



(51) International Patent Classification:

G16B 35/20 (2019.01) C12Q 1/6813 (2018.01)
G16B 40/00 (2019.01) C12Q 1/6876 (2018.01)
G16B 20/00 (2019.01) C40B 30/06 (2006.01)
G16B 25/10 (2019.01) C12N 9/22 (2006.01)
C12N 15/113 (2010.01) A61K 48/00 (2006.01)
C12N 15/85 (2006.01)

(21) International Application Number:

PCT/US2024/033813

(22) International Filing Date:

13 June 2024 (13.06.2024)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

63/507,870 13 June 2023 (13.06.2023) US

(71) Applicants: NEW YORK GENOME CENTER, INC.

[US/US]; 101 Avenue Of The Americas, 7th Floor, New York, NY 10013 (US). NEW YORK UNIVERSITY [US/US]; 70 Washington Square South, New York, NY 10012 (US).

(72) Inventors; and

(71) Applicants: WESSELS, Hans-Hermann [DE/US]; 304 Clermont Avenue, Brooklyn, NY 11205 (US). STIRN, Andrew [US/US]; 110 Castilian Way, San Mateo, CA 94402 (US). KNOWLES, David, A. [GB/US]; 11123 Taylor Court, West Windsor, NJ 08550 (US).

(72) Inventor: SANJANA, Neville, E.; 4 Washington Square Village, Apartment 4H, New York, NY 10012 (US).

(74) Agent: SCHALLER, Colleen, M. et al.; Howson & Howson LLP, 325 Sentry Parkway East, Five Sentry East, Suite 160, Blue Bell, PA 19422 (US).

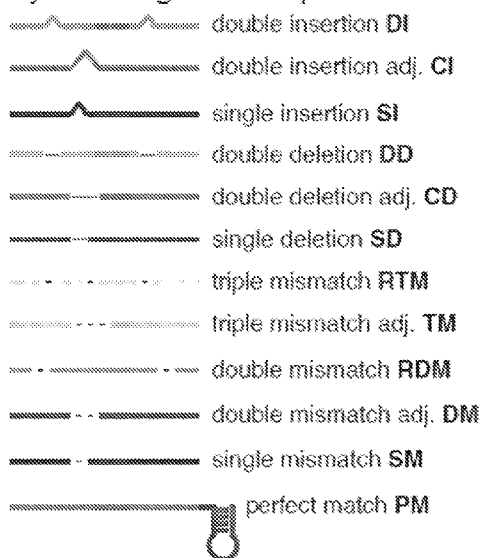
(81) Designated States (unless otherwise indicated, for every

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

(54) Title: A DEEP LEARNING FRAMEWORK TO PREDICT ON-TARGET AND OFF-TARGET ACTIVITY OF CRISPR GUIDE RNAS

FIG. 1A

Systematic guide RNA perturbations



(57) Abstract: A Class 2, Type VI clustered regularly interspaced short palindromic repeat (CRISPR) RNA (gRNA) which comprises a direct repeat (DR) stem loop sequence and a guide or spacer sequence, is provided. Also described are methods for generating, selecting, characterizing and optimizing a clustered regularly interspaced short palindromic repeats (CRISPR) RNA (gRNA) for use in the CRISPR-Cas1Sd system described herein. Also provided is a screening method to identify gRNA particularly suited for use with specified targets.



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

Published:

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

A DEEP LEARNING FRAMEWORK TO PREDICT ON-TARGET AND OFF-TARGET
ACTIVITY OF CRISPR GUIDE RNAS

STATEMENT REGARDING FEDERALLY SPONSORED RESEARCH OR
5 DEVELOPMENT

This invention was made with government support under grant numbers R00HG008171, DP2HG010099, and R01CA218668 awarded by the National Institutes of Health and grant number D18AP00053 awarded by the Defense Advanced Research Projects Agency. The government has certain rights in the invention.

10 BACKGROUND OF THE INVENTION

Programmable RNA-guided, RNA-targeting Type VI CRISPR Cas proteins (Cas13) enable direct manipulation of cellular RNAs with high precision compared to previous RNA-targeting technologies¹⁻⁵. A growing number of RNA-engineering technologies have been developed using nuclease active Cas13 or inactive dCas13 effector proteins⁶. These methods critically rely on the ability of Cas13 to distinguish between binding sites in target RNAs and closely related secondary (off-target) binding sites based on the complementarity between guide RNA (gRNA) sequence and bound RNA sequence. In general, the goal is to maximize on-target gRNA activity while minimizing off-target effects. While progress has been made on understanding Cas13 gRNA design rules for nuclease activation and on-target activity^{3,7-11}, relatively less is known about Cas13 off-target binding and activation. Our understanding is currently limited to Cas13a that has been used in diagnostics^{10,11} with only a few studies of off-target activity for Cas13d systems^{7,9}, which are more commonly used for *in vivo* perturbation. Several recent examples that use Cas13 effectors in pre-clinical model systems¹²⁻¹⁷, emphasizing the need for high precision for any potential human therapeutic.

Understanding the determinants of gRNA activity not only can improve target specificity but can also enable controlled/precise attenuation of gene dosage. Notably, biological systems often rely on relative gene dosage as opposed to a binary on-off state: Dose-dependent gene expression is crucial to maintain balanced stoichiometry between members of multiprotein complexes¹⁸ or during embryonic development¹⁹, and underlies X-chromosome inactivation²⁰.

Moreover, somatic copy number variation and subsequent gene amplification have been associated with cancer²¹ and a large number of human genetic diseases²².

Precise modulation of gene expression in mammalian systems can be achieved in multiple ways. For example, synthetic promoter sequences²³ or tetracycline-dependent promoter constructs²⁴ can be used to modulate gene expression. Similarly, insertion of *cis*-regulatory elements (i.e. miRNAs binding sites) in the 3'UTRs of endogenous genes renders them susceptible to the recruitment of the endogenous RNA surveillance and silencing machinery²⁵. Such approaches, however, require a considerable amount of engineering on an individual target basis. In contrast, programmable nuclease-null (dCas9) CRISPR systems provide a flexible and scalable alternative for systematic titration of gene expression²⁶. One caveat is that epigenetic effector domains commonly fused to dCas9 (e.g. KRAB domain) may act more in a switch-like fashion²⁷.

What is needed are more effective compositions for CRISPR-based therapies.

SUMMARY OF THE INVENTION

In one aspect, method of generating a set of guide RNAs (gRNAs) capable of driving a series of discrete expression levels of a target gene in a cell population using CRISPR interference (CRISPRi) or CRISPR activation (CRISPRa) is provided. The method includes (i) providing a first perfect match (PM) gRNA that comprises a spacer sequence that is 100% complementary to a target sequence within the gene; (ii) providing a second mismatch gRNA that targets the gene, wherein the spacer sequence of the second gRNA comprises one or more mismatches with the target DNA sequence such that the CRISPRi or CRISPRa activity on the gene obtained using the second gRNA is intermediate between that obtained using the first gRNA and that obtained using a scrambled gRNA providing no CRISPRi or CRISPRa activity on the gene; and (iii) providing a third indel mismatch gRNA that targets the gene, wherein the spacer sequence of the third sgRNA comprises one or more mismatches with the target DNA sequence such that the CRISPRi or CRISPRa activity on the gene obtained using the third gRNA is intermediate between that obtained using the second gRNA and that obtained using a scrambled gRNA providing no CRISPRi or CRISPRa activity on the gene; wherein the mismatches of the second and third gRNAs are selected according to the following rule:

(a) the CRISPRi or CRISPRa activity of the second gRNA is designed to be greater than that of the third gRNA based on the following positional relationships, wherein the positions correspond to the number of bases in the gRNA downstream to the DR sequence: $18 > 9 > 13 > 5 > 22$.

In certain embodiments, the method includes providing one or more additional mismatch

- 5 gRNAs, wherein each of the one or more additional gRNAs provide CRISPRi or CRISPRa activity on the gene that is intermediate between that obtained using the third gRNA and that obtained using a scrambled gRNA providing no CRISPRi or CRISPRa activity on the gene, and wherein the mismatches with the target sequence of each of the one or more additional gRNAs are selected according to rule (a).

10 In one aspect, a computational model termed TIGER is provided. The computational model comprises one or more of the following layers:

- i) 2D convolution with 32 4 x 4 filters and ReLU activations
- ii) 2D convolution with 32 4 x 4 filters and ReLU activations
- iii) 2D max pooling with a 1 x 2 pool size
- 15 iv) Flatten to vector
- v) Dropout with 0.25 dropout rate
- vi) Concatenation with a vector of non-sequence features
- vii) Dense layer with 128 sigmoid outputs
- viii) Dropout with 0.1 dropout rate
- 20 ix) Dense layer with 32 sigmoid outputs
- x) Dropout with 0.1 dropout rate
- xi) Linear output layer with scalar output;

wherein the non-sequence features comprise one or more of the following:

- a) target accessibility
- 25 b) hybridization MFE
- c) guide MFE
- d) target location within transcript
- e) junction proximity
- f) guide secondary structure

30 wherein the computational model predicts normalized LFC values using a log-hyperbolic-cosine loss.

In another aspect, a set of single guide RNAs (sgRNAs) for obtaining a series of discrete expression levels of a target gene using CRISPRi or CRISPRa is provided. The set comprises (i) a first perfect match (PM) gRNA that comprises a spacer sequence that is 100% complementary to a target sequence within the gene; (ii) a second mismatch gRNA that targets the gene, wherein the spacer sequence of the second gRNA comprises one or more mismatches with the target DNA sequence such that the CRISPRi or CRISPRa activity on the gene obtained using the second gRNA is intermediate between that obtained using the first gRNA and that obtained using a scrambled gRNA providing no CRISPRi or CRISPRa activity on the gene; and (iii) a third indel mismatch gRNA that targets the gene, wherein the space sequence of the third sgRNA comprises one or more mismatches with the target DNA sequence such that the CRISPRi or CRISPRa activity on the gene obtained using the third gRNA is intermediate between that obtained using the second gRNA and that obtained using a scrambled gRNA providing no CRISPRi or CRISPRa activity on the gene; wherein the mismatches of the second and third gRNAs are selected according to the following rule: (a) the CRISPRi or CRISPRa activity of the second gRNA is designed to be greater than that of the third gRNA based on the following positional relationships, wherein the positions correspond to the number of bases in the gRNA downstream to the DR sequence: $18 > 9 > 13 > 5 > 22$.

In another aspect, a method of selecting a gRNA for use in a Cas-based genome editing system capable of introducing a discreet level of expression of a target gene in a cell population is provided. The method includes transducing mammalian cells with a library of gRNAs, the library comprising a plurality of different gRNAs for one or more genes, the plurality of gRNAs comprising for each gene: i) perfect-match (PM) gRNAs that are 100% complementary to a target sequence within the gene; ii) mismatch gRNAs that contain 1, 2, or 3 nucleotide mismatches to the target sequence within the gene; iii) indel gRNAs that contain 1 or 2 nucleotide insertions or deletions to the target sequence within the gene; determining the activity of each gRNA in the library; determining the relative activity of all mismatch gRNAs and indel gRNAs relative to their cognate PM gRNAs; and selecting the gRNA based on the discreet level of expression desired.

In yet another aspect, a method for precise modulation of target RNA is provided. The method comprises providing a gRNA that allows a desired level of effector activity of a Cas13d

enzyme when the gRNA and Cas13d enzyme are contacted with the target RNA to which the gRNA binds, wherein the gRNA is identified by a method as described herein.

In yet another aspect, a method of predicting the efficacy of a subject gRNA for use in a Cas-based genome editing system is provided. The method includes analyzing the subject gRNA sequence with a computational model to predict efficacy of the subject gRNA, the computational model comprising one or more of the following layers:

- i) 2D convolution with 32 4 x 4 filters and ReLU activations
- ii) 2D convolution with 32 4 x 4 filters and ReLU activations
- iii) 2D max pooling with a 1 x 2 pool size
- 10 iv) Flatten to vector
- v) Dropout with 0.25 dropout rate
- vi) Concatenation with a vector of non-sequence features
- vii) Dense layer with 128 sigmoid outputs
- viii) Dropout with 0.1 dropout rate
- 15 ix) Dense layer with 32 sigmoid outputs
- x) Dropout with 0.1 dropout rate
- xi) Linear output layer with scalar output;

wherein the non-sequence features comprise one or more of the following:

- a) target accessibility
- 20 b) hybridization MFE
- c) guide MFE
- d) target location within transcript
- e) junction proximity
- f) guide secondary structure

- 25 wherein the computational model predicts normalized LFC values using a log-hyperbolic-cosine loss.

In yet another aspect, a system for selecting gRNA for titrating expression of a target gene in a target cell from a multiplicity of candidate guides is provided. The system includes a memory; and a processor coupled with the memory and configured to: analyze the subject gRNA sequence with a computational model to predict efficacy of the subject gRNA, the computational model comprising one or more of the following layers:

- i) 2D convolution with 32 4 x 4 filters and ReLU activations
 - ii) 2D convolution with 32 4 x 4 filters and ReLU activations
 - iii) 2D max pooling with a 1 x 2 pool size
 - iv) Flatten to vector
 - 5 v) Dropout with 0.25 dropout rate
 - vi) Concatenation with a vector of non-sequence features
 - vii) Dense layer with 128 sigmoid outputs
 - viii) Dropout with 0.1 dropout rate
 - ix) Dense layer with 32 sigmoid outputs
 - 10 x) Dropout with 0.1 dropout rate
 - xi) Linear output layer with scalar output;
- wherein the non-sequence features comprise one or more of the following:
- a) target accessibility
 - b) hybridization MFE
 - 15 c) guide MFE
 - d) target location within transcript
 - e) junction proximity
 - f) guide secondary structure
- wherein the computational model predicts normalized LFC values using a log-hyperbolic-cosine
- 20 loss.

Still other aspects and advantages of these compositions and methods are described further in the following detailed description of the preferred embodiments thereof.

BRIEF DESCRIPTION OF THE DRAWINGS

- 25 FIG. 1A – FIG. 1H show pooled CRISPR-Cas13 essentiality screen assaying Cas13d guide RNA (gRNA) efficacy. (FIG. 1A) Design of pooled CRISPR-Cas13d screen for mapping gRNA variants with mismatch and insertion-deletion changes to perfect-match (PM) gRNAs. (FIG. 1B) Composition of gRNA library containing 120,000 perfectly matching and mismatched gRNA sequences targeting the coding region of essential genes. (FIG. 1C) Abundance of
- 30 individual gRNAs was measured in TetO-*Rfx*Cas13d-HEK293FT cells over 30 days ($n = 3$ independent transduction replicates). (FIG. 1D) Pearson correlation of gRNA abundance as $\log_2$

fold-change ($\log_2FC$) at Day 15 and Day relative to the Day 0 input representation showing perfect match (PM) gRNAs as a mean of three replicates ($n = 13,782$). (FIG. 1E) Fraction of active gRNAs ($\log_2FC < -0.5$) for PM gRNAs separated by RF_{on} quartile predictions. (FIG. 1F) Fraction of active ($\log_2FC < -0.5$) predicted Quartile 4 (Q4) PM gRNAs for all 16 essential gene targets. (FIG. 1G) Relationship between median distance between neighboring PM gRNAs and autocorrelation of $\log_2FC$ at lag = 1 ($n = 16$ target gene transcripts). (FIG. 1H) Distribution of PM gRNAs along the coding region of the 16 target gene transcripts and their $\log_2FC$ enrichments. Negative $\log_2FC$ values indicate better transcript knock-down. ρ = Autocorrelation of $\log_2FC$ with lag = 1.

FIG. 2A – FIG. 2G show large-scale mapping of Cas13d guide RNA (gRNA) mismatch activity. (FIG. 2A) Empirical cumulative distribution of gRNA depletion for all gRNAs by introduced mutation type. Vertical line indicates cutoff for active gRNAs ($\log_2FC < -0.5$). (FIG. 2B) Fraction of active gRNAs for gRNAs as shown in FIG. 1A. (FIG. 2C-FIG. 2I) Relative targeting activity (fraction of parental PM gRNA $\log_2FC$) of gRNAs for all PM gRNA derivatives with mismatches or indels at the indicated position relative to their cognate perfect match gRNAs ($n = 388$ reference PM gRNA with $\log_2FC < -0.5$. Each cell indicates the mean of gRNAs). (FIG. 2C (*left*)) Heatmap depicting all mismatch types. (FIG. 2C (*right*)) Boxplot highlighting 820 the full distribution of relative gRNA activity for single mismatch gRNAs. (FIG. 2D (*left*)) Heatmap depicting all deletion types. (FIG. 2D (*right*)) Boxplot highlighting the single deletion gRNAs. (FIG. 2E (*left*)) Heatmap depicting all insertion types. (FIG. 2E (*right*)) Boxplot highlighting the single insertion gRNAs. (FIG. 2F) Detailed representation of relative activity for single mismatch (SM) gRNA separated by reference guide nucleotide (*bottom*) or substitution identity (*top*) for each mismatch position. (FIG. 2G) Relative targeting efficacy for gRNAs containing 0 or 1 G-U wobble base pairs compared to unpaired mismatches for all mismatched gRNA types ($n = 388$ reference PM gRNA with $\log_2FC < -0.5$). All gRNA abbreviations are defined in FIG. 1A.

FIG. 3A – FIG. 3I show a deep learning model to predict optimal Cas13d guide RNAs. (FIG. 3A) TIGER combines one-hot encoded guide and target sequences for sequence input, following an *AlexNet* architecture but allowing for non-sequence features as inputs to the first dense layer. (FIG. 3B) Correlation of predictions with additional sequence context (5' only, 3' only, combined 5' and 3') to the 23nt gRNA target site (10-fold cross-validation, target site-level)

using a sequence only model. H_0 denotes the best performing condition and all differences between other conditions and H_0 are significant ($p < 0.05$, Steiger's test⁶⁷). (FIG. 3C) The effect of including different feature groups (individually and cumulatively) on the correlation of predictions aggregated from the held-out target sites ($n = 10$ random folds). We present feature groups in descending order of increased correlation (individually). (FIG. 3D) Receiver-operator characteristic (ROC) curve and other performance metrics for predictions aggregated from the held-out genes ($n = 16$) of the survival screen of essential genes. (FIG. 3E) Receiver-operator characteristic (ROC) curve and other performance metrics for all gRNAs from a previously published screen using flow cytometry of cell surface proteins (Ref.⁷). FIG. 3D and FIG. 3E employ a Steiger's test⁶⁷ for Pearson and Spearman comparisons, DeLong's test^{68,69} for AUROC comparisons and a bootstrapped Kolmogorov-Smirnov test⁷⁰ for AUPRC comparisons. (FIG. 3F) Design of pooled CRISPR-Cas13d screen targeting 5,166 genes with 8 high efficacy gRNAs from TIGER_{combined} predictions. (FIG. 3G) The top 10 most depleted genes show consistent depletion in each cell 850 line ($n = 8$ gRNAs per target gene). Dotted lines (*black*) indicate the 1st and 99th percentiles for non-targeting gRNA distribution. (FIG. 3H) Fraction of active gRNAs (more depleted than the 99th percentile of the non-targeting gRNAs) as a function of gene depletion. Grey lines indicate 99th percentile of non-targeting gRNA distribution for HAP1 (*solid*) and HEK293FT (*dashed*) cells. (FIG. 3I) Receiver-operator characteristic (ROC) curve for DepMap essential genes for screens depicted in FIG. 3F – FIG. 3H ($n = 1,082$ essential genes, $n = 458$ non-essential genes). Number in parenthesis next to each cell line name indicates AUROC.

FIG. 4A – FIG. 4J show training TIGER using gRNAs with mismatches enables prediction of off-target activity and transcript modulation using gRNAs with single mismatches. (FIG. 4A) The correlation between observed and TIGER-predicted gRNA abundance by gRNA design type. (FIG. 4B) The correlation between change in observed and TIGER-predicted gRNA abundance by gRNA design type. The change in gRNA abundance is defined as the difference in fold-change for a particular gRNA with mismatches and its cognate perfect match (PM) gRNA. (FIG. 4C) Receiver-operator characteristic (ROC) curves for each gRNA design type from 10-fold target site cross-validation. (FIG. 4D) Averaged correlation (Pearson and Spearman) and aggregated areas under the ROC and precision-recall curves for each gRNA design type from 10-fold target-site cross-validation. (FIG. 4E) A framework for using gRNA with single-mismatches

to modulate Cas13 targeting activity. (FIG. 4F) Correlation of predicted and observed relative activity ratio for all 23 SM gRNAs per highly active ($\log_2[\text{FC}] < -1$) target sites ($n = 324$ target sites). (FIG. 4G) Confusion matrix for efficiency ratios between a gRNA with a single-mismatch to the intended target RNA binned by quintiles for all active target sites ($FDR < 0.01$) ($n = 393$ target sites with $n = 9,032$ SM gRNA variants). Each column is normalized. (FIG. 4H) Design of pooled CRISPR-Cas13d screen for TIGER_{combined} gRNA predictions targeting 1,082 common essential genes with 4 high efficacy PM gRNAs, and 10 SM gRNAs with varying relative activity. (FIG. 4I) Correlation of predicted and observed relative activity ratio 881 for all 10 SM gRNAs per highly active PM gRNA ($n = 3,161$ gRNAs with $\log_2[\text{FC}] < -1$). (FIG. 4J) Confusion matrix for efficiency ratios between a gRNA with a single-mismatch to the intended target RNA binned by quintiles ($n = 30,582$ SM gRNA variants). Each column is normalized.

FIG. 5 provides a schematic of the TIGER model architecture.

FIG. 6 provides a graphical summary of various validation folding strategies.

FIG. 7 provides a schematic of the numbering strategy for the gRNA (bottom) and target (top).

FIGs. 8A-8C show two non-sequence features used in the TIGER model. FIG. 8A shows the correlation of target site accessibility and gRNA efficacy. FIG. 8B shows the correlation of target site hybridization and gRNA efficacy. FIG. 8C shows the schematic of how these features are assessed with the top referring to target site accessibility, and the bottom referring to target site hybridization.

FIGs. 9A-9N are representative figures showing the data for 10,000 perfect match (PM) gRNA, and the assessment of their non-sequence features, as used in the TIGER model.

FIGs. 10A-10B demonstrate that combinatorial knockdown of synthetic lethal pair HDAC1 - HDAC2 confirms lack of collateral RNA cleavage after low copy-number transduction of RfxCas13d. A combinatorial titration pooled transcriptomic screen targeting a synthetic lethal gene pair (HDAC1 and HDAC2). Designed position-specific mismatches in the gRNA arrays titrate knockdown activity of RfxCas13d. After gRNA library transduction into human A375 cells, depletion of single or dual gene perturbations is quantified by comparing gRNA counts between day 14 and day 0 (FIG. 10A). FIG. 10B - Fold-change of gRNAs targeting HDAC1 or HDAC2 individually or together at 100% predicted gRNA activity 14 days after RfxCas13d induction ($n = 120$ gRNA arrays with 40 each for HDAC1 only, HDAC2 only and both HDAC1

and HDAC2 targeting). The distribution of the central 95% of non-targeting gRNAs is shown in gray (n = 492 non-targeting perturbations). Statistical significance determined by two-sided Mann-Whitney U test.

5 DETAILED DESCRIPTION

Described herein is a method of screening and selecting a clustered regularly interspaced short palindromic repeats (CRISPR) guide RNA (gRNA) suitable for forming a complex with a CRISPR-associated protein 13d (Cas13d) or a variant thereof and directing the complex to a target RNA. In certain embodiments, the method includes assessing the on-target and/or off-
10 target activity of the gRNA.

The inventors generated the largest Cas13d dataset to date measuring the activity of ~250,000 gRNAs across a total of 14 individual screens, and performed a comprehensive assessment of Cas13d gRNA on-target and off-target activity. Specifically, they sought to characterize perfect match (PM) gRNA activity determinants and gRNAs permutations across a
15 large set of nucleotide mismatches and indels relative to their cognate target sites. It was determined that a gRNA's ability to trigger Cas13d nuclease activity depends on the permutation position within the gRNA, the nucleotide identity, as well as target site context. Prior studies have not characterized certain gRNA permutations such as insertions or deletions. This analysis shows that mismatches are generally better tolerated compared to more disruptive insertions and
20 deletions in gRNA or target RNA sequences. Utilizing this unique dataset, the TIGER CNN model was trained for on-target activity and off-target activity. TIGER has best-in-class performance for Cas13d on-target activity, compared to existing Cas13d on-target models, including those with larger training sets. Of relevance for understanding impacts across the transcriptome, the TIGER model is the first compelling attempt to understand and model Cas13d
25 off-target binding and nuclease activation. And finally, the TIGER platform is applied to develop a novel approach for precise and massively parallel interrogation of gene dosage.

