comprising:

computer-executable program instructions to classify a CRISPR locus using an unsupervised machine learning applied to all or a subset of CRISPR locus elements as an initial set of inputs; and computer-executable program instructions to identify putative novel effector elements in the CRISPR locus, and optionally screening each identified putative novel effector element for one or more biological functions.

12. A system to identify novel CRISRP effector elements, the system comprising:

a processor communicatively coupled to the storage device; and

a processor communicatively coupled to the storage device, wherein the processor executes application code instructions that are stored in the storage device and that cause the system to:

classify a CRISPR locus using an unsupervised machine learning applied to all or a subset of CRISPR locus elements as an initial set of inputs; and

identify putative novel effector elements in the CRISPR locus, and optionally screening each identified putative novel effector element for one or more biological functions.

* * * * *