Novel CRISPR technologies hold great promise for a new generation of therapeutic agents. Among these, RNA-targeting CRISPR proteins have recently been shown to provide therapeutic value in disease models¹²⁻¹⁷. High precision is key to the safety of therapeutic RNA-
30 targeting CRISPR agents. TIGER predictions will enable ranking and ultimately avoidance of undesired off-target binding sites and nuclease activation, and further spur development of RNA-

targeting therapeutics. The ability to distinguish between closely related target sites may for example enable targeting of allelic variants and other nearly undruggable targets like fusion gene products⁴⁷.

Further, the TIGER model can be used for precise modulation target RNA knockdown at scale. Specifically, RNA-targeting CRISPR perturbations can be used to systematically study the effect of gene dosage at the RNA levels. This platform fundamentally extends on previous microRNA-based platforms²⁵ that, on one hand, lack scalability due to laborious target site engineering and, on the other hand, lack target-specificity if engineered microRNAs are provided exogenously due to their short target site recognition sequence. In addition, tuning of gene expression at the RNA level may be beneficial compared to modulation at the DNA level, as gene expression initiation is inherently stochastic⁴⁸ and biological systems have evolved in a way to fine-tune gene expression post-transcriptionally^{49,50}. Other DNA- targeting (e.g., dCas9-based) CRISPR approaches have been proposed for gene expression modulation²⁶. However, it is unclear if epigenetic effector domains (e.g., KRAB) fused dCas9 proteins are well suited as they may act more in a binary on-off fashion^{27,51}, and may lack precision for closely spaced genes due to spreading of chromatin modifications and DNA methylation⁵². Here, we provide precise measurements for the titration of 1,082 essential genes across thousands of target sites confirming the TIGER model's prediction accuracy.

Provided herein are compositions and methods for generating discrete, intermediate expression levels of any gene of interest when using CRISPRi or CRISPRa. In particular, the present compositions and methods involve the introduction of one or more mismatches or indels into the targeting sequence of gRNAs so as to achieve a level of CRISPRi or CRISPRa activity that is, e.g., intermediate between that obtained with a gRNA sharing 100% homology (perfect match, PM) with a target RNA sequence and/or an unmodified constant region and that obtained with a non-specific gRNA providing no CRISPRi or CRISPRa activity on the gene in question. Further, rules are provided by which the specific effects of a given mismatch or mutation on CRISPRi or CRISPRa activity can be determined, allowing the design of sets of gRNAs targeting a given gene and providing a series of discrete levels of expression of the gene. As described herein, such sets can be combined to form libraries targeting multiple genes, including large libraries targeting thousands of genes in the genome.

The present methods extend previously developed CRISPR-based transcriptional repression (CRISPR interference, or CRISPRi) and overexpression (CRISPR activation, or CRISPRa), in which dCas13 is fused to a transcriptional repressor or activator, respectively, and is targeted to endogenous genes via a single guide RNA (sgRNA). Jost, M., et al (March 2020).

5 Titrating gene expression using libraries of systematically attenuated CRISPR guide RNAs Nature Biotechnology, 38(3):355–364; WO 2021/021677A9 both of which are incorporated herein by reference. The Cas13-gRNA complex binds to RNA (e.g., mRNA) via base-pairing between the gRNA and RNA, i.e., the targeting sequence of the sgRNA and the target RNA sequence on the e.g., mRNA, and the fused transcriptional repressor or activator leads to
10 downregulation or upregulation of the gene, respectively. The present disclosure provides methods to control the activity of Cas13 at a given RNA target site, e.g., by introducing mismatches into the gRNA (e.g., within the targeting sequence of the gRNA) or by introducing mutations into the gRNA constant region.

The present disclosure also provides rules, factors, and parameters to determine how a
15 given mismatch in a gRNA targeting sequence affects the extent of repression or activation of a target gene by CRISPRi or CRISPRa, allowing the design of sets of mismatched gRNAs against the gene to allow its downregulation or upregulation to varying extents. In some embodiments, the information on the expression level of the target gene is encoded in the gRNA sequence or in the vector encoding the gRNA, and can therefore be read out by, e.g., deep sequencing and
20 matched to a resulting phenotype. In such embodiments, experiments involving systematically mismatched gRNAs can be conducted in a single pooled experiment, reducing experimental variation and enhancing reproducibility. It will be appreciated that any of the herein-described methods and compositions can be applied to both gene downregulation (using CRISPRi) and overexpression (using CRISPRa).

25 A. Guide RNA (gRNA)

The CRISPR system described herein has two main components: a Cas13 protein and the gRNA. The Cas protein has nuclease activity, and the gRNA leads the Cas protein to the specific target. A gRNA for Cas13 is a small RNA with a conserved sequence forming a stem-loop structure (also known as direct repeat or ‘DR’) and a spacer sequence. The spacer is sometimes
30 referred to as a guide. Cas13 binds to the DR, and the spacer sequence is complementary to the target RNA. The compositions and methods provided herein utilize spacer sequences that are a

perfect match (PM), i.e., 100% complementary to the target sequence, in addition to spacer sequences having one or more mismatches (mutations as compared to the PM) or indels (one or more nucleotide insertions or deletions as compared to the PM).

In one embodiment, the spacer is about 20 nucleotides (nt) to about 33 nt. In a further embodiment, the spacer is about 20 nt, about 21 nt, about 22 nt, about 23 nt, about 24 nt, about 25, nt, about 26, nt, about 27 nt, about 28 nt, about 29 nt, about 30 nt, about 31 nt, about 32 nt, or about 33nt. In one embodiment, the spacer is about 27 nt. In one embodiment, the spacer is about 23 nt. Existence of a sequence in a target RNA similar to a protospacer or protospacer adjacent motif (PAM) was not found in the CRISPR-Cas13d system.

As used herein, nucleotide residues in a gRNA or a portion of it (for example, a spacer, or a stem loop), unless specified, are numbered as illustrated in Figure 7. In one embodiment, the numbering is based on a numbering from 5' end of the gRNA to 3' end recognizing the guide match start as nt 1. The guide match start is the first nucleotide residue (nt) from the 5' end of the gRNA which is capable of matching to a nt of a target RNA. The nt numbering at the 3' side of the guide match start is a positive integer positively correlated to its distance to the guide match start, while the nt numbering at the 5' side of the guide match start is a negative integer whose absolute value is positively correlated to its distance to the guide match start. One exception is the last nt of the DR stem loop contiguously proceeding the first nt of the guide is numbered as nt 0. As one of skill in the art would appreciate, whenever an order of a nt is implying, for example, via using the terms "first" "last" "proceeding" or similar, the order is counted from the 5' end to the 3' end. Further, with respect to a target RNA or a portion of it (for example, a hybridization region of the target RNA), the nt numbering is from 5' end of the target RNA to its 3' end recognizing the nt which is capable of matching to the guide match start as nt 0. The nt numbering at the 3' side of the nt matching to the guide match start is a positive integer positively correlated to its distance to the guide match start, while the nt numbering at the 5' side of the nt matching to the guide match start is a negative integer whose absolute value is positively correlated to its distance to the guide match start. An illustration can be found in Figure 7. Based on the definitions herein, one of skill in the art would understand that nt x of a gRNA matches to nt -(x-1) of a corresponding target RNA, wherein x is a positive integer. In one embodiment, for guide RNA features nucleotide 1 defines the guide start site (GSS) being the most 5' guide RNA base matching the target RNA. Nucleotide 2 relative to GSS is the

subsequent base (moving in the 5' to 3' direction) in the guide RNA and so on. For target RNA features, we denote the target nucleotide opposite to the GSS as nucleotide 0. Moving in 5' to 3' direction target RNA nucleotide -1 is upstream to the GSS and pairs with guide nucleotide 2, while target RNA nucleotide +1 is downstream of the target site and so on. In a further embodiment, a range of nt is also illustrated as nucleotide position p over the distance d to the position p+d with its cognate sequence. In another embodiment, a nt range is noted as (nt x: y) indicating nt x to nt y, wherein x and y is an integer which may be positive, negative or zero.

As used herein, the term “match” or any other grammatical variation thereof, when referring to two nucleotide residues in one or two nucleic acid molecule(s), means a nt residue (which may be an RNA or a DNA) is complementary to the other (which may also be a RNA or a DNA). As it is known to one of skill in the art, guanine is the complementary base of cytosine, and adenine is the complementary base of thymine in DNA and of uracil in RNA. The nucleotide residues matching with each other, for example, in a secondary structure (observed or predicted) of the nucleic acid molecule(s) are a pair of nucleotide residues (nt), or paired nt. In a further embodiment, if no matching nt in the nucleic acid molecule(s), the nt is then referred to as an unpaired nt or a mismatch.

Hybridization is the process of complementary base pairs (nucleotide residues) binding to form a double helix. The term “hybridization” or any other grammatical variation hereof refers to at least two regions from one single nucleic acid molecule or of two or more nucleic acid molecules which comprises at least one nucleotide residue in one region matches a nucleotide residue in another region. In one embodiment, each of the nt in the first region matches to a nt in the second region. In a further embodiment, each of the nt in the first region matches to each of the nt in the second region. In another embodiment, one or more mismatch(es) may be found between two regions, for example one mismatch, two mismatches, two consecutive mismatches, two nonconsecutive mismatches, three or more mismatches (consecutive or nonconsecutive).

The gRNA may target any region of the target RNA. However, as Cas13 proteins only bind and cut ssRNA is essential that your gRNAs for Cas13 target the loops or single-stranded regions of the RNA, and not dsRNA portions generated due to secondary structure.

Nucleic acid secondary structure is the base pairing interactions within a single nucleic acid polymer or between two polymers. It can be represented as a list of bases which are paired in a nucleic acid molecule. Nucleic acid secondary structure can be determined from atomic

coordinates (tertiary structure) obtained by X-ray crystallography, often deposited in the Protein Data Bank. Current methods include 3DNA/DSSR and MC-annotatc. Methods for nucleic acid secondary structure prediction are also available, for example those relying on a nearest neighbor thermodynamic model. A common method to determine the most probable structures given a sequence of nucleotides makes use of a dynamic programming algorithm that seeks to find structures with low free energy. The lower the free energy is, the more stable the secondary structure is. Thus, minimum free energy (MFE) has been used in characterizing a secondary structure. In one embodiment, minimum free energies (MFEs) of a gRNA secondary structure were derived using RNAfold [--gquad] on the full-length gRNA sequence. Also, a MFE of a secondary structure form by two regions hybridizing to each other (for example a target RNA and its corresponding guide) is referred to as a hybridization MFE. In one embodiment, in order to calculate the hybridization MFE between regions of a target RNA and its corresponding guide, Target RNA unpaired probability (accessibility) was calculated using RNAplfold [-L 40 -W 80 -u 50] as described previously. RNA-RNA-hybridization was calculated using RNAhybrid [-s -c] using the di-nucleotide frequency derived from the target sequence. We calculated the RNA-hybridization minimum free energy for each spacer RNA nucleotide position p over the distance d to the position $p + d$ with its cognate target sequence.

Further, G-quadruplex is a secondary structure formed in nucleic acid by sequences that are rich in guanine. They are helical structures containing guanine tetrads that can form from one, two or four strands. Four guanine bases can associate through Hoogsteen hydrogen bonding to form a square planar structure called a guanine tetrad (G-tetrad or G-quartet), and two or more guanine tetrads (from G-tracts, continuous runs of guanine) can stack on top of each other to form a G-quadruplex. G-quadruplex structures can be computationally predicted from DNA or RNA sequence motifs or other method available publicly or commercially. Generally, a simple pattern match is used for searching for possible intrastrand quadruplex forming sequences: $d(G3+N1-7G3+N1-7G3+N1-7G3+)$, where N is any nucleotide base (including guanine). See, for example, rank-Kamenetskii MD, Mirkin SM (1995). "Triplex DNA structures". Annual Review of Biochemistry. 64 (9): 65–95. In one embodiment, RNAfold may be used to determine a presence or absence of a G-quadruplex.

As used herein a "target RNA" refers to an RNA molecule or a nucleic acid molecule to which a spacer sequence is designed to target, e.g., have complementarity, where hybridization

between a target RNA and a guide promotes the formation of a CRISPR-Cas13d complex. In one embodiment, the target RNA comprises at least 20 nt (or at least 23 nt, or at least 87 nt, or at least 100 nt) RNA residues or a modification thereof. In a further embodiment, the target RNA comprises at least 20 nt contiguous RNA residues or a modification thereof. The region of a target RNA which is capable of hybridizing to a spacer of a gRNA is referred to herein as a potential hybridization region. Such target RNA, a hybridization region therein, a gRNA which the hybridization region of the target RNA may hybridize to, and a spacer of the gRNA correspond to each other.

As used herein, the term “seed” region or any other grammatical variation thereof means a critical region of the target sequence of Class 2, Type VI enzymes (e.g., Cas13d) that must be strictly complementary to the CRISPR RNA guide to ensure knock-down efficacy. Mismatches between the target and CRISPR RNA guide sequence can contribute to off-target activity. In one embodiment, the critical Cas13d seed region is defined as the region located between spacer RNA nucleotides 15 to 21. In another embodiment, the seed region is defined as the region located between spacer RNA nucleotides 15 to 21, with its center at nucleotide 18 relative to the guide RNA 5' end. Within the seed region, single mismatches lead to diminished guide enrichment, while mismatches outside the seed region were better tolerated (see Figure 2F).

The gRNA also includes a short hairpin region, also referred to as a “direct repeat” (DR) sequence, via which the Cas13 protein complexes with the guide RNA. Like all CRISPR-Cas effectors described to date, Cas13 enzymes each recognize a direct repeat (DR) sequence containing a conserved stem loop structure within their cognate crRNA. However, between Cas13 subtypes, the DR sequence motifs, RNA fold, and DR position relative to the spacer sequence are each distinct. For Cas13a and Cas13d, the DR is located on the 5' end while the Cas13b DR is 3' of the spacer sequence. Thus, in one embodiment, the DR is 5' to the spacer sequence. In another embodiment, the DR is 3' to the spacer sequence. See, e.g., Cheng et al, Structural Basis for the RNA-Guided Ribonuclease Activity of CRISPR-Cas13d, Cell. 2018 Sep 20;175(1):212-223.e17, which is incorporated herein by reference. See also, WO 2022/063314, which describes DRs and modified DRs that are useful herein. This document is incorporated herein by reference.

In some embodiments, gRNAs are provided with one or more mismatches in the spacer sequence of the gRNA in order to generate intermediate levels of CRISPRi or CRISPRa activity.

Complementary nucleotides are, generally, A and T (or A and U), and G and C. In some embodiments, sets of gRNAs are provided with different mismatches to generate a series of different expression levels of a target gene. A set typically includes at least one gRNA in which the spacer sequence is a PM, as well as one or more gRNAs that comprise one or more mismatches within the spacer sequence. Mismatches in the targeting sequence selected according to the present methods reduce the CRISPRi or CRISPRa activity to an intermediate level between that of a PM gRNA with 100% homology to the target RNA (e.g., providing 100% CRISPRi or CRISPRa activity) and that of a non-targeting gRNA that does not target the target RNA (i.e., with a targeting sequence comprising insufficient homology to the target RNA sequence to promote Cas13 binding and consequent CRISPRi or CRISPRa activity). It will be appreciated that a given gene can be targeted using a single set of gRNAs that recognize a single target sequence within the gene, or with multiple sets that each target a different sequence within the target RNA.

In some embodiments, a gRNA comprising one or more mismatches in the spacer sequence provides about 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 80%, 85%, 90%, or 95% CRISPRi or CRISPRa activity, wherein 100% CRISPRi or CRISPRa activity corresponds to the activity in the presence of a PM sgRNA, and wherein 0% CRISPRi or CRISPRa activity corresponds to the activity in the presence of a non-targeting sgRNA with no, or only insignificant amounts of, homology to the target sequence.

B. *Method of Designing, Generating and Ranking gRNA(s)*

In one aspect, provided is a method for generating and selecting a gRNA which is capable of forming a complex with a CRISPR-associated protein 13d (Cas13d) or a variant thereof, and driving a discreet expression level of a target gene in a cell population using CRISPR interference (CRISPRi). The method comprises the following steps:

- (1) designating a potential hybridization region in the target RNA;
- (2) designing a subject gRNA which is capable of hybridizing to the hybridization region, wherein the gRNA is a perfect match gRNA, a mismatch gRNA, or an indel gRNA;
- (3) analyzing the subject gRNA with the TIGER model to predict efficacy of the gRNA.

In certain embodiments, the ranking is performed using deep learning, such as using the TIGER model described herein (see, e.g., Examples 4-5). The inventors have trained a convolutional neural network termed “TIGER” (Targeted Inhibition of Gene Expression via gRNA design) to predict efficacy from CRISPR guide sequence and context. TIGER

5 outperforms existing models at predicting on- and off-target activity on our dataset and published datasets. TIGER scoring combined with specific mismatches yields the first general framework to modulate transcript expression, enabling use of RNA-targeting CRISPRs to precisely control gene dosage.

Our model uses guide sequence, target sequence, and additional non-sequence features
10 (FIG. 3A). We one-hot encoded guide and target sequence. The sequence input to the convolutional neural network (CNN) consists of 23 nt target and gRNA sequences with 2 nt of upstream and downstream target context. We also one-hot encode flanking target sequence context and set the guide sequence encoding at context positions to zero. We treat the encoded sequences as a two-dimensional image (four nucleotides x sequence length) with two channels
15 (guide and target). Our neural network layers are:

1. 2D convolution with 32 4 x 4 filters and ReLU activations
2. 2D convolution with 32 4 x 4 filters and ReLU activations
3. 2D max pooling with a 1 x 2 pool size
4. Flatten to vector
- 20 5. Dropout with 0.25 dropout rate
6. Concatenation with a vector of non-sequence features
7. Dense layer with 128 sigmoid outputs (hidden)
8. Dropout with 0.1 dropout rate
9. Dense layer with 32 sigmoid outputs (hidden)
- 25 10. Dropout with 0.1 dropout rate
11. Linear output layer with scalar output

Layers:

1, 2 - This neural network layer refers to a 2D convolutional layer that uses 32 filters with
30 a size of 4x4 each. The input to this layer would typically be a 2 dimensional image, and the layer would apply the 32 filters to different parts of the input image to extract features. The

ReLU activation function is then applied element-wise to the output of each filter. The “ReLU” activation function stands for Rectified Linear Unit, which is a common activation function used in deep learning. ReLU applies the function $f(x) = \max(0, x)$ element-wise to the output of the convolutional operation, setting any negative values to 0 and leaving positive values unchanged.

5 This introduces non-linearity into the network and helps it learn more complex representations of the input data.

3 - At the next stage, we perform max pooling with a pooling size of two.

4 - We flattened the resulting 32 channels of 16x4 learned feature representations to a vector of length 2048 to which we concatenate any non-sequence features (6). To regularize, we use 25% dropout (5) before our dense layers, which consists of three layers with 128 (7), 32 (9), and 1 (11) neurons, respectively. Between each dense layer, we use 10% dropout (8, 10). The first two dense (hidden) layers apply sigmoid activations. The single output neuron does not apply any activation function.

15 In addition to the framework discussed above, in certain embodiments one or more of the following non-sequence features may be input as a vector at the first dense layer:

i. target accessibility – (See, FIG. 8A, 8C (top))

1. log unpaired (FIG. 9E)
2. log unpaired at position 11 (FIG. 9F)
3. log unpaired at position 19 (FIG. 9G)
4. log unpaired at position 25 (FIG. 9H)

ii. hybridization MFE (See, FIG. 8B, 8C (bottom))

1. positions 1 and 23 (FIG. 9B)
2. positions 3 and 12 (FIG. 9D)
3. positions 15 and 9 (FIG. 9C)

iii. guide MFE (FIG. 9L)

iv. target location within transcript (encoded per Wei et al. (2021)) (FIG. 9I)

v. junction proximity

1. $\log_{10}(\text{5p distance to junction} + 1)$ (FIG. 9J)
2. $\log_{10}(\text{3p distance to junction} + 1)$ (FIG. 9K)

vi. guide secondary structure

1. presence of direct repeat stem loop (a Boolean variable) (FIG. 9M)

2. presence of a g-quadruplex (a Boolean variable) (FIG. 9N)

We considered six groups of non-sequence features: 1) target accessibility (i.e. lack of predicted secondary structure in multiple windows), 2) the RNA-RNA hybridization MFE between spacer and target site (multiple positions), 3) gRNA folding minimum free energy (MFE), 4) the target site's location within the transcript (relative position in CDS), 5) the target site's proximity to an exon-exon junction (5' distance, 3' distance), and 6) binary features related to the guide RNA's secondary structure (folded repeat and G-quadruplex). The data and ranking information relating to the non-sequence features are shown in Table 1 below.

Quantile	Feature	min	1st Qu.	median	3rd Qu.	max	mean	stdev
Q1	fold change (log2) D30/D0	-5.526	-2.891	-2.12	-1.449	0.242	-2.182	1.053
Q1	Distance to 5' junction (log x + 1)	0	3.466	4.143	4.644	6.832	3.965	0.978
Q1	Distance to 3' junction (log x + 1)	0	3.664	4.277	4.762	8.234	4.154	1.058
Q1	TargetSiteAccessibility_UnpairedProbability_Log10P-11p.23w	-11.054	-4.797	3.813	-2.958	-0.33	-3.907	1.369
Q1	TargetSiteAccessibility_UnpairedProbability_Log10P-11p.4w	-4.817	-1.063	0.724	-0.434	0.012	-0.807	0.527
Q1	TargetSiteAccessibility_UnpairedProbability_Log10P-19p.4w	-3.255	-1.233	0.864	-0.547	0.017	-0.942	0.548
Q1	TargetSiteAccessibility_UnpairedProbability_Log10P-25p.4w	-5.578	-1.092	0.717	-0.398	0.007	-0.814	0.573
Q1	crRNA MFE	-32.1	-16.1	-14.5	-13.3	-12.8	-14.899	1.939
Q1	Direct Repeat folding probability	0	1	1	1	1	0.895	0.307
Q1	G-quadruplex folding probability	0	0	0	0	1	0.001	0.032
Q1	hybridization crRNA MFE (1p.23w)	-62.6	-49.7	-47	-44.1	-32.4	-46.967	4.093
Q1	hybridization crRNA MFE (3p.12w)	-31.3	-24.5	-22.6	-20.5	-13.6	-22.531	2.982
Q1	hybridization crRNA MFE (15p.9w)	-25.6	-19.6	-18.3	-16.9	-9.9	-18.235	1.885
Q1	Relative location in CDS	0	0.286	0.51	0.736	0.999	0.508	0.275

Q2	fold change (log2) D30/D0	-2.76	-1.012	0.463	-0.221	0.017	-0.699	0.627
Q2	Distance to 5' junction (log x + 1)	0	3.555	4.234	4.718	6.831	4.038	1.047
Q2	Distance to 3' junction (log x + 1)	0	3.526	4.277	4.787	8.245	4.129	1.221
Q2	TargetSiteAccessibility_UnpairedProbability_Log10P-11p.23w	-12.604	-4.888	3.721	-2.767	0.421	-3.904	1.637
Q2	TargetSiteAccessibility_UnpairedProbability_Log10P-11p.4w	-5.337	-1.117	0.729	-0.418	0.005	-0.835	0.578
Q2	TargetSiteAccessibility_UnpairedProbability_Log10P-19p.4w	-4.996	-1.178	0.762	-0.461	0.007	-0.886	0.62
Q2	TargetSiteAccessibility_UnpairedProbability_Log10P-25p.4w	-4.817	-1.158	0.755	-0.408	0.004	-0.862	0.613
Q2	crRNA MFE	-33.4	-16.5	-14.9	-13.5	-12.8	-15.395	2.542
Q2	Direct Repeat folding probability	0	1	1	1	1	0.878	0.327
Q2	G-quadruplex folding probability	0	0	0	0	1	0.006	0.08
Q2	hybridization crRNA MFE (1p.23w)	-64.2	-49.8	-46	-42.7	-30.3	-46.379	5.446
Q2	hybridization crRNA MFE (3p.12w)	-33.5	-25.2	-22.9	-20.6	-12.8	-22.995	3.461
Q2	hybridization crRNA MFE (15p.9w)	-25.7	-18.9	-17.1	-15.2	-8.6	-17.079	2.753
Q2	Relative location in CDS	0	0.262	0.531	0.765	1	0.513	0.291
Q3	fold change (log2) D30/D0	-1.558	-0.2	0.054	0.017	0.14	-0.153	0.304
Q3	Distance to 5' junction (log x + 1)	0	3.638	4.304	4.804	6.819	4.131	1.06
Q3	Distance to 3' junction (log x + 1)	0	3.526	4.234	4.779	8.244	4.093	1.234
Q3	TargetSiteAccessibility_UnpairedProbability_Log10P-11p.23w	-14.037	-5.141	3.875	-2.762	0.169	-4.114	1.889
Q3	TargetSiteAccessibility_UnpairedProbability_Log10P-11p.4w	-5.286	-1.19	0.789	-0.422	0.004	-0.887	0.622

Q3	TargetSiteAccessibility_UnpairedProbability_Log10P-19p.4w	-6.364	-1.175	0.745	-0.394	0.003	-0.861	0.627
Q3	TargetSiteAccessibility_UnpairedProbability_Log10P-25p.4w	-5.882	-1.203	0.796	-0.43	0.002	-0.898	0.644
Q3	crRNA MFE	-36.5	-17.4	-15.2	-13.6	-12.8	-15.88	2.948
Q3	Direct Repeat folding probability	0	1	1	1	1	0.881	0.323
Q3	G-quadruplex folding probability	0	0	0	0	1	0.007	0.08
Q3	hybridization crRNA MFE (1p.23w)	-65.5	-50.575	-45.9	-42	-30.1	-46.499	6.35
Q3	hybridization crRNA MFE (3p.12w)	-34.7	-26.1	-23.6	-20.8	-13	-23.497	3.775
Q3	hybridization crRNA MFE (15p.9w)	-25.7	-18.8	-16.5	14.125	-7.8	-16.649	3.232
Q3	Relative location in CDS	0	0.229	0.47	0.746	0.999	0.484	0.289
Q4	fold change (log2) D30/D0	-0.287	0.077	0.139	0.194	0.34	0.126	0.097
Q4	Distance to 5' junction (log x + 1)	0	3.638	4.344	4.86	6.743	4.161	1.109
Q4	Distance to 3' junction (log x + 1)	0	3.584	4.317	4.875	8.247	4.184	1.236
Q4	TargetSiteAccessibility_UnpairedProbability_Log10P-11p.23w	-13.51	-5.264	-3.87	-2.701	0.041	-4.176	2.099
Q4	TargetSiteAccessibility_UnpairedProbability_Log10P-11p.4w	-4.166	-1.242	0.795	-0.463	0.006	-0.915	0.631
Q4	TargetSiteAccessibility_UnpairedProbability_Log10P-19p.4w	-5.92	-1.133	0.707	-0.365	0.002	-0.839	0.663
Q4	TargetSiteAccessibility_UnpairedProbability_Log10P-25p.4w	-5.871	-1.207	0.777	-0.428	0.004	-0.902	0.667
Q4	crRNA MFE	-36.1	-17.7	-15.5	-13.7	-12.8	-16.202	3.261
Q4	Direct Repeat folding probability	0	1	1	1	1	0.89	0.313
Q4	G-quadruplex folding probability	0	0	0	0	1	0.012	0.107

Q4	hybridization crRNA MFE (1p.23w)	-66.3	-50.6	-45.6	-41.2	-29.7	-46.381	6.954
Q4	hybridization crRNA MFE (3p.12w)	-33.7	-26.5	-23.7	-20.9	-12.9	-23.758	3.961
Q4	hybridization crRNA MFE (15p.9w)	-25.7	-18.7	-15.6	-13.6	-8	-16.215	3.432
Q4	Relative location in CDS	0.001	0.228	0.479	0.755	1	0.49	0.296
Q5	fold change (log2) D30/D0	0.114	0.286	0.366	0.483	1.335	0.398	0.16
Q5	Distance to 5' junction (log x + 1)	0	3.651	4.331	4.852	6.771	4.16	1.084
Q5	Distance to 3' junction (log x + 1)	0	3.555	4.344	4.883	8.237	4.218	1.242
Q5	TargetSiteAccessibility_UnpairedProbability_Log10P-11p.23w	-13.401	-5.063	3.734	-2.62	0.367	-4.032	1.987
Q5	TargetSiteAccessibility_UnpairedProbability_Log10P-11p.4w	-4.818	-1.282	0.827	-0.436	0.001	-0.931	0.653
Q5	TargetSiteAccessibility_UnpairedProbability_Log10P-19p.4w	-5.797	-1.136	0.682	-0.355	0.002	-0.835	0.656
Q5	TargetSiteAccessibility_UnpairedProbability_Log10P-25p.4w	-5.271	-1.211	0.773	-0.418	0.002	-0.902	0.663
Q5	crRNA MFE	-34.9	-17.45	-15	-13.5	-12.8	-15.944	3.176
Q5	Direct Repeat folding probability	0	1	1	1	1	0.868	0.339
Q5	G-quadruplex folding probability	0	0	0	0	1	0.011	0.105
Q5	hybridization crRNA MFE (1p.23w)	-65.2	-49.15	-44	-40.3	-27.8	-45.332	7.191
Q5	hybridization crRNA MFE (3p.12w)	-35.6	-26.2	-23.1	-20.45	-12.9	-23.359	4.016
Q5	hybridization crRNA MFE (15p.9w)	-25.7	-17.9	-15	-13	-7.9	-15.642	3.496
Q5	Relative location in CDS	0	0.228	0.484	0.755	1	0.49	0.296

For certain embodiments, Table 1 provides blaze marks for the non-sequence features, and the blaze marks for each quintile, with quintile 1 providing the highest efficacy levels, and quintile 5 providing the lowest. See, FIGs. 9A-9N.

In another embodiment, a method of titrating activity of a gene of interest is provided. After identifying an active PM gRNA, knockdown activity is titrated using various single mismatches (SM) between the gRNA and target site. See, FIG. 4E. TIGER can accurately predict “activity ratio”, i.e., the ratio of the fold-change of a SM gRNA to the fold-change of its PM cognate gRNA. (FIG. 4E), and thus, can be used to generate one or more gRNA capable of providing a desired level of knockdown/expression of the gene. In certain embodiments, this is particularly useful for conditions where gene dosage matters, e.g., haploinsufficiency. See, e.g., Huang et al, Characterising and Predicting Haploinsufficiency in the Human Genome, PLoS Genetics, 2010, Vol 6, No. 10, e1001154.

5 In a further embodiment, nt 15 to nt 21 (or nt 17 to 18 or nt 18) of the gRNA matching with its corresponding hybridization region of the target RNA without mismatches ranks higher than those with mismatches. In yet a further embodiment, gRNA having one or two or more mismatches to its corresponding target RNA ranks lower comparing to those having zero mismatches (PM).

10 In one aspect, provided is a method for predicting on-target activity of a gRNA. The gRNA composed of a DR stem loop and a spacer is capable of forming a complex with a Cas13d or a variant thereof and directing the complex to the target RNA. The method comprises characterizing one or more of the features (any one or combination of the features as disclosed herein) of a plurality of gRNAs and their corresponding target RNAs; assessing on-target activity of each of the gRNAs; constructing a model using the characterization data and the on-target
15 activity data by a modeling method. In one embodiment, the modeling method comprises TIGER modeling as described herein, and as exemplified at huggingface.co/spaces/Knowles-Lab/tiger. In a further embodiment, input of the model comprises characterization(s) of one or more of features of a gRNA and its corresponding target RNA. In yet a further embodiment, output of the model is an on-target score of the gRNA. As used herein an on-target score is an assigned
20 number (for example, an integer, rational number or irrational number) which positively correlates to on-target activity of a gRNA.

In one embodiment, the predicting method further comprises applying the constructed model to a gRNA and generating an on-target score of the gRNA. In a further embodiment, the predicting method comprises applying the constructed model to two or more gRNAs (such as a first gRNA and a second gRNA), and generating on-target scores of the gRNAs. In yet a further embodiment, the gRNAs share the same target RNA. In one embodiment, the gRNA is capable of hybridizing to a different (overlapping or non-overlapping) hybridization region of the same target RNA. In certain embodiments, the predicting method further comprises comparing the generated on-target scores and selecting the gRNA having the higher/highest score for directing the gRNA-Cas13d complex to the target RNA.

As used herein, an on-target activity of a gRNA (*i.e.*, efficacy of a gRNA) may refer to one or more of the following: cell survival, efficacy of the gRNA in forming a complex with a Cas13d protein or a variant thereof; efficacy of the gRNA in hybridizing to the corresponding target RNA; efficacy of the gRNA in directing a Cas13d-gRNA complex to the target RNA; efficacy of the gRNA in reducing the corresponding target RNA; and enrichment or abundance or depletion of the gRNA (or the guide of the gRNA or the target RNA) after applying the gRNA and a Cas13d or a variant thereof to a cell or cell culture. As an illustrative embodiment shown in the Examples, the gRNA efficacy was determined by quantifying gRNA abundances in sorted and unsorted cell populations. The value represents the log₂ fold change of sorted divided by input (for example, unsorted) counts. Higher values depict higher efficacies/efficiencies for target knockdown owed to the screen design. As used herein, an on-target score may be used to quantify the on-target activity. In one embodiment, an on-target score is an efficiency quintile as used here in (Q1 to Q5 also shown as bin1 to bin5). In another embodiment, an on-target score is a measured or calculated efficacy, for example, a fold change of gRNA/guide/target RNA abundance before versus after applying the gRNA.

In one aspect, provided is a method for predicting off-target activity of a gRNA. As disclosed herein, the gRNA is composed of a DR stem loop and a guide, and is capable of forming a complex with a Cas13d or a variant thereof and directing the complex to the target RNA. The predicting method comprises characterizing one or more of the features of a plurality of gRNAs and their corresponding target RNAs; assessing off-target activity of each of the gRNAs; and constructing a model using the characterization and the off-target activity acquired by a modeling method, *e.g.*, TIGER.

In one embodiment, the predicting method further comprises applying the TIGER model to a gRNA and generating an activity ratio of the gRNA. In a further embodiment, the predicting method further comprises applying the constructed model to two or more gRNA (for example, a first gRNA and a second gRNA) and generating activity ratios of the gRNAs. In yet a further embodiment, the gRNAs share the same target RNA. In one embodiment, the gRNA is capable of hybridizing to a different (overlapping or non-overlapping) hybridization region of the same target RNA. In certain embodiments, the predicting method further comprises comparing the generated activity ratio and selecting the gRNA having appropriate predicted activity ratio for the desired gene expression level.

As used herein, an off-target activity refers to an activity of a gRNA-Cas13d complex binds to and optionally nicks an RNA which is not the target RNA. An off-target effect refers to binding of a gRNA-Cas13d complex with an RNA which is not the target RNA and any consequence(s) thereof, for example, reduction of a non-target RNA, reduction of a peptide or a protein encoded by the non-target RNA, increase or reduction of a peptide or a protein whose expression is regulated by the non-target RNA, and any physiological change(s) relating thereto.

In a further aspect, provided is a method for selecting a gRNA from two or more of gRNAs for directing a complex which comprises the gRNA and a CRISPR-associated protein 13d (Cas13d) or a variant thereof (*i.e.*, Cas13d-gRNA complex or gRNA-Cas13d complex) to a target RNA.

As used herein, a modeling method refers to a mathematical or statistical analysis, for example, random forest models, classification and regression tree models, boosting, Bayesian networks, Markov random field, linear and generalized linear models, boosted tree models, neural networks, support vector machines, general chi-squared automatic interaction detector models, interactive tree models, multiadaptive regression spline, machine learning classifiers, a multi hypothesis testing, a principal component analysis, and any combinations thereof. These statistical models are well known to those of skill in the art. Any other suitable algorithms in performing the characterization process. In variations, the analysis can be characterized by a learning style including any one or more of: supervised learning (e.g., using logistic regression, using back propagation neural networks), unsupervised learning (e.g., using an Apriori algorithm, using K-means clustering), semi-supervised learning, reinforcement learning (e.g., using a Q-learning algorithm, using temporal difference learning), and any other suitable

learning style. Furthermore, the analysis can implement any one or more of: a regression algorithm (e.g., ordinary least squares, logistic regression, stepwise regression, multivariate adaptive regression splines, locally estimated scatterplot smoothing, etc.), an instance-based method (e.g., k-nearest neighbor, learning vector quantization, self-organizing map, etc.), a regularization method (e.g., ridge regression, least absolute shrinkage and selection operator, elastic net, etc.), a decision tree learning method (e.g., classification and regression tree, iterative dichotomiser 3, C4.5, chi-squared automatic interaction detection, decision stump, random forest, multivariate adaptive regression splines, gradient boosting machines, etc.), a Bayesian method (e.g., naive Bayes, averaged one-dependence estimators, Bayesian belief network, etc.), a kernel method (e.g., a support vector machine, a radial basis function, a linear discriminate analysis, etc.), a clustering method (e.g., k-means clustering, expectation maximization, etc.), an associated rule learning algorithm (e.g., an Apriori algorithm, an Eclat algorithm, etc.), an artificial neural network model (e.g., a Perceptron method, a back-propagation method, a Hopfield network method, a self-organizing map method, a learning vector quantization method, etc.), a deep learning algorithm (e.g., a restricted Boltzmann machine, a deep belief network method, a convolution network method, a stacked auto-encoder method, etc.), a dimensionality reduction method (e.g., principal component analysis, partial least squares regression, Sammon mapping, multidimensional scaling, projection pursuit, etc.), an ensemble method (e.g., boosting, bootstrapped aggregation, AdaBoost, stacked generalization, gradient boosting machine method, random forest method, etc.), and any suitable form of analysis. The machine learning classifier may be a discriminant analysis (DA) machine learning classifier, a nearest neighbor (NN) machine learning classifier, a random forest (RF) machine learning classifier, or a support vector machine (SVM). A DA machine learning classifier may be a linear discriminant analysis (LDA) classifier, or a quadratic discriminant analysis (QDA) classifier. In one embodiment, the SVM classifier may have three kernels, including a linear kernel, a radial basis function (RBF) kernel, and a polynomial kernel. In another embodiment, the machine learning classifier may employ a convolutional neural network (CNN). In one embodiment, a modeling method may be performed on a computer.

As used herein, characterizing a feature or a grammatical variation thereof refers to a qualitative or quantitative manner of describing the feature. For example, it may be presence or

absence of the feature, a numeric range of the feature, or a parameter/number/percentage calculated.

In certain embodiments, the ranking and/or any of the predicting methods as disclosed herein are determined *in silico* in software. Such software is, for example, an R language
 5 program, a Python program or similar. Other codes performing the same function may also be used.

These methods can be applied to target RNA which is a messenger RNA (mRNA), a mature mRNA, a primary transcript mRNA (pre-mRNA), a ribosomal RNA (rRNA), a 5.8S rRNA, a 5S rRNA, a transfer RNA (tRNA), a transfer-messenger RNA (tmRNA), an enhancer
 10 RNA (eRNA), a small interfering RNA (siRNA), a microRNA (miRNA), a small nucleolar RNA (snoRNA), a Piwi-interacting RNA (piRNA), a tRNA-derived small RNA (tsRNA), a small rDNA-derived RNA (srRNA), a non-coding RNA (ncRNA), long (intergenic) non-coding RNA (lincRNA/lncRNA), a single-stranded RNA (ssRNA), a circular RNA (circRNA), a vault RNA (vRNA/vtRNA), a SmY RNA, a double-stranded RNA (dsRNA), a small Cajal body-specific
 15 RNA (scaRNA), an antisense RNA (aRNA/asRNA), a ribonuclease RNA (e.g. RNase P), a non-coding regulatory RNA (e.g. 7SK RNA), RNA-viruses, single stranded DNA, coding sequence (CDS) of a RNA, 5' untranslated region (UTR) of a RNA, 3' UTR of a RNA, or a intron of a RNA, or a satellite repeat sequence embedded in any of said RNA targets.

Also provided is a non-naturally occurring and/or synthesized and/or engineered gRNA
 20 ranked and selected by a method as disclosed herein.

C. Other Components and Compositions

Also provided is a non-naturally occurring, synthesized (chemically or recombinantly) or engineered nucleic acid molecule comprising one or more of the gRNA(s) as disclosed, or a nucleic acid sequence complementary to the gRNA(s), or a nucleic acid sequence encoding the
 25 gRNA(s), or a nucleic acid sequence complementary to the gRNA coding sequence. In one embodiment, the nucleic acid molecule is a DNA.

In one embodiment, the nucleic acid molecule is a mature RNA. In one embodiment, the nucleic acid molecule comprises a DNA sequence encoding the gRNA(s). In a further embodiment, the nucleic acid molecule further comprises a first regulatory sequence directing
 30 expression of the gRNA(s). For example, the first regulatory sequence may comprise without

limitation, a Pol III promoter, for example, a U6 promoter, a H1 promoter, a T7 promoter, and a 7SK promoter.

In the embodiments relating to a nucleic acid molecule or a vector or uses thereof, a nucleic acid molecule encoding a gRNA may be in operative association with an RNA pol III promoter. A variety of Polymerase III promoters which can be used are publicly or commercially available, for example the U6 promoter, the promoter fragments derived from H1 RNA genes or U6 snRNA genes of human or mouse origin or from any other species. In addition, pol III promoters can be modified/engineered to incorporate other desirable properties such as the ability to be induced by small chemical molecules, either ubiquitously or in a tissue-specific manner. For example, in one embodiment the promoter may be activated by tetracycline. In another embodiment, the promoter may be activated by IPTG (lacI system). See, US5902880A and US7195916B2. In another embodiment, a Pol III promoter from various species might be utilized, such as human, mouse or rat.

In another embodiment, the nucleic acid molecule further comprises a DNA sequence encoding a Class 2, Type VI effector protein or a variant thereof. In one embodiment, the encoded protein is any Class 2, Type VI protein. In a further embodiment, the protein is a Cas13d protein. In another embodiment, the effector protein is a *RfxCas13d* from *Ruminococcus flavefaciens* strain XPD3002. In another embodiment, other Cas13d proteins may be utilized, for example, an *AdmCas13d* from *Anaerobic digester metagenome 15706*, *EsCas13d* from *Eubacterium siraeum DSM15702*, *PIE0Cas13d* from *Gut metagenome assembly PIE0-k21*, *UrCas13d* from *Uncultured Ruminococcus sp.*, *RffCas13d* from *Ruminococcus flavefaciens FDI*, and *RaCas13d* from *Ruminococcus albus*. In a further embodiment, the feature(s), ranges of the features(s), and any combination thereof may be adjusted according to a Cas13d other than *RfxCas13d*.

In a further embodiment, the Cas13d or a variant thereof further comprises a nuclear localization signal (NLS) or a cytosolic signal or a nuclear-export signal (NES). In another embodiment the Cas13d or a variant thereof is fused to an endoplasmic reticulum localization element, an Outer Mitochondrial membrane localization element, a Mitochondria localizing element, a Nucleolus localizing element (NIK3x), a Nuclear lamina localizing element (LMNA) or a Nuclear pore complex localizing element (SENP2). In yet a further embodiment, the Cas13d or a variant thereof is capable of nicking a target RNA. In one embodiment, the Cas13d or a

variant thereof has been engineered and does not have a nuclease activity, therefore referred to as a dead Cas13d.

In one embodiment, the DNA sequence encoding the effector, e.g., Cas13d, protein is under the control of a regulatory sequence directing expression thereof in a mammalian cell. In yet a further embodiment, the nucleic acid molecule comprises a second regulatory sequence which directs expression of the Cas13d protein or a variation thereof. In one embodiment, the second regulatory sequence comprises an RNA polymerase II (Pol II) promoter, for example, an EF-1 Alpha Short (EFS) promoter, or a Tet operator (tetO) promoter. In a further embodiment, the second regulatory sequence comprises one or more of the following: a polyadenylation (poly(A)) sequence, a selectable marker, a tag, and a Woodchuck Hepatitis Virus (WHP) Posttranscriptional Regulatory Element (WPRE) sequence. In certain embodiments, the tag is selected from one or more of the following: a FLAG tag, a poly(His) tag, a chitin binding protein (CBP) tag, a maltose binding protein (MBP) tag, a Strep tag, a glutathione-S-transferase (GST) tag, a thioredoxin (TRX) tag, a poly(NANP) tag, a V5 tag, a HA tag, a Spot tag, a T7 tag, a NE tag, a fluorescence tag, a Green Fluorescent Protein (GFP) tag, and a MYC tag. In one embodiment, the FLAG tag has a sequence of DYKDDDK, SEQ ID NO:47. In certain embodiments, the selectable marker is a puromycin resistance gene, a kanamycin resistance gene, a chloramphenicol resistance gene, a blasticidin S resistance gene, an ampicillin resistance gene, a tetracycline resistance gene, or a G418 resistance gene.

In one embodiment, the Cas13d coding sequence is operably linked to a regulatory element to ensure expression in a target cell. In a further embodiment, the promoter is an inducible promoter, such as a doxycycline inducible promoter. In a preferred embodiment, the regulatory element(s) comprises an RNA pol II promoter. A RNA pol II promoter is a promoter that is sufficient to direct accurate initiation of transcription by the RNA polymerase II machinery, wherein the RNA polymerase II (RNAP II and Pol II) is a RNA polymerase found in the nucleus of eukaryotic cells, catalyzing the transcription of DNA to synthesize precursors of messenger RNA (mRNA) and most small nuclear RNA (snRNA) and microRNA.

A variety of Polymerase II promoters that can be used within the compositions and methods described herein are publicly or commercially available to a skilled artisan, for example, viral promoters obtained from the genomes of viruses including promoters from polyoma virus, fowlpox virus (UK 2,211,504), adenovirus (such as Adenovirus 2 or 5), herpes

simplex virus (thymidine kinase promoter), bovine papilloma virus, avian sarcoma virus, cytomegalovirus (CMV), a retrovirus (*e.g.*, MoMLV, or RSV LTR), Hepatitis-B virus, Myeloproliferative sarcoma virus promoter (MPSV), VISNA, and Simian Virus 40 (SV40); other heterologous mammalian promoters including the actin promoter, β -actin promoter, immunoglobulin promoter, heat-shock protein promoters, human Ubiquitin-C promoter, PGK promoter. Additional promoters are readily known and available. See, *e.g.*, (Kadonaga, 2012), WO 2014/15134, and WO 2016/054153. In one particular embodiment, the promoter is a CMV promoter. In a further embodiment, the promoter is an EF-1 Alpha Short (EFS) promoter, or a Tet operator (tetO) promoter.

The term “regulatory element” or “regulatory sequence” refers to expression control sequences which are contiguous with the nucleic acid sequence of interest (for example, a Cas13d coding sequence or a sequence for expressing a gRNA) and expression control sequences that act *in trans* or at a distance to control the nucleic acid sequence of interest. As described herein, regulatory elements comprise but not limited to: promoter; enhancer; transcription factor; transcription terminator; efficient RNA processing signals such as splicing and polyadenylation signals (polyA); sequences that stabilize cytoplasmic mRNA, for example Woodchuck Hepatitis Virus (WHP) Posttranscriptional Regulatory Element (WPRE); sequences that enhance translation efficiency (*i.e.*, Kozak consensus sequence); sequences that enhance protein stability; and when desired, sequences that enhance secretion of the encoded product. Also, see Goeddel; Gene Expression Technology: Methods in Enzymology 185, Academic Press, San Diego, CA (1990). Regulatory sequences include those which direct constitutive expression of a nucleic acid sequence in many types of target cell and those which direct expression of the nucleic acid sequence only in certain target cells (*e.g.*, tissue-specific regulatory sequences). Furthermore, the Cas13d can be delivered by way of a vector comprising a regulatory sequence to direct synthesis of the Cas13d at specific intervals, or over a specific time period. It will be appreciated by those skilled in the art that the design of the vector can depend on such factors as the choice of the target cell, the level of expression desired, and the like.

As used herein, “operably linked” sequences or sequences “in operative association” include both expression control sequences that are contiguous with the nucleic acid sequence of interest (for example, a Cas13d coding sequence or a sequence for expressing a gRNA) and

expression control sequences that act *in trans* or at a distance to control the nucleic acid sequence of interest.

Optionally, the nucleic acid sequence encoding a Cas13d protein further comprises a reporter gene or a nucleic acid encoding a selectable marker, which may include sequences encoding geneticin, hygromycin, ampicillin or purimycin resistance, among others. A reporter gene, which is used as an indication of whether the Cas13d coding sequence has been incorporated into and/or expressed as a functional protein in the target cell or not, is readily selected by one of skill in the art, including without limitation, the *E. coli lacZ* gene, the chloramphenicol acetyltransferase (CAT) gene, or a gene encoding a fluorescent protein such as Green fluorescent protein (GFP).

In certain embodiments, one nucleic molecule comprises the sequence for the gRNA and a separate nucleic molecule encodes the sequence of the Cas13d protein.

Also provided is a vector comprising a gRNA and or a nucleic acid molecule as disclosed. In one embodiment, the vector is a viral vector, a retrovirus vector, a lentiviral vector, an adenovirus vector, an adeno-associated virus vector, or a hybrid viral vector. In another embodiment, the vector is a non-viral vector or an analogous carrier, such as a nanoparticle, a lipid complex, a polymer, a quantum dot, a carbon nanotube, a magnetic nanoparticle, or a gold nanoparticle. Further, a vector (for example, a plasmid) for producing of the vector is provided.

In still another embodiment, a ribonucleoprotein (RNP) complex as described herein includes a Class 2, Type VI effector protein and a gRNA, as defined herein. In still another embodiment, a cell is provided which contains one or more of the gRNA, nucleic acid molecules, RNP or compositions described herein. The cell may be mammalian, preferably a human cell. In other embodiments, the cell may be bacterial.

Additionally, provided is a library comprising a plurality of gRNAs or nucleic acid molecules or RNPs or vectors or cells as disclosed. In one embodiment, each of the gRNAs is capable of directing a Cas13d or a variant thereof to a different target RNA or a different region of one target RNA. In another embodiment, the library is a lentiviral library.

A composition is also provided comprising a pharmaceutical acceptable carrier and one or more gRNA(s), RNPs, or nucleic acid molecule(s) or vector(s), or cells as disclosed. These compositions may be for pharmaceutical use and thus useful in the treatment of a disease

associated with an abnormal RNA or misregulation of an RNA transcript. Some examples of these diseases are the diseases mentioned specifically above.

In yet a further embodiment, the gRNA, RNPs, pharmaceutical compositions, cells, vectors and libraries may also comprise gRNA having guide sequences which mismatch the target and allow the Class 2, Type VI effector protein to bind the target, but not elicit target degradation when used in the methods known to those of skill in the art as well as the methods described and exemplified specifically herein.

E. Methods of Use

One or more of the gRNAs, nucleic acid molecules, RNPs, vectors, cells, and libraries described herein are useful in a variety of methods including without limitation, treating a disease associated with an abnormal RNA; screening functional RNA(s); knocking-down, detecting, or editing a target RNA; or detecting or editing splicing, alternative isoforms, intron retention or differential UTR usage, or binding but not degrading the target.

In one aspect, the gRNA(s), nucleic acid molecule(s), RNP(s), vector(s), cell(s), or composition(s) containing one of more of them are used as a medicament, for example, in the treatment of a disease associated with an abnormal RNA such as by reducing the level of the abnormal RNA. Such disease may be a cancer/tumor, a virus infection, or a genetic disorder. In one embodiment, the treatment comprises contacting a target cell, and/or a biological sample from a subject having or suspected of having the disease with the gRNA(s), nucleic acid molecule(s), RNP(s), or vector(s) described herein. In further embodiment, target RNA of the gRNA(s) is/are the abnormal RNA(s) associated with the disease. In yet a further embodiment, the level of the abnormal RNA(s) in the target cell and/or in the biological sample is reduced. In one embodiment, the level of the abnormal RNA(s) after the treatment is reduced to at least about 10%, about 20%, about 30%, about 40%, about 50%, about 60%, about 70%, about 80%, about 90%, or about 95% of the level before the treatment or the level of a subject having this disease. In another embodiment, the level of the abnormal RNA(s) after the treatment is about 10%, about 20%, about 30%, about 40%, about 50%, about 60%, about 70%, about 80%, about 90%, about 100%, about 1.2 fold, about 1.5 fold, about 2 fold, about 3 fold or about 4 fold of a control level of a subject who is free of the disease. In still other methods, the targets are blocked but not degraded. In still other embodiments, the targets are modified temporarily. In other embodiment, the targets are modified permanently.

In another aspect, the gRNA(s), nucleic acid molecule(s), RNP(s), vector(s), cell(s), or composition(s) containing one or more of them are used as a medicament, for example, in the treatment of a disease associated with an insufficient amount of a gene or gene product, e.g., haploinsufficiency, such as by increasing the level of RNA. In one embodiment, the treatment comprises contacting a target cell, and/or a biological sample from a subject having or suspected of having the disease with the gRNA(s), nucleic acid molecule(s), RNP(s), or vector(s) described herein. In further embodiment, target RNA of the gRNA(s) RNA(s) associated with the insufficiency. In yet a further embodiment, the level of the RNA(s) in the target cell and/or in the biological sample is increased. In one embodiment, the level of the RNA(s) after the treatment is increased by at least about 10%, about 20%, about 30%, about 40%, about 50%, about 60%, about 70%, about 80%, about 90%, or about 95% of the level before the treatment or the level of a subject having this disease. In another embodiment, the level of the RNA(s) after the treatment is about 10%, about 20%, about 30%, about 40%, about 50%, about 60%, about 70%, about 80%, about 90%, about 100%, about 1.2 fold, about 1.5 fold, about 2 fold, about 3 fold or about 4 fold of a control level of a subject who is free of the disease. In certain embodiments, the level after treatment is a clinically therapeutic level.

In one embodiment, a method of treating a disease associated with an abnormal RNA or misregulation of an RNA transcript, comprises administering to a subject in need thereof the gRNA, nucleic acid molecules, vectors, RNAs, cells, or pharmaceutical compositions described herein. The administering step involves in one embodiment, delivering the selected or designed gRNA as a mature RNA to a cell that expresses an RNA-targeting CRISPR-associated protein, e.g., a Class 2, Type VI protein, such as Cas13d or a variant. In one embodiment the cell has been conditioned or modified to express the Cas13d or variant, and the administering occurs ex vivo. In another embodiment, the administering step involves delivering the gRNA described herein in a vector which co-expresses the RNA-targeting CRISPR-associated protein. In still another embodiment, the administering step involves delivering the gRNA and RNA-targeting CRISPR-associated protein as a ribonucleoprotein complex to the subject. In still a further embodiment, the administering step involves delivering the nucleotide molecule containing the gRNA with a separate nucleotide molecule that expresses the RNA-targeting CRISPR-associated protein.

As used herein, the terms “cancer” and “tumor” are used interchangeably and refer to an abnormal cell growth invading or spreading to other parts of the subject or having a potential of the invading or spreading. In order to achieve the abnormal cell growth, to invade other parts of the subject, and/or to evade an immune reaction, abnormal RNAs may be present in a

5 tumor/cancer cell. The cancer/tumor includes, but is not limited to, a solid tumor (*e.g.*, breast, colon, ovarian, lung, liver and glioma, Mesothelioma, and non-small cell lung cancer), a B cell lymphoma, a Cutaneous T cell lymphoma and a Lymphoid leukemia.

Upon virus infection, a target cell may generate abnormal RNA(s) in order to neutralize the virus. Additionally or alternatively, after the virus' entry to a target cell, the virus may utilize
10 the RNA producing machinery of the target cell producing abnormal RNA(s) in order to replicate the virus, or to lyse the target cell, or to perform other function(s) required by fulfilling the virus life cycle. Such virus infection may include HCV infection and related liver diseases, smallpox, the common cold and different types of flu, corona virus infections, measles, mumps, rubella, chicken pox, and shingles, hepatitis (HCV, HBV, or HAV), HIV, herpes and cold sores, polio,
15 rabies, Ebola and Hanta fever.

Abnormal RNA(s) may also be found in other diseases, including, without limitation, Atherosclerosis, Polycystic Kidney Disease, Cardiac disease, Cardiac stress, Myocardial infarction, Kidney fibrosis, Cardiac fibrosis, diabetes, Diabetes-related kidney complications, type 2 diabetes, non-alcoholic fatty liver diseases, mycosis fungoides, and Scleroderma. Other
20 representative examples of disease-causing defects associated with misregulation or defects in RNA include without limitation Prader Willi syndrome, Spinal muscular atrophy (SMA), Dyskeratosis congenita (X-linked), Dyskeratosis congenita (autosomal dominant), Dyskeratosis congenita (autosomal dominant), Diamond-Blackfan anemia, Shwachman-Diamond syndrome, Treacher-Collins syndrome, Prostate cancer, Myotonic dystrophy, type 1 (DM1), Myotonic
25 dystrophy, type 2 (DM2), Spinocerebellar ataxia 8 (SCA8), Huntington's disease-like 2 (HDL2), Fragile X-associated tremor ataxia syndrome (FXTAS), Fragile X syndrome, X-linked mental retardation, Oculopharyngeal muscular dystrophy (OPMD), Human pigmentary genodermatosis, Retinitis pigmentosa, Cartilage-hair hypoplasia (recessive), Autism, Beckwith-Wiedemann syndrome (BWS), Charcot-Marie-Tooth (CMT) Disease, Charcot-Marie-Tooth (CMT) Disease,
30 Amyotrophic lateral sclerosis (ALS), Leukoencephalopathy with vanishing white matter, Wolcott-Rallison syndrome, Mitochondrial myopathy and sideroblastic anemia (MLASA),

Encephalomyopathy and hypertrophic cardiomyopathy, Hereditary spastic paraplegia, Leukoencephalopathy 2, susceptibility to diabetes, deafness, and MELAS syndrome. See, for example, Thomas A. Cooper, et al, RNA and Disease, *Cell*, 136(4):2009, 777-793, ISSN 0092-8674; Scotti, M., Swanson, M. RNA mis-splicing in disease. *Nat Rev Genet* 17, 19–32 (2016).

- 5 <https://doi.org/10.1038/nrg.2015.3>; Matsui, M., Corey, D. Non-coding RNAs as drug targets. *Nat Rev Drug Discov* 16, 167–179 (2017). <https://doi.org/10.1038/nrd.2016.117>; Rupaimoole, R., Slack, F. MicroRNA therapeutics: towards a new era for the management of cancer and other diseases. *Nat Rev Drug Discov* 16, 203–222 (2017). <https://doi.org/10.1038/nrd.2016.246>; and Rohilla, K.J., Gagnon, K.T. RNA biology of disease-associated microsatellite repeat expansions. *Acta Neuropathol Commun* 5, 63 (2017). <https://doi.org/10.1186/s40478-017-0468-y>,
10 incorporated by reference herein.

In certain embodiments, the abnormal RNA(s) is/are presented in a biological sample. In a further embodiment, the abnormal RNA(s) may not be within a cell.

- In another aspect, a functional screening method is provided. The method comprises
15 contacting one or more gRNA(s), and/or nucleic acid molecule(s), and/or vector(s), and/or a library as disclosed with a target cell of a cell culture, a tissue, or a subject. In one embodiment, the method comprises amplifying the nucleic acid molecule or the vector in the target cell, and optionally quantifying the nucleic acid molecule or the vector.

- In one embodiment, a Cas13d protein is expressed by a nucleic acid molecule or a vector
20 in the target cell. Thus in the target cell, the gRNA forms a complex with a Cas13d or a variation thereof, and directs the complex to a target RNA. In a further embodiment, the nucleic acid molecule or vector is the same nucleic acid molecule or vector which comprises or expresses the gRNA(s). In another embodiment, the nucleic acid molecule or vector expresses the Cas13d protein but not the gRNAs and thus, is referred to as “Cas13d molecule” or “Cas13d vector” as
25 used herein. In one embodiment, the ratio of the Cas13d molecule (or Cas13d vector) to a gRNA (or nucleic acid molecule and/or vectors providing the gRNA) is about 100 to 1 to about 1 to 100, including each ratio therebetween. In one embodiment, the ratio is about 10 to 1, about 5 to 1, about 4 to 1, about 3 to 1, about 2 to 1, about 1 to 1, about 1 to 2, about 1 to 3, about 1 to 4, about 1 to 5, or about 1 to 10. In a further embodiment, the ratio is a molar ratio.

- 30 In one embodiment, the encoded Cas13d protein is a *RfxCas13d* from *Ruminococcus flavefaciens* strain XPD3002. Other Cas13d may also be utilized, for example, *AdmCas13d* from

Anaerobic digester metagenome 15706, *EsCas13d* from *Eubacterium siraeum* DSM15702, *PIE0Cas13d* from *Gut metagenome assembly PIE0-k21*, *UrCas13d* from *Uncultured Ruminococcus sp.*, *RffCas13d* from *Ruminococcus flavefaciens* FD1, and *RaCas13d* from *Ruminococcus albus*. In a further embodiment, the Cas13d or a variant thereof further comprises a nuclear localization signal (NLS) or a cytosolic signal or a nuclear-export signal (NES). In yet a further embodiment, the Cas13d or a variant thereof is capable of nicking a target RNA. In one embodiment, the Cas13d or a variant thereof has been engineered and does not have a nuclease activity. In one embodiment, the Cas13d is conjugated to a reporter molecule.

In one embodiment, the method reduces level of one or more of target RNA(s) in a target cell. In a further embodiment, the method functionally knocks down or knocks out one or more gene(s) expressing the target RNA(s). In yet a further embodiment, the method knocks down or knocks out one or more gene(s) in a plurality of target cells in parallel.

In certain embodiments, a selective pressure or a stimulus is applied to the target cells prior to, during or after the contacting step, which is referred to as a perturbation step. Such selective pressure or a stimulus includes, for example, a chemical agent or a biological agent or actively physically disturbing the target cell(s).

In certain embodiments, the method further comprises assessing cell viability, cell proliferation, cell apoptosis, cell death, cell phenotype, existence or concentration of a molecule (for example, the target RNA(s)), protein or cell marker expression, or response to a stimulus of a target cell, or a function which may be achieved by the cell culture, tissue, or subject comprising the target cell(s).

In yet a further aspect, a method for editing or modifying a target RNA is provided, comprising contacting a gRNA-Cas13d RNB complex with a target RNA. In one embodiment, this method or any composition used in the method is used for treatment of a disease associated with the target RNA.

In one embodiment, the gRNA of the complex is as disclosed herein. In a further embodiment, the complex is produced by a vector or a nucleic acid sequence disclosed. In one embodiment, the Cas13d nicks the target RNA. In another embodiment, the Cas13d has been engineered to have no nuclease activity. Other suitable Cas13d variants have been discussed in other sections of this application.

In a further embodiment, the Cas13d of the complex is engineered to edit or modify an RNA, for example. For example, the Cas13d may be conjugated to an RNA aminase, deaminase (*e.g.*, ADAR, ADAR1, ADAR2), methylase, or demethylase (*e.g.*, ALKBH5). In another embodiment, the Cas13d is conjugated to a splicing factor, for example a RBFOX1 or RBM38, whereby exon inclusion in the target RNA is induced when the hybridization region is at the downstream intron (*i.e.*, intron at the 3' side of an exon), and whereby exon exclusion in the target RNA is induced when the hybridization region is within the target exon. In yet another embodiment, the Cas13d is conjugated to a polyadenylation factor, for example, Nudix hydrolase 21 (NUDT21), whereby polyadenylation of RNA is induced at the hybridization region of the target protein.

A “nucleic acid” or a “nucleotide”, as described herein, can be RNA, DNA, or a modification thereof, and can be selected, for example, from a nucleic acid encoding a protein of interest, oligonucleotides, nucleic acid analogues, for example peptide- nucleic acid (PNA), pseudocomplementary PNA (pc-PNA), locked nucleic acid (LNA) etc. In certain embodiments, the terms “nucleotide” “nucleic acid” “nucleotide residue” and “nucleic acid residue” are used interchangeably, referring to a nucleotide in a nucleic acid polymer. In a further embodiment, consecutive nucleotide residues refer to nucleotide residues in a contiguous region of a nucleic acid polymer.

A nucleic acid molecule (RNA or DNA) or a nucleotide therein may be modified or edited. In one embodiment, such modification or edition includes 5' capping, 3' polyadenylation, and RNA splicing. In another embodiment, the modification or edition includes methylation (for example on a A residue resulting in a m⁶A), demethylation (for example, on a m⁶A, optionally via a RNA demethylase, including but not limited to ALKBH5), deamination (for example, from adenosine (A) to inosine (I), optionally via a tRNA-specific adenosine deaminase (ADAT), or from C to U, optionally via a pentatricopeptide repeat (PPR) protein), or amination (for example, from U to C or from G to A).

Ribonucleic acid (RNA) is a polymeric molecule essential in various biological roles in coding, decoding, regulation and expression of genes. As used herein, RNA may refer to a CRISPR guide RNA, a messenger RNA (mRNA), a mitochondrial RNA, short hairpin RNAi (shRNAi), small interfering RNA (siRNA), a mature mRNA, a primary transcript mRNA (pre-mRNA), a ribosomal RNA (rRNA), a 5.8S rRNA, a 5S rRNA, a transfer RNA (tRNA), a

transfer-messenger RNA (tmRNA), an enhancer RNA (eRNA), a small interfering RNA (siRNA), a microRNA (miRNA), a small nucleolar RNA (snoRNA), a Piwi-interacting RNA (piRNA), a tRNA-derived small RNA (tsRNA), a small rDNA-derived RNA (srRNA), a non-coding RNA (ncRNA), long (intergenic) non-coding RNA (lincRNA/lncRNA), a single-stranded RNA (ssRNA), a circular RNA (circRNA), a vault RNA (vRNA/vtRNA), a SmY RNA, a double-stranded RNA (dsRNA), a small Cajal body-specific RNA (scaRNA), an antisense RNA (aRNA/asRNA), a ribonuclease RNA (e.g. RNase P), a non-coding regulatory RNA (e.g. 7SK RNA), RNA-viruses or single stranded DNA. In one embodiment, the target RNA is an endogenous RNA. Additionally, or alternatively, the target RNA comprises/is a CDS. In another embodiment, the target RNA comprises/is a UTR (including a 5' UTR or a 3' UTR). In yet another embodiment, the target RNA comprises/is an intron.

As used herein, deoxyribonucleic acid (DNA) is a polymeric molecule formed by deoxyribonucleic acid, including, but not limited to, genomic DNA, double-strand DNA, single-strand DNA, DNA packaged with a histone protein, complementary DNA (cDNA which is reverse-transcribed from an RNA), mitochondrial DNA, and chromosomal DNA.

A “vector” as used herein is a biological or chemical moiety comprising a nucleic acid sequence which can be introduced into an appropriate cell for replication or expression of said the nucleic acid sequence. Common vectors include naked DNA, phage, transposon, plasmids, viral vectors, cosmids (Phillip McClean, www.ndsu.edu/pubweb/~mcclean/plsc731/cloning/cloning4.htm) and artificial chromosomes (Gong, Shiaoqing, *et al.* “A gene expression atlas of the central nervous system based on bacterial artificial chromosomes.” *Nature* 425.6961 (2003): 917-925). One type of vector is a “plasmid”, which refers to a circular double stranded DNA loop into which additional nucleic acid segments can be ligated. Another type of vector is a viral vector, wherein additional nucleic acid segments can be ligated into the viral genome. Certain vectors are capable of autonomous replication in a cell into which they are introduced (e.g., bacterial vectors having a bacterial origin of replication and episomal mammalian vectors). In certain embodiments, the vector is a lentiviral vector. Other vectors (e.g., non-episomal mammalian vectors) are integrated into the genome of a cell upon introduction into the cell, and thereby are replicated along with the cell genome.

A “viral vector” refers to a synthetic or artificial viral particle in which an expression cassette containing a nucleic acid sequence of interest is packaged in a viral capsid or envelope.

Examples of viral vector include but are not limited to lentivirus, adenoviruses (Ads), retroviruses (γ -retroviruses and lentiviruses), poxviruses, adeno-associated viruses (AAV), baculoviruses, herpes simplex viruses. In one embodiment, the viral vector is replication defective. A “replication-defective virus” refers to a viral vector, wherein any viral genomic sequences also packaged within the viral capsid or envelope are replication-deficient; *i.e.*, they cannot generate progeny virions but retain the ability to infect cells.

As used herein, “carrier” includes any and all solvents, dispersion media, vehicles, coatings, diluents, antibacterial and antifungal agents, isotonic and absorption delaying agents, buffers, carrier solutions, suspensions, colloids, and the like. The use of such media and agents for pharmaceutical active substances is well known in the art. Supplementary active ingredients can also be incorporated into the compositions. The phrase “pharmaceutically acceptable” refers to molecular entities and compositions that do not produce an allergic or similar untoward reaction when administered to a subject. Delivery vehicles such as lipid particle, liposomes, nanocapsules, nanosphere, nanoparticle, microparticles, microspheres, lipid particles, vesicles, and the like, may be used for the introduction of the compositions of the present invention into suitable target cells.

By “biological sample” is meant any biological fluids, cells or tissues of a subject that is suitable for use, such as, for example, cell-containing body fluids such as blood, sperm, cerebral spinal fluid, saliva, sputum or urine, leukocyte fractions, buffy coat, feces, swabs, puncture fluids, skin fragments, whole organisms or parts thereof, organs, organ fragments, tissues and tissue parts of a subject. Still other suitable samples are in the form of sections, biopsies, fine needle aspirates or tissue sections, isolated cells, for example in the form of adherent or suspended cell cultures, plants, plant parts, plant tissues from the fractions may be carried out at the same time or one or plant cells, bacteria, viruses, yeasts and fungi, without being limited thereto. In one embodiment, the biological sample contains a target RNA. In one embodiment, a suitable biological sample is a tissue section from human tissue, such as a tumor.

Pooled viral CRISPR “libraries” (in certain embodiments, lentiviral libraries) are a heterogenous population of viral transfer vectors, each containing an individual gRNA targeting a single gene in a given genome.

As used herein, the term “tag” refers to a peptide or polypeptide whose presence can be readily detected. In certain embodiments, the tag is selected from one or more of the following: a

FLAG tag, a poly(His) tag, a chitin binding protein (CBP) tag, a maltose binding protein (MBP) tag, a Strep tag, a glutathione-S-transferase (GST) tag, a thioredoxin (TRX) tag, a poly(NANP) tag, a V5 tag, a HA tag, a Spot tag, a T7 tag, a NE tag, a fluorescence tag, a Green Fluorescent Protein (GFP) tag, and a MYC tag. In certain embodiments, the tag is a fluorescent protein such as
5 Green fluorescent protein (GFP).

A “reporter molecule”, which is used to indicate the presence of a molecule to which it is conjugated (for example, a gRNA, a nucleic acid molecule, a protein, or a Cas13d), is readily known by one of skill in the art. In one embodiment, the reporter molecule may be a tag or a nucleic acid molecule encoding a tag. In another embodiment, the reporter molecule may be an
10 enzyme or a nucleic acid molecule expressing the enzyme, such as an *E. coli* lacZ enzyme, or a chloramphenicol acetyltransferase (CAT), or a luciferase.

As used herein, the term “selectable marker” refers to a molecule, a peptide or polypeptide whose presence can be readily detected in a target cell when selective pressure is applied to the cell. In certain embodiments, the selectable marker is a puromycin resistance gene,
15 a kanamycin resistance gene, a chloramphenicol resistance gene, a blasticidin S resistance gene, a geneticin resistance gene, a hygromycin resistance gene, an ampicillin resistance gene, a tetracycline resistance gene, or a G418 resistance gene.

The term “target cell” may refer to any cell of interest. Thus, a target cell may refer to a cell having a target RNA or suspected of having a target RNA. In certain embodiments herein,
20 the term “target cell” refers to a cell of various mammalian species. In one embodiment, the target cell is a mammalian cell. In a further embodiment, the target cell might be a eukaryotic cell, a prokaryotic cell, an embryonic stem cell, a cancer cell, a neuronal cell, an epithelial cell, an immune cell, an endocrine cell, a muscle cell, an erythrocyte, or a lymphocyte.

The term “mammal” or grammatical variations thereof, are intended to encompass a
25 singular “mammal” and plural “mammals,” and includes, but is not limited to, humans; primates such as apes, monkeys, orangutans, and chimpanzees; canids such as dogs and wolves; felids such as cats, lions, and tigers; equids such as horses, donkeys, and zebras; food animals such as cows, pigs, and sheep; ungulates such as deer and giraffes; rodents such as mice, rats, hamsters and guinea pigs; wild animals, such as bears, domesticated animals, livestock and laboratory
30 animals. In some preferred embodiments, a mammal is a human.

As used herein, the term “subject” includes any mammal in need of these methods or compositions, including particularly humans. The subject may be male or female.

As used herein, the terms “therapy”, “treatment” and any grammatical variations thereof shall mean any of prevention, delay of outbreak, reducing the severity of the disease symptoms, and/or removing the disease symptoms (to cure) in a subject in need.

The terms “a” or “an” refers to one or more. For example, “an expression cassette” is understood to represent one or more such cassettes. As such, the terms “a” (or “an”), “one or more,” and “at least one” are used interchangeably herein. In certain embodiment, the term “one or more” refers to any integer from one to the maximum including any integer therebetween.

The terms “another”, “first”, “second”, “third”, “fourth”, “fifth” and “sixth” are used throughout this specification as reference terms to distinguish between various forms and components of the compositions and methods, for example, first or second promoter.

As used herein, the term “about” means a variability of plus or minus 10 % from the reference given, unless otherwise specified.

The words “comprise”, “comprises”, and “comprising” are to be interpreted inclusively rather than exclusively, i.e., to include other unspecified components or process steps. The words “consist”, “consisting”, and its variants, are to be interpreted exclusively, rather than inclusively, i.e., to exclude components or steps not specifically recited.

As described herein, the terms “reduce” “decrease” “alleviate” “ameliorate” “improve” “delay” “earlier” “low” “high” “mitigate”, any grammatical variation thereof, or any similar terms indication a change, means a variation of about 5 fold, about 2 fold, about 1 fold, about 90%, about 80%, about 70%, about 60%, about 50%, about 40%, about 30%, about 20%, about 10%, about 5 % compared to a reference (e.g., a guide generated without using the disclosed methods, or a non-targeting control), unless otherwise specified.

Also, it is noted that any range as disclosed herein, (for example, an MFE range, a hybridization MFE range, a nucleotide (nt) range, a percentage range, or a log10 value range) includes the endpoint and every number/nt/percentage/value therebetween, unless specified.

It is also noted that any embodiment listed with respect to a gRNA, a nucleic acid molecule, a vector, a library, a composition, any other component, a method, or a use, may be combined with any other embodiments with respect to a gRNA, a nucleic acid molecule, a vector, a library, a composition, any other component, a method, or a use.

Technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs and by reference to published texts, which provide one skilled in the art with a general guide to many of the terms used in the present application. The definitions contained in this specification are provided for clarity in describing the components and compositions herein and are not intended to limit the claimed invention.

EXAMPLES

Transcriptome engineering in living cells with RNA-targeting CRISPR effectors is empowering a diverse and expanding set of applications, including transcript knockdown, editing, post-transcriptional modification, labeling, isoform targeting, affinity proteomics, and high-throughput perturbations. Crucially, all of these applications depend on accurate prediction of on-target activity and off-target avoidance. Here, we designed and tested ~120,000 RfxCas13d guide RNAs targeting essential genes in human cells with systematically-designed mismatches, insertions and deletions (indels). We found that mismatches and indels have a position- and context-dependent impact on Cas13d activity: single nucleotide mismatches at the center of a sensitive seed region lead to full gRNA inactivation, while extended mismatches at gRNA ends are largely tolerated, and thus are the main reason for unexpected off-target activity.

We also found that target mismatches that result in G:U wobble pairings are better tolerated than other single-base mismatches. Using this large-scale dataset, we trained a convolutional neural network that we term TIGER (Targeted Inhibition of Gene Expression via gRNA design) to predict efficacy from guide sequence and context. TIGER outperforms existing models at predicting on- and off-target activity on our dataset and published datasets. We show that TIGER scoring combined with specific mismatches yields the first general framework to modulate transcript expression, enabling use of RNA-targeting CRISPRs to precisely control gene dosage.

The following examples disclose both general and specific embodiments of the disclosed compositions and methods described herein, which should be construed to encompass any and all variations that become evident as a result of the teaching provided herein.

Example 1: Methods

Cell culture

The generation of doxycycline-inducible *RfxCas13d-NLS* HEK293FT and HAP1 cells was described before (Addgene #138149)⁷. *RfxCas13d-NLS* HEK293FT and HAP1 cells were maintained at 37°C with 5% CO₂ in D10 media or I10, respectively: DMEM with high glucose and stabilized L-glutamine (Caisson DML23) or IMDM (Caisson IML02) supplemented with 10% fetal bovine serum (Serum Plus II Sigma-Aldrich 14009C) and 5µg/mL Blasticidin S (ThermoFisher A1113903).

Screen library design and pooled oligo cloning

For this study we designed three CRISPR Cas13d gRNA libraries: 1) A 120,000 gRNA library tiling 16 essential genes with perfect match, mismatch, and indel gRNAs (Supplementary Data 1-2). 2) A 42,326 gRNA on-target library designed using our TIGER_{combined} model targeting 5,166 target genes with 8 PM gRNAs (Supplementary Data 5). 3) A 48,608 gRNA titration library targeting 1,082 essential genes with four perfect match and ten single mismatch gRNAs (Supplementary Data 10). For each target gene, we designed gRNAs against the most abundant isoform using publicly available transcript isoform quantifications from the Cancer Cell Line Encyclopedia (CCLE) for HEK-TE quantification (<https://sites.broadinstitute.org/ccle/>).

For the first library, we selected 16 genes (Supplementary Data 1) previously found to be essential using RNA-targeting CRISPR screens⁷. We predicted gRNA efficacies for all possible 23mer gRNAs using our RF_{on} random-forest model with minimal constraints (T-homopolymer < 4, V-homopolymer < 5)⁷. We only considered gRNAs falling within the coding region boundaries, which have been shown to be the most active previously⁷. For ten genes we selected, if possible, ~100 evenly spaced perfectly matching (PM) gRNAs for each prediction quartile to a total of 400 gRNAs. For six genes, we selected 1000 PM gRNAs in the same way across all prediction quartiles. For each of the six genes we selected 100 PM gRNA in the most effective quartiles Q3 (30 gRNAs) and Q4 (70 gRNAs) and designed gRNA variants per parental reference PM gRNA. These 181 variant gRNAs included: 23 single mismatch gRNAs (SM), 50 random double mismatch gRNAs, 8 consecutive double mismatches (DM), 50 random triple mismatches (RTM), 8 consecutive triple mismatchesTM, 8 single nucleotide deletions (SD), 8 random double nucleotide deletions (DD), 5 consecutive double deletions (CD), 8 single nucleotide insertions (SI), 8 random double nucleotide insertions (DI) and 5 consecutive double

insertions (CI). For nucleotide substitution or insertions, we chose a random base avoiding self-substitutions and terminal thymidine/uridine (T/U) bases. We filtered PM gRNAs and derivatives to not end with a T/U at position 23, as it would be interpreted as the start of the PolIII-terminator sequence directly downstream of the gRNAs, and would lead to gRNAs truncations.

5 For SM gRNA, we randomly sampled one nucleotide substitution per position to evenly cover all 23 positions. For all other categories, we randomly sampled from all possible variants. In total we designed 120,000 gRNAs (Supplementary Data 2; 10,000 PM gRNAs, 108,600 PM gRNA variants and 1,400 non-targeting (NT) control guides with >3 mismatches to the hg19 transcriptome). For the second and third library, we predicted PM and SM gRNA efficacies
10 using our TIGER_{combined} model. In total, we designed gRNAs for 5,166 genes. We included several (not mutually exclusive) gene groups: 1,082 common essential genes (DepMap release 08-2021; DepMap score < -1 in ≥ 500 cell lines or scored in all five screens in Hart *et al.*⁵³), 1,052 other genes (DepMap score < -0.5 in ≥ 100 cell lines or scored in at least three screens in Hart *et al.*⁵³), 1,477 RNA binding protein genes⁵⁴, 1,706 Transcription factor genes⁵⁵, and 458
15 control genes (DepMap score between -0.1 and +0.1 in ≥ 700 cell lines, spanning a wide range of gene expression values from 1 to >1,000 TPM). For all 5,166 target genes, we designed eight top PM gRNAs targeting the genes coding region. For the 1,082 essential genes, we designed for four top PM gRNAs ten SM gRNA variants randomly sampling across the 69 possible variants per PM gRNA with a roughly even spread of predicted low to high activity SM gRNA variants.
20 We added 1,000 non-targeting (NT) control guides with >3 mismatches to the hg19 transcriptome. On-target and titration libraries were designed together, padded by library specific priming sites enabling separate PCR amplification and plasmid library cloning.

The pooled gRNA libraries were synthesized as single-stranded oligonucleotides (Twist Biosciences) and then PCR amplified in 1 reaction per 10,000 gRNAs with a 50 μ l reaction
25 volume: 0.5 μ l Q5 polymerase (NEB), 10 μ l 5x reaction buffer, 2 μ l oligo pool (1ng/ μ l), 2.5 μ l of each fwd and rev primer (10 μ M), 2.5 μ l dNTPs (10mM), 30 μ l H₂O. Thermocycling conditions for the PCRs were 98C/30s, 8x or 9x[98C/10s, 63C/10s, 72C/15s], 72C/3min. The PCR product was gel-purified or purified using Zymo Clean and Concentrator 25 kit and Gibson-cloned into BsmBI-digested pLentiRfxGuide-Puro (Addgene #138151) using each time eight Gibson
30 reactions with a 20 μ l reaction volume: 500ng digested plasmid (0.088 pmol), 123.15ng purified oligo pool (1.3245 pmol, 15:1 molar ratio), 10 μ l 2x Gibson Assembly Master Mix (NEB),

incubated for 1 hour at 50°C. Each gRNA was represented with >200 colonies. Complete library representation with minimal bias (90th percentile/10th percentile gRNA read ratios of ~2.5 for all libraries) was verified by Illumina sequencing (MiSeq).

5 *Pooled lentiviral production and screening*

Lentivirus was produced via transfection of library plasmid pool and appropriate packaging plasmids (psPAX2: Addgene 12260; pMD2.G: Addgene 12259) using linear polyethylenimine MW25000 (Polysciences 23966). Specifically, we seeded ten million HEK293FT cells in 10cm dishes, transfected the cells the next evening (per dish: 60µl PEI, 10 9.2µg plasmid pool, 6.4 µg psPAX2, 4.4µg pMD2.G) and changed the medium the morning after. At 3 days post-transfection, viral supernatant was collected and passed through a 0.45 µm filter and stored at -80°C until use.

Doxycycline-inducible *RfxCas13d*-NLS HEK293FT and HAP1 cells were transduced with the pooled library lentivirus in separate infection replicates ensuring at least 1000x guide representation in the selected cell pool per infection replicate using a standard spinfection 15 protocol. We performed three or four independent infection replicate screens/experiments. After 24 hours, cells were selected with 1 µg/mL puromycin (ThermoFisher A1113803), resulting in ~30% cell survival. Puromycin-selection was complete ~48-72 hours post puromycin-addition. Assuming independent infection events (Poisson), we determined that ~83% of surviving cells 20 received a single sgRNA construct⁵⁶. *RfxCas13d* expression was induced by addition of 1µg/ml doxycycline (Sigma D9891) upon complete puromycin selection at the time of input sample collection (Day 0). Cells were passaged every two to three days (maintaining full representation) and supplemented with fresh doxycycline. For our initial screen, we collected genomic DNA (at least 1000 cells per construct representation) from each sample at day 0, day 15 and day 30. For 25 the TIGER on-target and titration screens we collected samples at day 0, day 7 and day 14.

Screen readout and read analysis

We used a two-step PCR protocol (PCR1 and PCR2) to amplify the guide RNA cassette for Illumina sequencing from genomic DNA (gDNA). The gDNA was extracted from screen 30 cells using the following protocol⁵⁶: Per 100 million cells, 12 mL of NK Lysis Buffer (50 mM Tris, 50 mM EDTA, 1% SDS, pH 8) were used for cell lysis. Once cells were resuspended, 60

mL of 20 mg/ml Proteinase K (Qiagen) was added and the sample was incubated at 55°C overnight. The next day, 60 mL of 20 mg/ mL RNase A (Qiagen) was added, mixed, and samples were incubated at 37°C for 30 min. Then 4 mL of pre-chilled 7.5 M ammonium acetate was added, samples were vortexed and spun at 4,000 g for 10 min. The supernatant was placed in a new tube, mixed well with 12 mL isopropanol and spun at 4,000 g for 10 min. DNA pellets were washed with 12 mL of 70% ethanol, spun, dried and pellets were resuspended with 0.2x TE buffer (Sigma). In addition, we also generated linearized plasmid library input and diluted it down to mimic similar copy number conditions as the genomic DNA samples.

For the PCR1 reaction, we used 960µg (screen 1) or 880µg (screen 461 2 and 3) gDNA for each sample. Per sample we performed 96 (screen 1) or 88 (screens 2 and 3) 100µL PCR1 reactions with a 100µl reaction volume: 10µl 10x Taq buffer, 0.02U/µl Taq-B enzyme (Enzymatics #P7250L), 0.2mM dNTPs, 0.2µM forward and reverse primers and 100ng gDNA/µl; PCR1 thermocycling conditions were: 94°C/30s, 10x (screen 1) or 18x (screens 2 and 3) [94°C/10s, 55°C/30s, 68°C/45s], 68°C/3min. Since our screen 1 library contained a large number of gRNAs with hamming distance of one to one another, we decided to only perform 10 cycles of PCR1. For each sample, all PCR1 products were pooled and mixed.

For each sample, we performed 24 (screen 1) or 6 (screen 2 and 3) PCR2 reactions with a 100µl reaction volume: 20µl 5x Q5 buffer (NEB), 0.01U/µl Q5 enzyme, 20µl PCR1-product, 0.2mM dNTPs, 0.4µM forward and reverse PCR2-primers in 100µl. PCR2 thermocycling conditions were: 98°C/30s, 18x (screen1) or 8x (screen 2 and 3) [98°C/10s, 63°C/30s, 72°C/45s], 72°C/5min. For screen 1 we performed an additional PCR2 on the linearized plasmid pool sample with either Q5 or Taq-B polymerase. We found raw counts to be highly correlated with no obvious influence due to the choice of polymerase. PCR primers can be found in Supplementary Data 14.

For each sample, PCR2 products were pooled, followed by normalization (gel-based band densitometry quantification), before combining equal amounts of uniquely-barcoded samples. The pooled product was then purified using SPRI beads. First, we performed a 0.6x vol/vol SPRI to remove gDNA carryover, followed by addition of a 0.3x vol/vol SPRI (0.6 + 0.3 = 0.9x final) to the supernatant to purify the ~260bp PCR product. Oligonucleotides can be found in Supplementary Data 14. The final amplicons were sequenced on Illumina NextSeq 500

– II MidOutput 1x150 v2.5 (screen 1) and Illumina NextSeq 500 – II HighOutput 1x150 v2.5 (screen 2 and 3).

Reads were first demultiplexed based on Illumina i7 barcodes present in PCR2 reverse primers using *bcl2fastq* and then by their custom in-read i5 barcode allowing for 1 mismatch.

5 Reads were trimmed to the expected gRNA length by searching for known anchor sequences relative to the guide sequence. Reads were collapsed (*FASTX-Toolkit*) to count perfect duplicates followed by exact string-match intersection with the reference to retain only perfectly matching and unique alignments. The raw guide RNA counts (Supplementary Data 3, 6 and 11) were normalized using a median of ratios method⁵⁷ and then batch corrected for biological replicates
10 using combat implemented in the *SVA* R package⁵⁸. Non-reproducible technical outliers were removed by flagging individual values with high variance within replicate samples of each timepoint (D0, D15, D30 for screen 1, and D0, D7, D14 for screen 2 and 3). These outlier counts are a common contaminant in early-passage input samples due to plasmid carryover from virus production in the viral supernatant used for infection⁵⁹. Specifically, we calculated the log₂-
15 transformed variance across all samples for each gRNA. Then, we calculated the variance within each time point across all replicate samples and flagged individual gRNA counts within the upper 0.6% variance percentile (e.g. cutoff = -1.366 for screen 1). We only flagged individual counts if those were present in the upper half of count distribution to avoid masking variance within depleting gRNAs. Since we had three replicates, we replaced the flagged count with NA,
20 but kept the other two unflagged replicates. For screen 1, we removed 154 guide RNAs due to all filtering steps (28 gRNAs were not detected; 126 gRNAs were only lowly represented in the plasmid library with less than 60 normalized counts).

Guide RNA enrichments (Supplementary Data 4, 7 and 12) were calculated building the count ratios between a timepoint and the corresponding input (Day 0) sample for each replicate
25 followed by log₂-transformation (log₂FC). Consistency between replicates was estimated using Pearson correlations and robust rank aggregation (RRA)⁶⁰. For data representation and modeling, we used the mean log₂FC across replicates. Delta log₂FC for mismatching guides was calculated by subtracting the log₂FC of the permuted gRNA from the PM reference guide. For data representations in FIG. 1 and FIG. 2 we normalized the observed log₂FC guide values in the
30 following way: For each gene *D* in the dataset, we computed the upper and lower quartiles of the guide log₂FC (*UQD* and *LQD*, respectively) as well as the corresponding quartiles for the log₂FC

among all datasets pooled together (UQP and LQP). We then updated each fold change x as follows:

$$\hat{x} = \left[\frac{x - LQ_D}{(UQ_D - LQ_D)} (UQ_P - LQ_P) + LQ_P \right]$$

5

Cell surface marker flow cytometry pooled screens

For certain analyses (e.g., entire dataset hold-out), we used a set of published pooled gene-tiling Cas13d screens using flow cytometry of cell surface markers from our previous study⁷. In these screens, library transduced cells were sorted based on expression of cell surface markers (CD46, CD55 or CD71). Specifically, we used the processed count data (available in
10 Supplementary Data 8 (ref.7)) for the CD46, CD55 and CD71 tiling screens, calculated gRNA fold-change by computing the count ratios between sorted bins (high vs. low) for each replicate followed by $\log_2$ -transformation ($\log_2FC$).

15 *Predicting RNA secondary structures and RNA-RNA hybridization energies*

gRNA secondary structure and minimum free energies (MFEs) was derived using RNAfold [--gquad] on the full-length gRNA (DR + guide) sequence⁶¹. Target RNA pairing probability (accessibility) was calculated using RNAplfold [-L 40 -W 80 -u 50] as described before⁷. These parameters specify a moving window of 80 nucleotides and a maximal base
20 pairing span of up to 40 nucleotides. We chose these parameters because prior studies^{3,62,63} both in the context of Cas13 and RNA interference have found optimal performance for a local window around the target site.

We performed a grid-search calculating the RNA accessibility for each target nucleotide in a window of minus 20 bases downstream of the target site to plus 20 bases upstream of the
25 target site assessing the unpaired probability of each nucleotide over 1 to 50 bases for all perfectly matching guides. Then, we calculated the Pearson correlation coefficient between the $\log_{10}$ -transformed unpaired probabilities and the observed gRNA $\log_2FC$ for each point and window relative to the gRNA. We selected four centers of high correlation to feed into the model. Target RNA accessibility features: 1. Position -11 upstream to the first spacer nucleotide

with width 23 nt; 2. Position -11 with width 4 nt; 3. Position -19 with width 4 nt; 4. Position -25 with width 4 nt.

RNA-RNA-hybridization between the guide RNA (PM and gRNA with nucleotide substitutions, but not for indel gRNAs) and its target site was calculated using RNAhybrid [-s - c]⁶⁴. We calculated the RNA hybridization minimum free energy for each guide RNA nucleotide position p over the distance d to the position $p + d$ with its cognate target sequence. All measures were either directly correlated with the observed gRNA log₂FC or using partial correlation to account for the gRNA folding MFE. In each case, we computed the Pearson correlation. We selected three centers of high correlation to feed into the model. Hybridization MFE features: 1. Position $p=1$ and $d=23$; 2. Position $p=3$ and $d=12$, 3. Position $p=15$ and $d=9$.

Assessing target RNA context

To assess the target RNA context, we calculated the nucleotide probability at each position (p) over a window (w) of 1 to 50 nucleotides centered around the position of interest (e.g. $p = -18$ with $w = 11$ summarizes the nucleotide content in a window from -23 to -13 with +1 being the first base of the gRNA). We evaluated p for all positions within 75 nucleotides upstream and downstream of the guide RNA. The nucleotide context of each point was then correlated with the observed log₂FC gRNA enrichments for all perfect match gRNAs, either directly or using partial correlation accounting for gRNA folding MFE. In each case we used Pearson correlation. We used the same positions p and window sizes that have been used before in RF_{on}⁷. These RNA nucleotide context features were only used in the RF_{on}, but not in the CNN models.

Gene essentiality normalization and determination of active guides

Given that essentiality varies across the 16 genes in our first pooled library (i.e., some genes are more essential than others), we experimented with different per-gene normalization approaches (including no normalization) to equalize survival effects across all genes in our survival screen prior to data modeling. All of our considered transformations are location-scale transformations, where the location and scale were derived separately for each gene. Data were transformed as follows: $(x - \text{location}) / \text{scale}$. We found the best generalization performance was

obtained using median fold-change as the location and the distance between the 10th and 90th percentiles as the scale.

We used the distribution of non-targeting gRNAs to determine which gRNAs targeting essential genes to consider as being active. We selected the most depleted 1% of non-targeting gRNAs as the threshold for activity after testing for normality of the non-targeting gRNA distribution (Lilliefors test, $p < 0.001$). This threshold corresponds to a $\log_2\text{FC} < -0.50$ for screen 1 (HEK293FT), $\log_2\text{FC} < -0.44$ and $\log_2\text{FC} < -0.29$ for screen 2 in HEK293FT and HAP1, respectively.

10 *Convolutional neural network for deep learning*

The sequence input to the convolutional neural network (CNN) consists of 23 nt target and gRNA sequences with 2 nt of upstream and downstream target context. The input initially is processed with two consecutive convolution layers each with 32 4x4 kernels followed by a rectified linear unit (ReLU). At the next stage, we perform max pooling with a pooling size of two. We flattened the resulting 32 channels of 16x4 learned feature representations to a vector of length 2048 to which we concatenate any non-sequence features. To regularize, we use 25% dropout before our dense layers, which consists of three layers with 128, 32, and 1 neurons, respectively. Between each dense layer, we use 10% dropout. The first two dense (hidden) layers apply sigmoid activations. The single output neuron does not apply any activation function. This design is similar to prior deep learning models in computer vision³⁸ and sequence analysis²⁶.

We use a log hyperbolic cosine (log-cosh) loss function, which is similar to an L1 loss, but is continuously differentiable. We optimize our models with *Adam* (adaptive moment estimation), an adaptive stochastic optimization algorithm that requires only first-order gradients⁶⁵, with a learning rate of 0.001. We employ early stopping with a patience of 100 epochs and restore model parameters of the best epoch. We implemented our models using TensorFlow via the Docker image *tensorflow:2.8.0-gpu*. Additionally, our code imports the following python packages: *biopython* (version 1.79), *pandas* (version 1.4.5), *tensorflow586* *probability* (version 0.16.0), *matplotlib* (version 3.5.1), *seaborn* (version 0.11.2), *shap* (version 0.41.0), and *statsmodels* (version 0.12.2).

30 *Cross-validation across genes, target 589 sites, gRNAs and datasets*

We consider three cross-validation (CV) approaches: gene level, target level and individual gRNAs. For gene-level CV, we create sixteen folds where each fold contains all (gRNA, target) tuples specific to that gene. This CV approach promotes transcriptome-wide generalization by holding out entire genes (and all of the corresponding target sites and gRNAs for that gene). The second approach (target-level CV) divides target sites into ten, non-overlapping folds. Holding out a target site places all perfect match and mismatched gRNAs designed for that target site into the holdout set, ensuring that a target sequence never appears both in training and validation sets. This is important as active PM and SM gRNAs for the same target sequence can have similar activity. The third approach (gRNA-level CV) holds out individual gRNAs; importantly, related gRNAs such as mismatch gRNAs (for a particular PM gRNA in the holdout group) may still be included in the training set. As expected, we observe better performance with gRNA600 level CV than target- or gene-level CV.

In some experiments, we train on the 120,000 gRNA library and then test on a separately collected pooled screen: flow cytometry of cell surface proteins from three genes (entire dataset holdout)⁷. For the entire dataset holdout, we used the log₂-transformed gRNA depletion in the fluorophore-low bin divided by the fluorophore-high bin. Supplementary Note 1 contains a formal description of all of these cross-validation strategies. All CV folds used in the study are presented in Supplementary Data 2.

Comparison with linear regression, random forest and recurrent neural networks

To quantify the random forest (RF_{on}) performance for predicting PM gRNA efficacy on the essentiality screen, we used the same 16 gene holdouts described above and iteratively re-trained the model with its previously-described architecture⁷ on 15 genes to predict the 16th gene. In addition, we re-trained RF_{on} on the entire essentiality data set to evaluate its performance on the phenotype selection data (flow cytometry). For the recent Cas13 deep learning models^{8,9}, we uploaded each of the sixteen transcripts for our targeted essential genes to their respective web portals in order to generate predictions for all potential target sites along the transcript (1 nt tiling). We only compute performance at target sites for which we measured Grna activity. Cheng *et al.*⁹ transform fold-change data via a parameterized sigmoid function. We exactly applied this transform to our fold-changes prior to computing their performance metrics. Wei *et*

*al.*⁸ generate predictions for 30 nt spacers. When computing their performance, we use the first 23 nt of their 30 nt spacer sequence to match their predictions to our 23 nt gRNA sequences.

We also designed a Bidirectional Gated Recurrent Unit (BiGRU)⁶⁶ model as an alternative deep learning approach that is based on recurrent units instead of convolution units.

5 We designed the BiGRU architecture to closely mimic the dense layers of TIGER. The BiGRU model uses a convolution kernel of length one to learn a 32-dimensional embedding from the one-hot-encoded target and gRNA sequences (16 unique 32-dimensional embeddings for each possible guide-target pair). This embedding feeds a bidirectional gated recurrent unit (GRU) layer, which outputs a 32-dimensional representation for each sequence position. We concatenate
10 the 32-dimensional outputs for both directions and all sequence positions and thereafter flatten it to a vector of length 64 x sequence length and concatenate non-sequence features. To regularize, we apply 25% dropout on this vector before feeding it to our dense layers, which consists of three layers with 128, 32, and 1 neurons. Between each dense layer, we use 10% dropout. The first two dense (hidden) layers apply sigmoid activations. The single output neuron does not
15 apply any activation function.

For linear regression, we used one-hot encoded sequences (flattened into a vector) and concatenate non-sequence features to the same vector.

Example 2: A massively-parallel *RfxCas13d* screen for perfect-match and variant guide RNAs

20 To systematically assess the efficacy of *RfxCas13d* guide RNA (gRNAs), we designed ~120,000 Cas13d gRNAs with a diverse set of mismatches, insertions, and deletions (indels) to target known essential genes (FIG. 1A, FIG. 1B, Supplementary Data 1, 2). This gRNA pool contains 10,000 perfect-match (PM) gRNAs for 16 genes and 108,600 gRNAs with designed mismatches for 600 PM gRNAs for six of these genes. In this manner, we can compare how each
25 engineered gRNA mutation impacts Cas13d activity related to its cognate PM gRNA. We designed the mismatch gRNAs to contain 1, 2 or 3 nucleotide mismatches and the indel gRNAs to contain 1 or 2 nucleotides indels. For both mismatches and indels, we designed separate groups of gRNAs with adjacent placement of mismatches/indels or random spacing of the mismatches/indels.

30 Targeting known essential genes in human cells, we performed cellular fitness (dropout) screens with the expectation that cells will drop out of the population over time depending on the

relative gRNA activity and corresponding degree of essential gene depletion (FIG. 1C). We lentivirally transduced a library of 120,000 gRNAs into a monoclonal HEK293FT cell line with doxycycline-inducible nuclear-localized *RfxCas13d* nuclease⁷. We found that gRNA counts and gRNA depletion (fold-change relative to the gRNA abundance at an early time point) were highly reproducible between replicate screens, timepoints and gRNA categories (FIG. 1D, Supplementary Fig. 1a-d, Supplementary Data 3, 4).

We first validated the performance of our previous Random Forest on-target model (RF_{on})⁷

for PM gRNAs. For PM gRNAs, the 23 nt of the spacer region contain the reverse complement of

the intended RNA target site. Specifically, 70.1% of predicted most-active quartile (Q4) gRNAs depleted more strongly than expected based on a false-discovery rate calculated using the non targeting (negative control) gRNAs ($FDR < 0.01$) (FIG. 1E, Supplementary Fig. 1e). This fraction

of active Q4-gRNAs ranged from ~95% (104 out of 110 gRNAs) for EIF3B to ~49% for NUP133

(53 out of 109 gRNAs) (FIG. 1F, Supplementary Fig. 1f). We found that active gRNAs were distributed in clusters along the target transcript sequence, yielding significant similarities between

neighboring gRNA efficacies (autocorrelations of $r = 0.13 - 0.60$) (FIG. 1G, FIG. 1H), in agreement with our prior study (CD46 $r = 0.66$; CD55 $r = 0.65$; CD71 $r = 0.4$)⁷.

Example 3: Indels are more deleterious than base substitutions for Cas13d gRNAs

For 600 perfect-match (PM) gRNAs predicted to have high activity by RF_{on} (quartile Q3 or Q4), we designed 108,600 gRNA variants (18,100 per gene). These variant gRNAs include 83,400 gRNAs with single, double, or triple base substitutions. We also included 25,200 gRNAs containing single or double insertions or deletions (indels) (FIG. 1A, FIG. 1B). We found 66.1% of PM gRNAs to be active ($\log_2$ fold-change < -0.5 , $FDR < 0.01$) (FIG. 2A) (see Methods).

Accumulating base substitutions gradually decreased gRNA efficacy from single mismatches (34.4% active gRNAs) to random triple mismatches (3.3% active gRNAs) (FIG. 2B). Overall, base substitutions were better tolerated than indel variants such as single nucleotide deletions

(SD) or single nucleotide insertions (SI) within the gRNA sequence (guides active: SM 34.4% > SD 20.7% > SI 17.9%).

Next, we calculated the relative activity for all gRNA variants relative to their cognate PM gRNAs (see Methods). Most single mismatch variants resulted in modest decreases in activity compared to the cognate PM gRNA (FIG. 2C). In contrast, single indels resulted in greater loss of activity compared to single mismatch variants with the greatest loss of activity for insertions in the central region of the gRNA (FIG. 2D, FIG. 2E). Unlike base substitutions, indels introduce bulges on either gRNA or target side, respectively. Since the gRNA is embedded within the Cas13d enzyme²⁸, gRNA positioning is likely more constrained than target sequence arrangement within the ternary Cas13-gRNA-target complex. This may explain why RNA bulges on the gRNA side (as introduced by nucleotide insertions) are the most disruptive.

We also confirmed the presence of the single mismatch-intolerant seed sequence centered on guide nucleotide positions 18⁷ (FIG. 2F). Single base substitutions outside the seed region led to a milder attenuation of relative gRNA efficacy for all single-mismatch gRNAs. We noted that A-to-G and to a lesser extent C-to-U substitutions within the gRNA had a milder effect than other substitutions, including in the seed region (FIG. 2F). A-to-G substitutions lead to G-U wobble pairing with the G on the gRNA side, while C-to-U substitutions have the G in the target site. We found that for all mismatch types, the contribution of G-U wobbles ameliorated the relative decrease in efficacy from the mismatch (FIG. 2G, Supplementary Fig. 1h).

Example 4: A deep learning model to predict guide RNA efficacy

Existing approaches to model Cas13d efficacy⁷⁻⁹ predict knockdown efficiencies only for gRNAs perfectly matching their target site. Although large-scale off-target screens have empowered predictive modelling for DNA-targeting CRISPRs like Cas9^{26,29-36} and Cas13a mediated RNA diagnostics *in vitro*¹⁰, there have not been systematic efforts to learn an off-target model for RNA-targeting CRISPR *in vivo*. Towards this end, we adapted a convolutional neural network (CNN) architecture similar to those pioneered in computer vision^{37,38} to predict a guide's efficacy from a one-hot encoding of its sequence. Similar to a previous CNN for Cas9 off-target prediction²⁶, our model has two convolution layers followed by a max-pooling layer and interleaves three dropout and dense layers for a total of two hidden layers plus an output layer.

Our model has two architectural augmentations beyond those used in prior work: additional sequence context flanking the 23nt target site and the flexibility to input a vector of non-sequence features at our first dense layer. We considered six groups of non-sequence features: 1) gRNA folding minimum free energy (MFE), 2) the RNA-RNA hybridization MFE between spacer and target site (multiple positions), 3) target accessibility (i.e. lack of predicted secondary structure in multiple windows), 4) the target site's proximity to an exon-exon junction (5' distance, 3' distance), 5) the target site's location within the transcript (relative position in CDS) and 6) binary features related to the guide RNA's secondary structure (folded repeat and G-quadruplex). We term this deep learning approach TIGER (Targeted Inhibition of Gene Expression via gRNA design) (FIG. 3A).

We first sought to determine the optimal flanking target sequence context when using only nucleotide sequence or all features (FIG. 3B, Supplementary Fig. 2a). We added additional context to just the 5'-end, just the 3'-end, and equally to both ends and predicted gRNA efficacy using 10-fold cross-validation (CV) over target sites (Supplementary Note 1). Consistent with findings from another recent study⁸, we found that additional 5' target site context of three nucleotides was optimal for the sequence-only model. However, the impact of this additional context is reduced when including non-sequence features, which capture 5'-end target accessibility and likely make the extra 5' context redundant. When analyzing the impact of each non-sequence feature in TIGER, we find that target RNA accessibility yields the greatest increase in performance (FIG. 3C).

Although our CNN is a regression model that predicts fold-change in guide abundance, we can also threshold our predictions to classify active guides versus inactive ones. We again identified active gRNAs based on an empirical false-discovery rate calculated from the non-targeting gRNAs ($FDR < 0.01$). To best assess generalization across the transcriptome, we aggregated predictions across gene-level CV (holding all gRNAs for each of the 16 genes in turn) to compute correlations between predictions and observations (Pearson and Spearman correlations), area under the receiver-operator characteristic curve (AUROC), and area under the precision-recall curve (AUPRC) (FIG. 3D, Supplementary Fig. 2b). For each of these metrics, we estimated an upper-bound by taking advantage of the three biological replicates of our Cas13d essential gene screen, quantifying the error when using a held-out replicate to predict the mean of the other two replicates. Across all four evaluation metrics, we find that TIGER trained

on PM gRNAs in combination with mismatched gRNAs (TIGER_{combined}) outperforms or matches TIGER trained on PM alone (TIGER_{on-target}) and yields superior performance compared to our previously published model (RF_{on})⁷ and deep learning models from two recent preprints^{8,9}. The gene-level CV (all gRNAs for a target gene are held-out from the training set) ensures that, during training, TIGER does not have access to any data related to the particular gene used for evaluation.

In addition to holding out all gRNAs for individual genes, we repeated these experiments with target site-level or gRNA-level CV (Supplementary Fig. 2c-h). Holding out target sites ensures a PM gRNA's mismatched variants will not be used in training (e.g. active SM gRNA are highly informative of their PM parents) (Supplementary Note 1). Under these CV strategies, the on-target and combined models see a boost in performance relative to gene-level CV. Here too, the combined model outperforms all other considered models. Both TIGER models perform close to this estimated upper-bound for the Spearman correlation, suggesting that they are adept at filtering variation due to technical noise.

In addition to the survival screen with essential gene-targeting gRNAs, we further evaluated model performance using a separate gRNA dataset with a different phenotypic selection, flow cytometry of cell surface proteins⁷, which has been used by other groups to assess generalization performance^{8,9}. Specifically, we use PM gRNAs from our survival screen as training data (PM and mismatched gRNAs in case of TIGER_{combined}), holding out PM gRNAs from the cell surface protein screen as test data⁷. To compare with our previous RF_{on} model, we re-trained it solely on the larger essentiality screen dataset (FIG. 1A). Importantly, this validation dataset contains no genes or target sites in common with our survival screen. When testing generalization performance on this validation dataset, we find that TIGER_{combined} and TIGER_{on-target} yield best-in-class predictions (FIG. 3E, Supplementary Fig. 2i). To further benchmark TIGER, we also trained a linear regression model and a recurrent neural network (Bidirectional Gated Unit) model and, across both datasets, we found that TIGER was either superior or comparable to these approaches (Supplementary Fig. 3). Taken together, we find that TIGER's predictions generalize over different screen modalities (cell proliferation and surface marker expression) and target genes (essential and non-essential).

Example 5: TIGER Model Training

Data Generation

To systematically assess the efficacy of *Rfx*Cas13d guide RNA (gRNAs), we designed
 5 ~120,000 Cas13d gRNAs with a diverse set of mismatches, insertions, and deletions (indels) to
 target known essential genes (FIG. 1A, FIG. 1B, Supplementary Data 1, 2). This gRNA pool
 contains 10,000 perfect-match (PM) gRNAs for 16 genes and 108,600 gRNAs with designed
 mismatches for 600 PM gRNAs for six of these genes. In this manner, we can compare how each
 engineered gRNA mutation impacts Cas13d activity relative to its cognate PM gRNA. We
 10 designed the mismatch gRNAs to contain 1, 2 or 3 nucleotide mismatches and the indel gRNAs
 to contain 1 or 2 nucleotides indels. For both mismatches and indels, we designed separate
 groups of gRNAs with adjacent placement of mismatches/indels or random spacing of the
 mismatches/indels.

Targeting known essential genes in human cells, we performed cellular fitness (dropout)
 15 screens with the expectation that cells will drop out of the population over time depending on the
 relative gRNA activity and corresponding degree of essential gene depletion (FIG. 1C). We
 lentivirally transduced a library of 120,000 gRNAs into a monoclonal HEK293FT cell line with
 doxycycline-inducible nuclear-localized *Rfx*Cas13d nuclease⁷. We found that gRNA counts and
 gRNA depletion (fold-change relative to the gRNA abundance at an early time point) were
 20 highly reproducible between replicate screens, timepoints and gRNA categories (FIG. 1D,
 Supplementary Fig. 1a-d, Supplementary Data 3, 4).

Data Preparation

Our guides target sixteen essential genes for knockdown. Let response variable y_{gij}
 denote j 'th replicate of the i 'th gRNA's observed log2-fold change (LFC) at day 21 for
 25 gene g . We first average the replicates: $y_{gi} = \text{mean}(y_{gi1}, \dots)$.

Because different essential genes vary in how quickly their knockdown affects
 cell survival, we normalize LFC on a per-gene basis. We compute for each gene g in our
 set of sixteen genes G

$$\begin{aligned} m_g &\triangleq \text{quantile}(\{y_{gi} \in g : y_{gi}'\text{'s guide is PM}\}, q = 0.5) \\ s_g &\triangleq \text{quantile}(\{y_{gi} \in g : y_{gi}'\text{'s guide is PM}\}, q = 0.9) - \\ 30 &\quad \text{quantile}(\{y_{gi} \in g : y_{gi}'\text{'s guide is PM}\}, q = 0.1), \end{aligned}$$

where the quantiles only consider perfectly-matched (PM) guides. We use m_g and s_g to parameterize a location scale normalization:

$$y_{gi}^{\wedge} \triangleq \frac{y_{gi} - m_g}{s_g} \quad (1)$$

where $y_{gi}^{\wedge}$ is the normalized LFC. We use normalized LFC values to train our model to encourage gene-agnostic predictions. As such, our model outputs normalized estimates. We assess performance in the original space via eq. (1)'s inverse. Equation (1) defines the variable $y_{gi}^{\wedge}$, which is the normalized LFC. m_g is the mean of variable g across all samples in the dataset, and s_g is the standard deviation of variable g across all samples in the dataset.

To compute $y_{gi}^{\wedge}$, we subtract the mean m_g from the original value y_{gi} , and then divide by the standard deviation s_g . This operation standardizes the variable g such that it has a mean of 0 and a standard deviation of 1 across the dataset.

Our model uses guide sequence, target sequence, and additional non-sequence features (FIG. 5). We one-hot encoded guide and target sequence. The sequence input to the convolutional neural network (CNN) consists of 23 nt target and gRNA sequences with 2 nt of upstream and downstream target context. We also one-hot encode flanking target sequence context and set the guide sequence encoding at context positions to zero. We treat the encoded sequences as a two-dimensional image (four nucleotides x sequence length) with two channels (guide and target). Our neural network layers are:

1. 2D convolution with 32 4 x 4 filters and ReLU activations
2. 2D convolution with 32 4 x 4 filters and ReLU activations
3. 2D max pooling with a 1 x 2 pool size
4. Flatten to vector
5. Dropout with 0.25 dropout rate
6. Concatenation with a vector of non-sequence features
7. Dense layer with 128 sigmoid outputs (hidden)
8. Dropout with 0.1 dropout rate
9. Dense layer with 32 sigmoid outputs (hidden)
10. Dropout with 0.1 dropout rate
11. Linear output layer with scalar output

In addition to the framework discussed above, we consider one or more of the following non-sequence features, which may be input as a vector at the first dense layer:

- i. target accessibility
 - 3. log unpaired
 - 4. log unpaired at position 11
 - 5. log unpaired at position 19
 - 6. log unpaired at position 25
 - ii. hybridization MFE
 - 7. positions 1 and 23
 - 8. positions 3 and 12
 - 9. positions 15 and 9
 - iii. guide MFE
 - iv. target location within transcript (encoded per Wei et al. (2021))
 - v. junction proximity
 - 10. $\log_{10}(\text{5p distance to junction} + 1)$
 - 11. $\log_{10}(\text{3p distance to junction} + 1)$
 - vi. guide secondary structure
 - 12. presence of direct repeat stem loop (a Boolean variable)
 - 13. presence of a g-quadruplex (a Boolean variable)

Model Training

With this feature set or a subset thereof (e.g. sequence only), we train TIGER to predict normalized LFC values using a log-hyperbolic-cosine loss. We perform cross-validation with a $1e-3$ learning rate and a 2048 batch size. We perform early-stopping on the validation set's loss.

Validation Strategies

We consider three validation folding strategies, which Fig. 6 summarizes graphically. The first creates sixteen folds along gene boundaries. In particular, it holds out all guides targeting a single gene for testing while training on guides for the remaining fifteen genes in our screen. This strategy is the most difficult to perform well on as it requires transcriptome-wide generalization. The second randomly assigns target sites into K (10 in our case) approximately equal sized folds. Folding by target site places a target sequence and all its guides (matched and mismatched) into the same fold, thus ensuring a target sequence cannot simultaneously appear

both in training and testing sets; enforcing this is important since active perfectly matched (PM) and singly mismatched (SM) guides that share a target site can convey significant information about one another. For example, training on an active PM guide and its associated target sequence and testing on a highly active SM guide targeting the same sequence is akin to testing on the training set (i.e., cheating). For this reason, the final validation strategy, which randomly splits guides into K (10 in our case) folds regardless of the target sequence, is the easiest to perform well on because it does not make such an enforcement.

Model Assessment

We use model pre-validation (Höfling and Tibshirani, 2008) to assess model performance. After assigning folds, we aggregate predictions from each held-out fold during cross-validation such that we have a held-out prediction for every data point. It is upon these aggregated predictions we compute performance metrics.

Example 5: SHapley Additive exPlanations (SHAP) analysis of feature importance

To determine TIGER's learned gRNA design rules, we performed a 10-fold CV of our combined model (TIGER_{combined}) with target site-level CV. For each holdout, we collected SHapley Additive exPlanations (SHAP)³⁹ values for sequence and non-sequence features. For sequence features of PM gRNAs, we observed a strong contribution of G and C nucleotides in the seed region (nucleotides 15-21) of the gRNA (Supplementary Fig. 4a, b) reflecting the local importance of G and C nucleotides^{7,8}. Similarly, we examined Pearson correlations and SHAP values for the gRNA substitutions (Supplementary Fig. 4c-d). We found that the CNN model correctly learned the increased importance of single-mismatches in the seed-region including the differential contribution of GU mismatched base pairing. We also examined SHAP values for non-sequence features for PM guides alone or all PM plus mismatched guides (Supplementary Fig. 4e-f). Among non-sequence features, we found that RNA-RNA hybridization (as in target site accessibility, gRNA-RNA hybridization MFE, and gRNA folding MFE) had the largest contributions to model predictions, consistent with our earlier findings (FIG. 3C).

Example 6: TIGER consistently predicts highly active gRNAs

Next we sought to test the generalizability of our TIGER_{combined} model at scale. We predicted 8 high-efficacy gRNAs for 5,166 genes and performed proliferation screens in two

different cell lines (HEK293FT and HAP1) (FIG. 3F, Supplementary Data 5-8). We noticed high consistency between gRNAs that target the same gene (FIG. 3G). Extending this analysis, we found that TIGER_{combined} correctly predicted active gRNAs (defined as $\log_2(\text{fold-change}) < -1$) in the HEK293 and HAP1 screens for 91% and 95%, respectively, of the chosen gRNAs (FIG. 3H), highlighting the robustness and generalizability of our on-target model across cell lines and thousands of unseen genes.

Having confirmed robust targeting, we next assessed whether RNA-targeting *RfxCas13d* CRISPR screens could discriminate essential genes from non-essential genes. Among the 5,166 target genes included in these screens, we embedded a set of 1,082 common essential genes and 458 non-essential genes based on DepMap classifications (see Methods). We found that our TIGER_{combined} model could successfully discriminate essential genes against control genes (FIG. 3I); AUROC 0.86 and 0.95 in HEK293FT and HAP1 cells). Gene depletion in HAP1 cells was generally more pronounced compared to that in HEK293FT cells (Supplementary Fig. 5a).

Previous reports have suggested that cell fitness (and therefore cell proliferation) may be affected by non-specific collateral *RfxCas13d* activity as a function of target gene expression^{40,41}. It is unclear, however, if collateral activity affects cell fitness in a controlled setting with single integration of *RfxCas13d* effector protein and gRNA expression cassettes. To explore this relationship, we analyzed gene dropout across a large span of expression levels (1 to >1,000 TPM). For common essential genes, we also observed a known dependency of gene expression and gene depletion^{42,43}, which was much reduced for the non-essential gene group (Supplementary Fig. 5b). For non-essential genes, we saw no evidence of dropout as a function of gene expression in HEK293FT. For the haploid HAP1 cells, we noticed that a relationship between gene expression and control gene depletion was mainly driven by genes expressed >100 transcripts per million (TPM). It is worth noting that only ~5% of 19,177 protein coding genes are expressed at levels above 100 TPM in human cell lines ($n = 1,377$ CCLE cell lines) and the majority of these genes are classified as common essential genes in DepMap. Non-specific collateral activity may only affect a small subset (< 6%) of very highly expressed genes (> 100TPM) in sensitive cells with single-integration Cas13d screens. Despite this, we found that HAP1 cells showed higher sensitivity (AUROC) for the identification of essential genes (FIG. 3I).

Taken together, we find that TIGER_{combined} gRNA predictions are robust and generalizable across thousands of genes in multiple cell lines, and that undesired viability defects based on target gene expression may be cell type dependent and usually only occur for a subset of highly expressed genes.

5

Example 7: Predicting off-target performance and titrating gene knockdown effects

Although multiple groups have developed predictive models of on-target Cas13d activity, there has been comparatively little work on off-target activity and no predictive models exist for Cas13d. Similarly, for CRISPR-Cas9 guide RNAs, nearly all deep learning approaches focus on
10 predicting on-activity and do not include separate inputs for gRNA and target site sequences^{26,29–36,44–46}. TIGER's architecture easily accommodates mismatches between target and gRNA (FIG. 3A). In addition to the ability to predict on-target efficacy for perfect match gRNA when trained on perfect match and mismatched data (TIGER_{combined}), we sought to extend the usability of TIGER to enable precise off-target predictions via target site-level CV (Supplementary Note 1).
15 This validation strategy avoids PM and SM guides for the same target appearing in training and validation.

When predicting changes in abundance for mismatch variant gRNAs, we find that the correlation between TIGER's predictions and observed values decreases as the number of mismatches and the distance between them increases (FIG. 4A, Supplementary Figure 6a). We
20 wondered whether this might be due in part to the variability in effect size between different PM gRNAs (FIG. 2D, FIG. 2E). To test this, we instead measured the difference in the predicted gRNA abundance between variant gRNAs and their cognate PM gRNA (FIG. 4B, Supplementary Figure 6b). Here, we find that the correlation no longer decreases with increasing mismatches, suggesting that when we explicitly account for the variability in PM gRNAs,
25 TIGER is able to predict the effect of different mismatches. As we did for PM gRNAs, we computed four performance metrics for each category of mismatch variant gRNAs (FIG. 4C, FIG. 4D, Supplementary Fig. 6c). This time, however, we compared our TIGER_{combined} model (trained on PM and mismatches) to a TIGER_{off-target} model (trained on mismatched gRNAs only). As is the case with on-target prediction, we find that the combined model is superior
30 (Supplementary Fig. 6d). Using an independent pooled screen (surface protein expression)⁷, which also contains gRNAs with single and double mismatch variants, we further confirmed our

model's ability to predict guide RNA efficacy after being trained on the survival screen dataset (Supplementary Fig. 6c).

Given that TIGER can predict a PM gRNA's efficacy and how this efficacy changes when mismatches are introduced, we can both identify mismatched target sites with off-target activity and engineer mismatches to precisely reduce gRNA efficacy and titer knockdown (FIG. 4E). To this end, we defined the *efficacy ratio* as the ratio of the fold-change of a single-mismatch gRNA to the fold-change of its PM cognate gRNA. Using target site-level CV, we compared predicted and observed relative gRNA activity for the 23 single mismatch gRNA variants designed for each individual target site. We found a high correlation between predicted and observed relative activities (median $r = 0.76$, median $\rho = 0.74$, $n = 324$ active target sites with PM $\log_2$ fold change < -1) (FIG. 4F). We binned all observed and predicted efficacy ratios for single-mismatch gRNAs with active cognate PM gRNAs into quintiles to compute a confusion matrix (FIG. 4G, Supplementary Data 9). For each quintile, we found that single mismatch gRNAs were most often correctly classified. In particular, TIGER achieves the best performance at the extremes (0-20 and 80-100) and can determine with high accuracy those single-mismatches that minimally impact or maximally disrupt activity.

Lastly, we sought to test the generalizability of our TIGER off-target model for gene essentiality titration across thousands of genes and target sites. Specifically, we predicted 4 high efficacy gRNAs for 1,082 common essential genes and designed ten gRNA variants with single nucleotide mismatches for each PM gRNA target site ($n = 47,608$ PM and SM gRNAs). Using a pooled proliferation screen in a different cell line (HAP1), we measured relative activity loss compared to the cognate PM gRNA via depletion (FIG. 4H, Supplementary Data 10-12). We found a high correlation between predicted and observed relative activities (median $r = 0.83$, median $\rho = 0.81$, $n = 3,161$ target sites with cognate PM $< -1 \log_2\text{FC}$;) (FIG. 4I, Supplementary Figure 6f, Supplementary Data 13). In this independent screen, we found strong agreement across quintiles (FIG. 4J). This suggests that our model is able to predict gRNA variants with a defined relative activity with high accuracy for unseen target sites and generalizes across cell lines.

Example 8: Validation of TIGER model

We validated our TIGER model in an investigation of whether modifying the guide RNA sequence could diminish the recognition of the target RNA transcript and hinder the conformational changes necessary for cis and trans RNA cleavage. We performed a combinatorial RfxCas13d screen in A375 cells targeting HDAC1 and/or HDAC2 using ~6,000 pairs of gRNAs. These two genes form a synthetic lethal gene pair, where targeting both genes (but not either singly) results in lethality and drop-out, as shown in a recent study using DNA-targeting perturbations. We took advantage of our TIGER (Targeted Inhibition of Gene Expression via gRNA design) model to design titrated gRNAs with single mismatches in the parent sequence that alter the expected transcript knockdown activity over a range of partial knockdowns (FIG. 10A).

For the HDAC1 and HDAC2 synthetic lethality titration screen, we first selected 20 pairs of target sites for each gene. For each target site (in HDAC1 or HDAC2), we designed 1 perfect-match gRNA (i.e. the target site itself) and 10 additional single-mismatch variant gRNAs with titrated transcript knockdown activities ranging from 5-95% (incrementing at every decile) that were designed using the TIGER model library. Each target site in HDAC1 was paired with each target site in HDAC2 in a dual gRNA array. For each target site pair, we used 11 HDAC1 titration gRNAs $\times$ 11 HDAC2 titration gRNAs $\times$ 2 positions (HDAC1-HDAC2 gRNA order or HDAC2-HDAC1 gRNA order), which yielded 242 combinations for each target site pair. Given the 20 pairs of target sites, we generated a total of 4,839 HDAC1/HDAC2 arrays. We also paired each gRNA targeting HDAC1 or HDAC2 with a non-targeting (NT) gRNA: This yielded an additional 880 arrays = 11 gRNAs per target site $\times$ 40 total target sites (20 in HDAC1, 20 in HDAC2) $\times$ 2 positions (HDAC1/2-NT gRNA order or NT-HDAC1/2 gRNA order). We also included 492 arrays with NT gRNAs in both positions. In total, the library for Screen 2 consisted of 6,211 arrays. Each gRNA was flanked with constant regions (for PCR amplification and Gibson cloning) and synthesized as 148mer single-stranded oligonucleotides (Twist Biosciences).

Targeting a single gene from HDAC1 and HDAC2 should decrease cell fitness due to the high expression levels of these genes —147 TPM for HDAC1 and 130 TPM for HDAC2 — if unintended collateral activity occurs. Consistent with our previous data, we found that targeting only one gene, HDAC1 or HDAC2, from this synthetic lethal gene pair does not cause significant gRNA depletion in our lentiviral, low-copy RfxCas13d cell lines (FIG. 10B). As

expected, targeting both genes in the same cell resulted in gRNA depletion. Further, we found that the level of depletion from the screen for all titration pairs correlates with the predicted activity for partial knockdown gRNAs (Data not shown).

To validate that the lack of collateral activity in this screen is a result of low copy number of RfxCas13d, we performed a competition assay and targeted HDAC1 and/or HDAC2 using either transfected or transduced RfxCas13d. As before, we observed depletion when both HDAC1 and HDAC2 were targeted, regardless of how RfxCas13d was delivered. However, when targeting HDAC1 alone, we only observed depletion of cells when RfxCas13d was delivered via transfection. Notably, targeting HDAC1 using perfect match gRNAs resulted in a significant decrease in GFP in the transfection condition. Even with imperfect match gRNAs (predicted knockdown of 55%), we observed a slight decrease in GFP signal in the transfection condition. This confirms that modulating the rate of target recognition with designed mismatches in the gRNA not only alters cis RNA cleavage but also trans collateral cleavage.

Sydney K. Hart, Hans-Hermann Wessels, Alejandro Méndez-Mancilla, Simon Müller, Gediminas Drabavicius, Olivia Choi, Neville E. Sanjana, Low copy CRISPR-Cas13d mitigates collateral RNA cleavage, bioRxiv 2024.05.13.594039 is incorporated herein by reference in its entirety.

References

1. Abudayyeh, O. O. et al. C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science* 353, (2016).
2. Abudayyeh, O. O. et al. RNA targeting with CRISPR-Cas13. *Nature* 550, 280–284
5 (2017).
3. Smargon, A. A. et al. Cas13b Is a Type VI-B CRISPR-Associated RNA-Guided RNase Differentially Regulated by Accessory Proteins Csx27 and Csx28. *Mol. Cell* 65, 618–630 (2017).
4. Konermann, S. et al. Transcriptome Engineering with RNA-Targeting Article
10 Transcriptome Engineering with RNA-Targeting. *Cell* 173, 1–12 (2018).
5. Yan, W. X. et al. Cas13d Is a Compact RNA-Targeting Type VI CRISPR Effector Positively Modulated by a WYL-Domain-Containing Accessory Protein. *Mol. Cell* 70, 327–339 (2018).
6. Smargon, A. A., Shi, Y. J. & Yeo, G. W. RNA-targeting CRISPR systems from
15 metagenomic discovery to transcriptomic engineering. *Nat. Cell Biol.* 22, 143–150 (2020).
7. Wessels, H. H. et al. Massively parallel Cas13 screens reveal principles for guide RNA design. *Nat. Biotechnol.* 38, 722–727 (2020).
8. Wei, J. et al. Deep learning and CRISPR-Cas13d ortholog discovery for optimized RNA targeting. *bioRxiv* (2022). doi:10.1101/2021.09.14.460134
9. Cheng, X. et al. Modeling CRISPR-Cas13d on-target and off-target effects. *bioRxiv*
20 (2021). doi:10.1101/2021.09.02.458773
10. Metsky, H. C. et al. Designing sensitive viral diagnostics with machine learning. *Nat. Biotechnol.* 40, 1123–1131 (2022).
11. Tambe, A., East-seletsky, A., Knott, G. J., Connell, M. R. O. & Doudna, J. A. RNA
25 Binding and HEPN-Nuclease Activation Are Decoupled in CRISPR-Cas13a. *Cell Rep.* 24, 1025–1036 (2018).
12. Powell, J. E. et al. Targeted gene silencing in the nervous system with CRISPR-Cas13. *Sci. Adv.* 8, 1–12 (2022).
13. Morelli, K. H. et al. An RNA-targeting CRISPR–Cas13d system alleviates disease-
30 related phenotypes in Huntington’s disease models. *Nat. Neurosci.* 26, 27–38 (2023).

14. Méndez-Mancilla, A. et al. Chemically modified guide RNAs enhance CRISPR-Cas13 knockdown in human cells. *Cell Chem. Biol.* 1–7 (2021). doi:10.1016/j.chembiol.2021.07.011
15. Rotolo, L. et al. Species-agnostic polymeric formulations for inhalable messenger RNA delivery to the lung. *Nat. Mater.* (2022). doi:10.1038/s41563-022-01404-0
- 5 16. Fan, N. et al. Hierarchical self-uncloaking CRISPR-Cas13a–customized RNA nanococoons for spatial-controlled genome editing and precise cancer therapy. *Sci. Adv.* 8, 1–18 (2022).
17. Guo, Y. et al. Specific knockdown of Htra2 by CRISPR-CasRx prevents acquired sensorineural hearing loss in mice. *Mol. Ther. - Nucleic Acids* 28, 643–655 (2022).
- 10 18. Nasim, M. T. et al. Stoichiometric imbalance in the receptor complex contributes to dysfunctional BMPR-II mediated signalling in pulmonary arterial hypertension. *Hum. Mol. Genet.* 17, 1683–1694 (2008).
19. Gurdon, J. B. & Bourillot, P. Y. Morphogen gradient interpretation. *Nature* 413, 797–803 (2001).
- 15 20. McHugh, C. A. et al. The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. *Nature* 521, 232–236 (2015).
21. Fehrmann, R. S. N. et al. Gene expression analysis identifies global gene dosage sensitivity in cancer. *Nat. Genet.* 47, 115–125 (2015).
22. Collins, R. L. et al. A cross-disorder dosage sensitivity map of the human genome. *Cell* 20 185, 3041–3055.e25 (2022).
23. Patwardhan, R. P. et al. High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat. Biotechnol.* 27, 1173–1175 (2009).
24. Gossen, M. & Bujard, H. Tight control of gene expression in mammalian cells by tetracycline-responsive promoters. *Proc. Natl. Acad. Sci. U. S. A.* 89, 5547–5551 (1992).
- 25 25. Michaels, Y. S. et al. Precise tuning of gene expression levels in mammalian cells. *Nat. Commun.* 10, 1–12 (2019).
26. Jost, M. et al. Titrating gene expression using libraries of systematically attenuated CRISPR guide RNAs. *Nat. Biotechnol.* 38, 355–364 (2020).
27. Bintu, L. et al. Dynamics of epigenetic regulation at the single-cell level. *Science* (80-.). 30 351, 720–724 (2016).

28. Zhang, C. et al. Structural Basis for the RNA-Guided Ribonuclease Activity of CRISPR-Cas13d. *Cell* 175, 212–223 (2018).
29. Charlier, J., Nadon, R. & Makarenkov, V. Accurate deep learning off-target prediction with novel sgRNA-DNA sequence encoding in CRISPR-Cas9 gene editing. *Bioinformatics* 37, 2299–2307 (2021).
30. Kim, H. K. et al. SpCas9 activity prediction by DeepSpCas9, a deep learning-based model with high generalization performance. *Sci. Adv.* 5, 1–10 (2019).
31. Lin, J. & Wong, K. C. Off-target predictions in CRISPR-Cas9 gene editing using deep learning. *Bioinformatics* 34, i656–i663 (2018).
32. Lin, J., Zhang, Z., Zhang, S., Chen, J. & Wong, K. C. CRISPR-Net: A Recurrent Convolutional Network Quantifies CRISPR Off-Target Activities with Mismatches and Indels. *Adv. Sci.* 7, 1–17 (2020).
33. Liu, Q., Cheng, X., Liu, G., Li, B. & Liu, X. Deep learning improves the ability of sgRNA off-target propensity prediction. *BMC Bioinformatics* 21, 1–15 (2020).
34. Luo, J., Chen, W., Xue, L. & Tang, B. Prediction of activity and specificity of CRISPR-Cpf1 using convolutional deep learning neural networks. *BMC Bioinformatics* 20, 1–10 (2019).
35. Niu, R., Peng, J., Zhang, Z. & Shang, X. R-CRISPR: A deep learning network to predict off-target activities with mismatch, insertion and deletion in CRISPR-Cas9 system. *Genes (Basel)*. 12, (2021).
36. Zhang, G., Zeng, T., Dai, Z. & Dai, X. Prediction of CRISPR/Cas9 single guide RNA cleavage efficiency and specificity by attention-based convolutional neural networks. *Comput. Struct. Biotechnol. J.* 19, 1445–1457 (2021).
37. LeCun, Y. et al. Backpropagation applied to digit recognition. *Neural computation* 1, 541–551 (1989).
38. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.* 25, 1097–1105 (2012).
39. Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *Adv. Neural Inf. Process. Syst.* 30, (2017).
40. Shi, P. et al. RNA-guided cell targeting with CRISPR/RfxCas13d collateral activity in human cells. *bioRxiv* (2021). doi:10.1101/2021.11.30.470032

41. Kelley, C. P., Haerle, M. C. & Wang, E. T. Negative autoregulation mitigates collateral RNase activity of repeat-targeting CRISPR-Cas13d in mammalian cells. *Cell Rep.* 40, 111226 (2022).
42. Wang, T. et al. Identification and characterization of essential genes in the human
5 genome. *Science* (80-.). 350, 1096–1101 (2015).
43. Hart, T., Brown, K. R., Sircoulomb, F., Rottapel, R. & Moffat, J. Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. *Mol. Syst. Biol.* 10, 1–16 (2014).
44. Kim, H. K. et al. High-throughput analysis of the activities of xCas9, SpCas9-NG and
10 SpCas9 at matched and mismatched target sequences in human cells. *Nat. Biomed. Eng.* 4, 111–124 (2020).
45. Kim, N. et al. Prediction of the sequence-specific cleavage activity of Cas9 variants. *Nat. Biotechnol.* 38, 1328–1336 (2020).
46. Xiang, X. et al. Enhancing CRISPR-Cas9 gRNA efficiency prediction by data integration
15 and deep learning. *Nat. Commun.* 12, 1–9 (2021).
47. Hu, W. et al. Single-base precision design of CRISPR-Cas13b enables systematic silencing of oncogenic fusions. *bioRxiv* 2022.06.22.497105 (2022).
48. Raj, A. & van Oudenaarden, A. Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences. *Cell* 135, 216–226 (2008).
49. Lai, E. C., Tomancak, P., Williams, R. W. & Rubin, G. M. Computational identification
20 of *Drosophila* microRNA genes. *Genome Biol.* 4, R42 (2003).
50. Stoeger, T., Battich, N. & Pelkmans, L. Passive Noise Filtering by Cellular Compartmentalization. *Cell* 164, 1151–1161 (2016).
51. Noviello, G., Gjaltema, R. A. F. & Schulz, E. G. CasTuner: a degron and CRISPR/Cas-
25 based toolkit for analog tuning of endogenous gene expression. *bioRxiv* 2022.10.05.511019 (2022).
52. Lensch, S. et al. Dynamic spreading of chromatin-mediated gene silencing and reactivation between neighboring genes in single cells. *Elife* 11, (2022).
53. Hart, T. et al. High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-
30 Specific Cancer Liabilities. *Cell* 163, 1515–1526 (2015).

54. Gerstberger, S., Hafner, M. & Tuschl, T. A census of human RNA-binding proteins. *Nat. Rev. Genet.* 15, 829–845 (2014).
55. Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. a & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.* 10, 252–263 (2009).
56. Chen, S. et al. Genome-wide CRISPR screen in a mouse model of tumor growth and metastasis. *Cell* 160, 1246–1260 (2015).
57. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 550 (2014).
58. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 28, 882–883 (2012).
59. Sack, L. M., Davoli, T., Xu, Q., Li, M. Z. & Elledge, S. J. Sources of error in mammalian genetic screens. *G3 Genes, Genomes, Genet.* 6, 2781–2790 (2016).
60. Kolde, R., Laur, S., Adler, P. & Vilo, J. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics* 28, 573–580 (2012).
61. Lorenz, R. et al. ViennaRNA Package 2.0. *Algorithms Mol. Biol.* 6, 26 (2011).
62. Agarwal, V., Subtelny, A. O., Thiru, P., Ulitsky, I. & Bartel, D. P. Predicting microRNA targeting efficacy in *Drosophila*. *Genome Biol.* 19, 1–23 (2018).
63. Agarwal, V., Bell, G. W., Nam, J.-W. & Bartel, D. P. Predicting effective microRNA target sites in mammalian mRNAs. *Elife* 4, 1–38 (2015).
64. Krueger, J. & Rehmsmeier, M. RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic Acids Res.* 34, 451–454 (2006).
65. Kingma, D. P. & Ba, J. L. Adam: A method for stochastic optimization. *arXiv* 1–15 (2014).
66. Cho, K., van Merriënboer, B., Bahdanau, D. & Bengio, Y. On the properties of neural machine translation: Encoder–decoder approaches. *Proc. SSST 2014 - 8th Work. Syntax. Semant. Struct. Stat. Transl.* 103–111 (2014). doi:10.3115/v1/w14-4012
67. Steiger, J. H. Tests for comparing elements of a correlation matrix. *Psychol. Bull.* 87, 245–251 (1980).

68. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* 44, 837–845 (1988).
69. Sun, X. & Xu, W. Fast implementation of DeLong’s algorithm for comparing the areas
5 under correlated receiver operating characteristic curves. *IEEE Signal Process. Lett.* 21, 1389–1393 (2014).
70. Massey, F. J. J. The Kolmogorov-Smirnov Test for Goodness of Fit. *J. Am. Stat. Assoc.* 46, 68–78 (1951).
71. Hofling, H. and Tibshirani, R. A Study of Pre-Validation. *The Annals of Applied*
10 *Statistics.* 2(2), 643-664 (2008).

CLAIMS:

1. A method of generating a set of guide RNAs (gRNAs) capable of driving a series of discrete expression levels of a target gene in a cell population using CRISPR interference (CRISPRi) or CRISPR activation (CRISPRa), each gRNA comprising a spacer sequence, the method comprising:

(i) providing a first perfect match (PM) gRNA that comprises a spacer sequence that is 100% complementary to a target sequence within the gene;

(ii) providing a second mismatch gRNA that targets the gene, wherein the spacer sequence of the second gRNA comprises one or more mismatches with the target DNA sequence such that the CRISPRi or CRISPRa activity on the gene obtained using the second gRNA is intermediate between that obtained using the first gRNA and that obtained using a scrambled gRNA providing no CRISPRi or CRISPRa activity on the gene; and

(iii) providing a third indel mismatch gRNA that targets the gene, wherein the spacer sequence of the third sgRNA comprises one or more mismatches with the target DNA sequence such that the CRISPRi or CRISPRa activity on the gene obtained using the third gRNA is intermediate between that obtained using the second gRNA and that obtained using a scrambled gRNA providing no CRISPRi or CRISPRa activity on the gene; wherein the mismatches of the second and third gRNAs are selected according to the following rule:

(a) the CRISPRi or CRISPRa activity of the second gRNA is designed to be greater than that of the third gRNA based on the following positional relationships, wherein the positions correspond to the number of bases in the gRNA downstream to the DR sequence: $18 > 9 > 13 > 5 > 22$.

2. The method of claim 1, further comprising providing one or more additional mismatch gRNAs, wherein each of the one or more additional gRNAs provide CRISPRi or CRISPRa activity on the gene that is intermediate between that obtained using the third gRNA and that obtained using a scrambled gRNA providing no CRISPRi or CRISPRa activity on the gene, and wherein the mismatches with the target sequence of each of the one or more additional gRNAs are selected according to rule (a) of claim 1.

3. The method of claim 1 or 2, wherein each of the mismatch gRNAs is assigned to a predicted activity quintile using a computational model.

4. The method of claim 3, wherein the computational model comprises one or more of the following layers:

- i) 2D convolution with 32 4 x 4 filters and ReLU activations
- ii) 2D convolution with 32 4 x 4 filters and ReLU activations
- iii) 2D max pooling with a 1 x 2 pool size
- iv) Flatten to vector
- v) Dropout with 0.25 dropout rate
- vi) Concatenation with a vector of non-sequence features
- vii) Dense layer with 128 sigmoid outputs
- viii) Dropout with 0.1 dropout rate
- ix) Dense layer with 32 sigmoid outputs
- x) Dropout with 0.1 dropout rate
- xi) Linear output layer with scalar output;

wherein the non-sequence features comprise one or more of the following:

- a) target accessibility
- b) hybridization MFE
- c) guide MFE
- d) target location within transcript
- e) junction proximity
- f) guide secondary structure

wherein the computational model predicts normalized LFC values using a log-hyperbolic-cosine loss.

5. The method of any of claims 1 to 4, wherein the target gene is a mammalian gene.

6. The method of claim 5, wherein the mammalian gene is a human gene.

7. A set of single guide RNAs (sgRNAs) for obtaining a series of discrete expression levels of a target gene using CRISPRi or CRISPRa, comprising

(i) a first perfect match (PM) gRNA that comprises a spacer sequence that is 100% complementary to a target sequence within the gene;

(ii) a second mismatch gRNA that targets the gene, wherein the spacer sequence of the second gRNA comprises one or more mismatches with the target DNA sequence such that the CRISPRi or CRISPRa activity on the gene obtained using the second gRNA is intermediate between that obtained using the first gRNA and that obtained using a scrambled gRNA providing no CRISPRi or CRISPRa activity on the gene; and

(iii) a third indel mismatch gRNA that targets the gene, wherein the spacer sequence of the third sgRNA comprises one or more mismatches with the target DNA sequence such that the CRISPRi or CRISPRa activity on the gene obtained using the third gRNA is intermediate between that obtained using the second gRNA and that obtained using a scrambled gRNA providing no CRISPRi or CRISPRa activity on the gene; wherein the mismatches of the second and third gRNAs are selected according to the following rule:

(a) the CRISPRi or CRISPRa activity of the second gRNA is designed to be greater than that of the third gRNA based on the following positional relationships, wherein the positions correspond to the number of bases in the gRNA downstream to the DR sequence: $18 > 9 > 13 > 5 > 22$.

8. A method of selecting a gRNA for use in a Cas-based genome editing system capable of introducing a discreet level of expression of a target gene in a cell population, the method comprising:

transducing mammalian cells with a library of gRNAs, the library comprising a plurality of different gRNAs for one or more genes, the plurality of gRNAs comprising for each gene:

i) perfect-match (PM) gRNAs that are 100% complementary to a target sequence within the gene;

ii) mismatch gRNAs that contain 1, 2, or 3 nucleotide mismatches to the target sequence within the gene;

iii) indel gRNAs that contain 1 or 2 nucleotide insertions or deletions to the target sequence within the gene;

determining the activity of each gRNA in the library;

determining the relative activity of all mismatch gRNAs and indel gRNAs relative to their cognate PM gRNAs;

selecting the gRNA based on the discreet level of expression desired.

9. A method for precise modulation of target RNA, the method comprising providing a gRNA that allows a desired level of effector activity of a Cas13d enzyme when the gRNA and Cas13d enzyme are contacted with the target RNA to which the gRNA binds, wherein the gRNA is identified by the method of claim 8.

10. The method of claim 8 or 9, wherein the effector activity is RNA knockdown (CRISPRi).

11. The method of any one of claims 8 to 10, wherein the desired level of effector activity is 10, 20, 30, 40, 50, 60, 70, 80, 90, or 100% of native expression (or control).

12. The method according to any preceding claim, wherein the mismatch gRNAs comprise 2 or 3 nucleotide mismatches adjacent to each other.

The method according to any preceding claim, wherein the mismatch gRNAs comprise 2 or 3 nucleotide mismatches contain at least one perfectly matched nucleotide between them.

13. The method according to any preceding claim, wherein the indel gRNAs comprise 2 or 3 nucleotide insertions or deletions adjacent to each other.

14. The method according to any preceding claim, wherein the indel gRNAs comprise 2 or 3 nucleotide insertions or deletions contain at least one perfectly matched nucleotide between them.

15. The method according to any preceding claim, wherein the transduced mammalian cells further comprise a Cas13d nuclease.

16. The method according to any preceding claim, wherein the gRNAs are 18-30nt.

17. The method according to any preceding claim, wherein the CRISPRi activity of the mismatch gRNA is designed to be lesser than that of the perfect match gRNA based on the following positional relationships, wherein the positions correspond to the number of bases in the gRNA downstream to the DR sequence: $PM > 18 > 9 > 13 > 5 > 22$.

18. A method of predicting the efficacy of a subject gRNA for use in a Cas-based genome editing system, the method comprising:

analyzing the subject gRNA sequence with a computational model to predict efficacy of the subject gRNA, the computational model comprising one or more of the following layers:

- i) 2D convolution with 32 4 x 4 filters and ReLU activations
- ii) 2D convolution with 32 4 x 4 filters and ReLU activations

- iii) 2D max pooling with a 1 x 2 pool size
- iv) Flatten to vector
- v) Dropout with 0.25 dropout rate
- vi) Concatenation with a vector of non-sequence features
- vii) Dense layer with 128 sigmoid outputs
- viii) Dropout with 0.1 dropout rate
- ix) Dense layer with 32 sigmoid outputs
- x) Dropout with 0.1 dropout rate
- xi) Linear output layer with scalar output;

wherein the non-sequence features comprise one or more of the following:

- a) target accessibility
- b) hybridization MFE
- c) guide MFE
- d) target location within transcript
- e) junction proximity
- f) guide secondary structure

wherein the computational model predicts normalized LFC values using a log-hyperbolic-cosine loss.

19. A system for selecting gRNA for titrating expression of a target gene in a target cell from a multiplicity of candidate guides, the system comprising:

a memory; and

a processor coupled with the memory and configured to:

analyze the subject gRNA sequence with a computational model to predict efficacy of the subject gRNA, the computational model comprising one or more of the following layers:

- xii) 2D convolution with 32 4 x 4 filters and ReLU activations
- xiii) 2D convolution with 32 4 x 4 filters and ReLU activations
- xiv) 2D max pooling with a 1 x 2 pool size
- xv) Flatten to vector
- xvi) Dropout with 0.25 dropout rate
- xvii) Concatenation with a vector of non-sequence features
- xviii) Dense layer with 128 sigmoid outputs

- xix) Dropout with 0.1 dropout rate
- xx) Dense layer with 32 sigmoid outputs
- xxi) Dropout with 0.1 dropout rate
- xxii) Linear output layer with scalar output;

wherein the non-sequence features comprise one or more of the following:

- g) target accessibility
- h) hybridization MFE
- i) guide MFE
- j) target location within transcript
- k) junction proximity
- l) guide secondary structure

wherein the computational model predicts normalized LFC values using a log-hyperbolic-cosine loss.

FIG. 1A

Systematic guide RNA perturbations

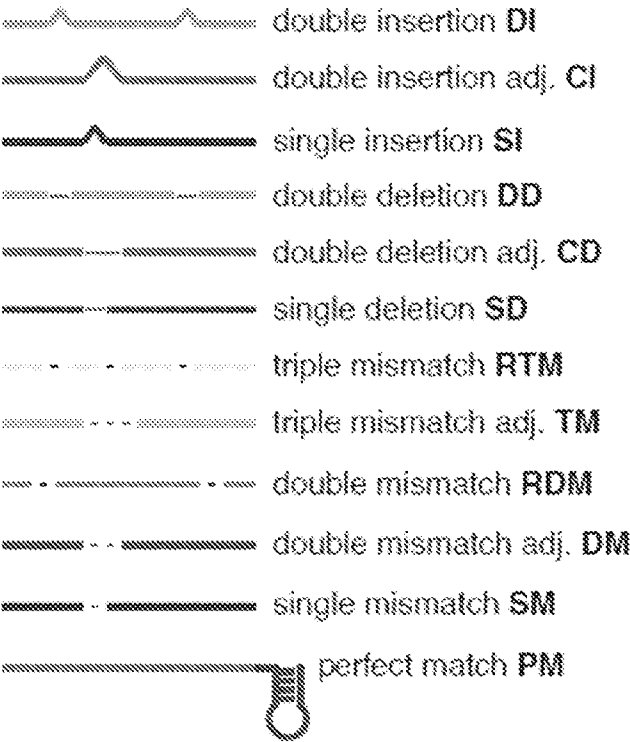


FIG. 1B

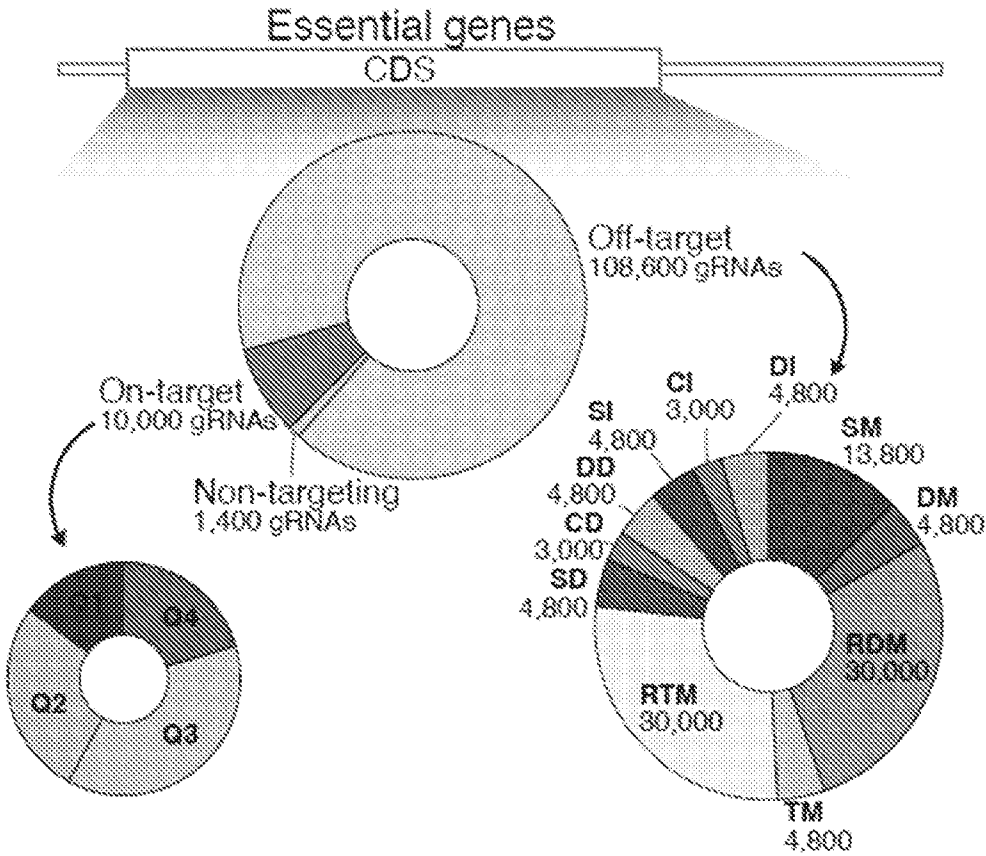


FIG. 1C

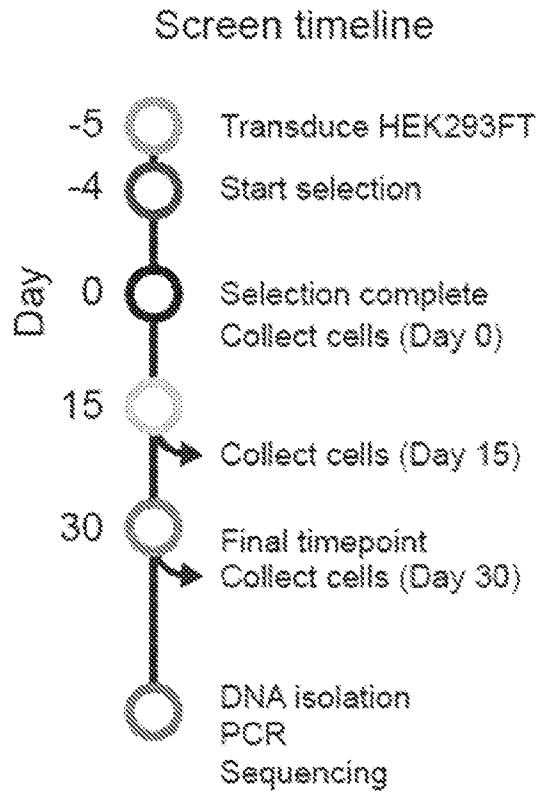


FIG. 1D

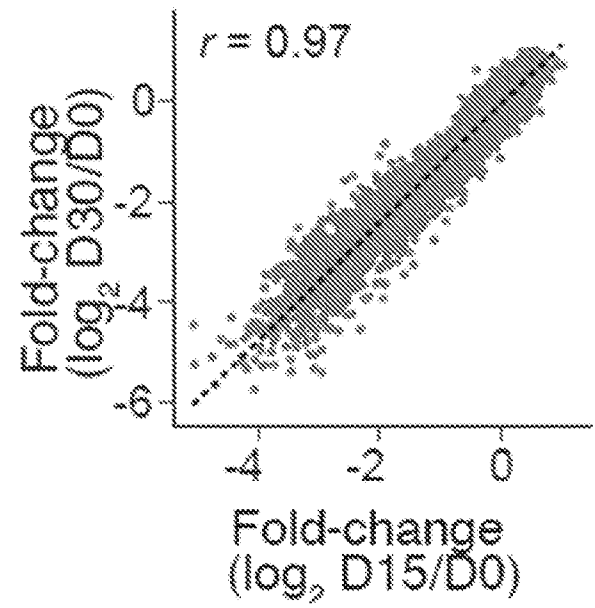


FIG. 1E

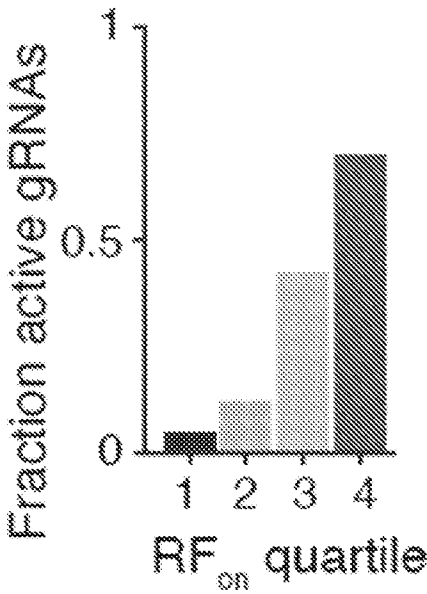


FIG. 1F

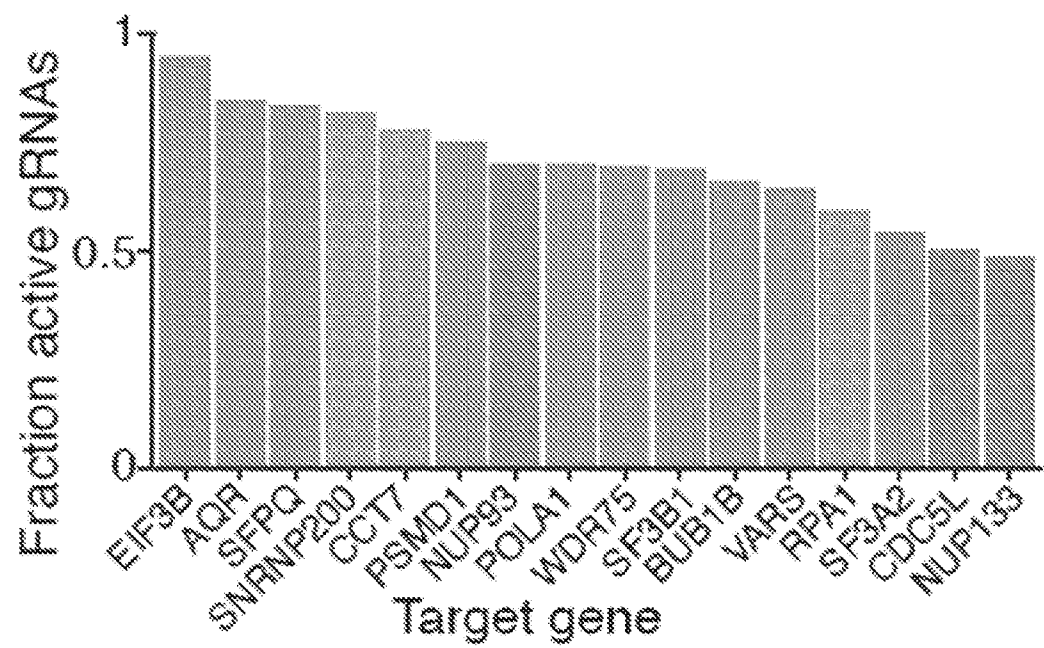


FIG. 1G

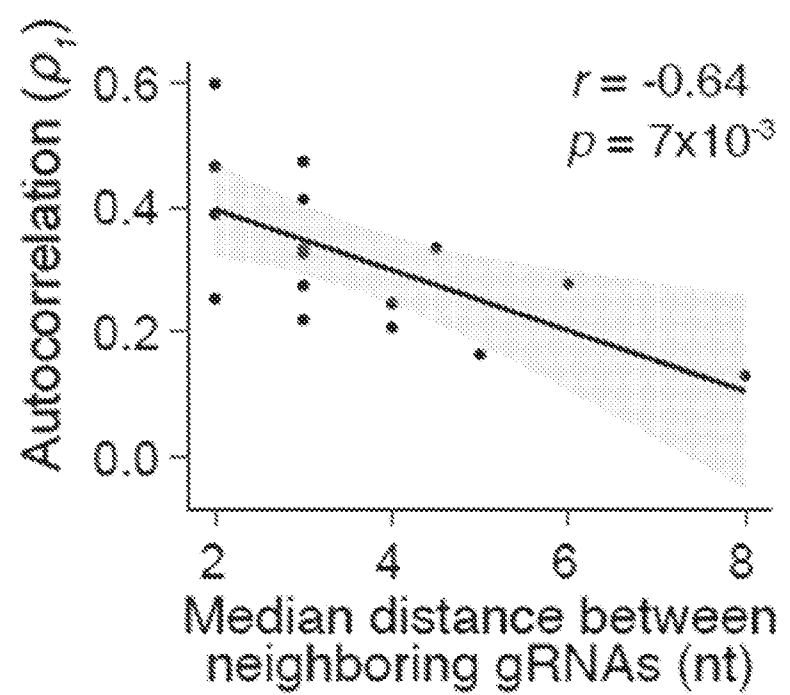


FIG. 1H

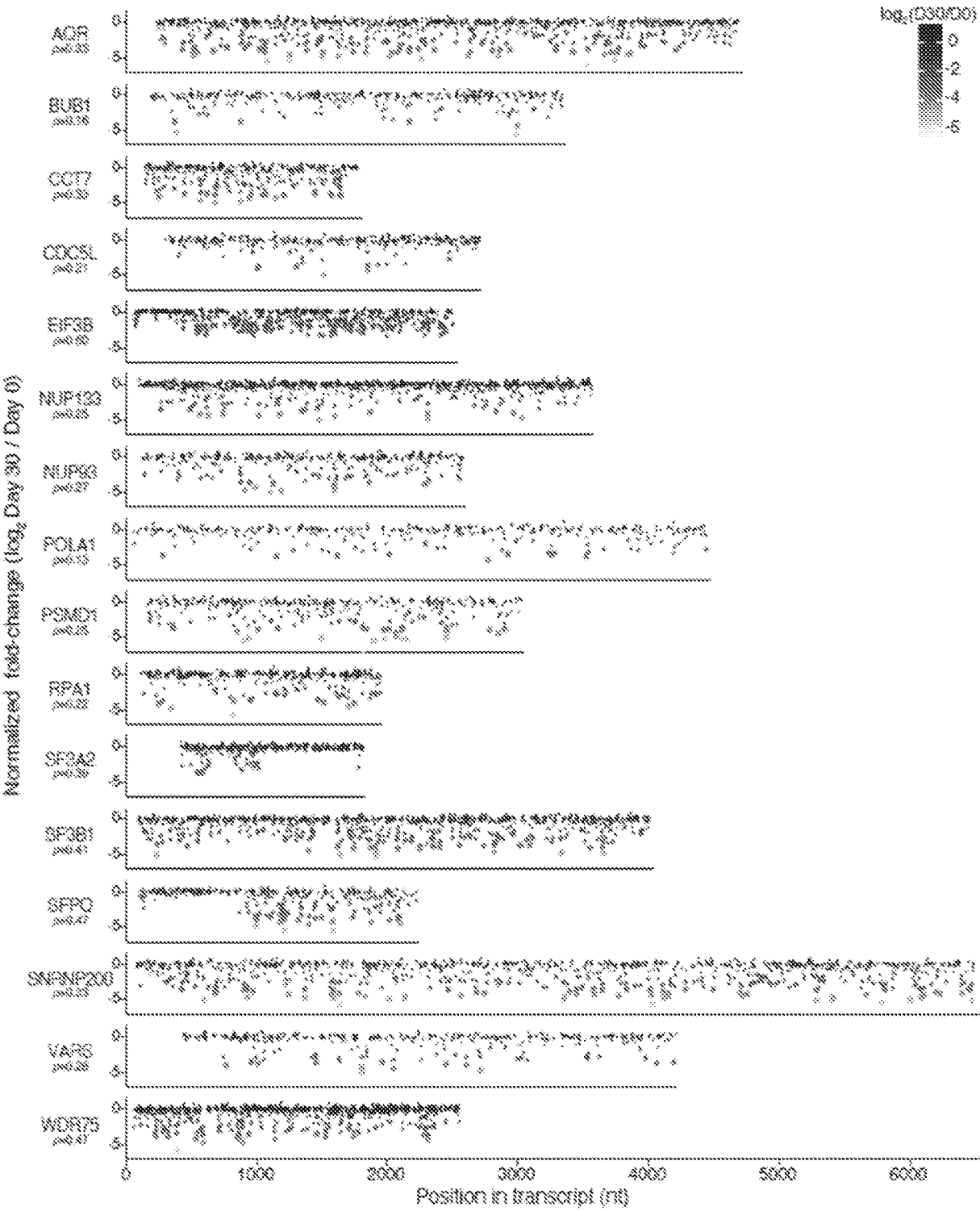


FIG. 2A

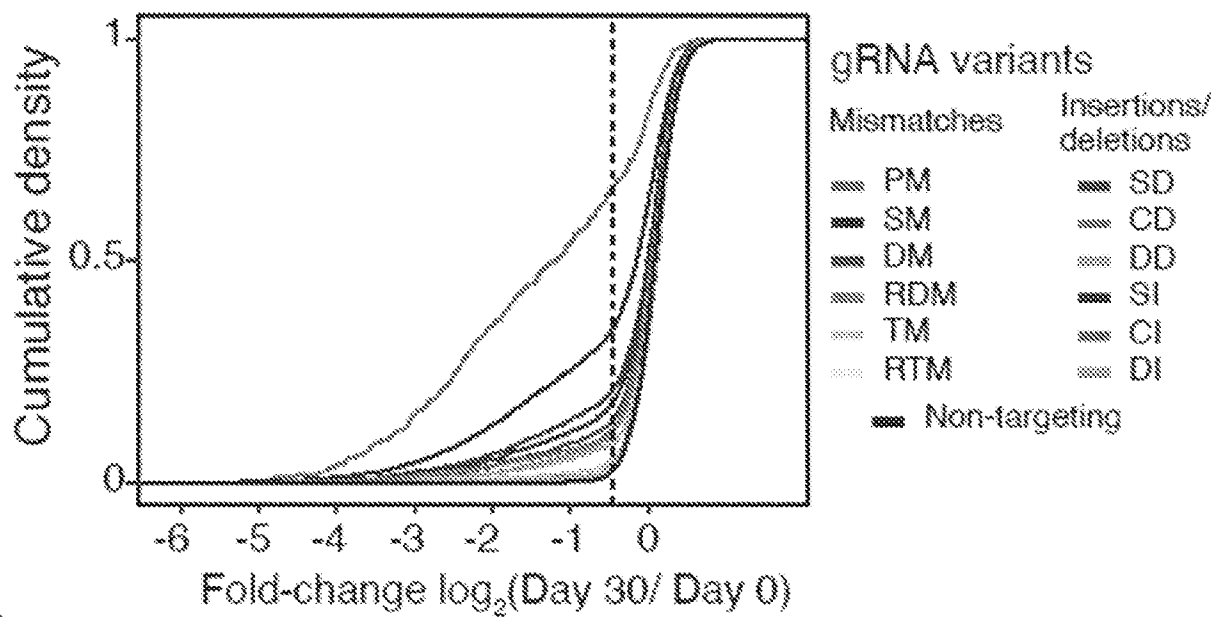


FIG. 2B

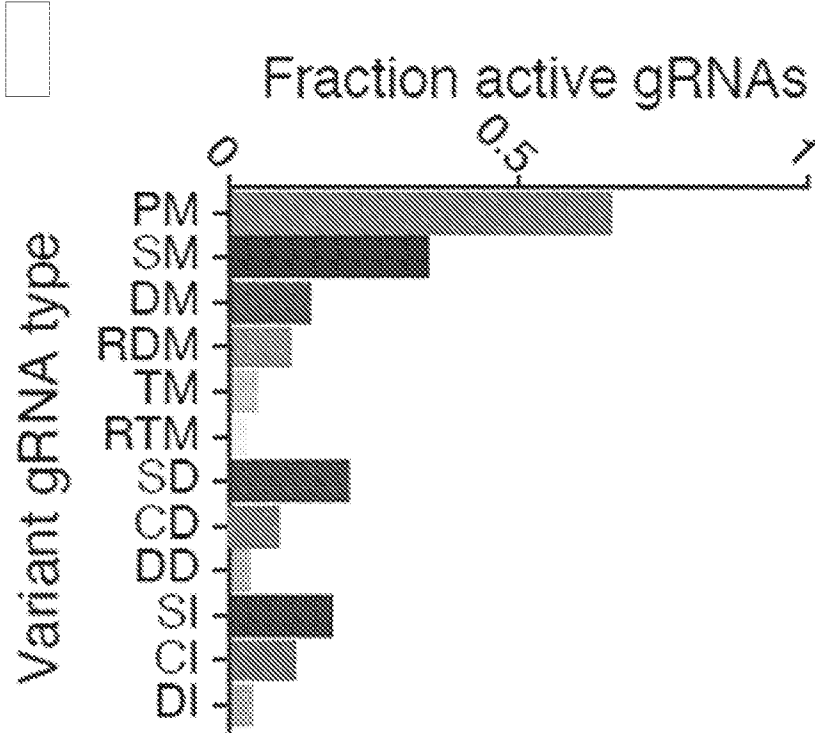


FIG. 2C

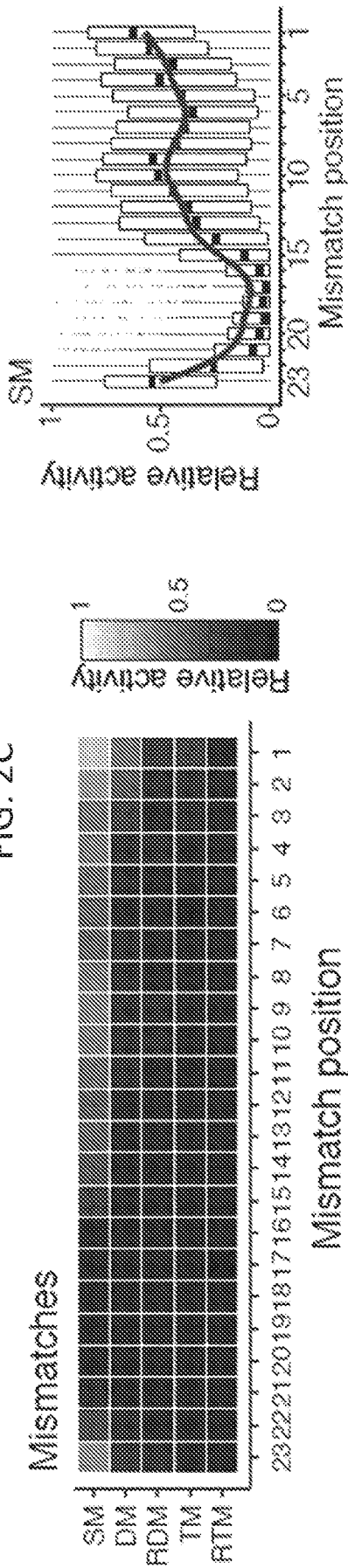


FIG. 2D

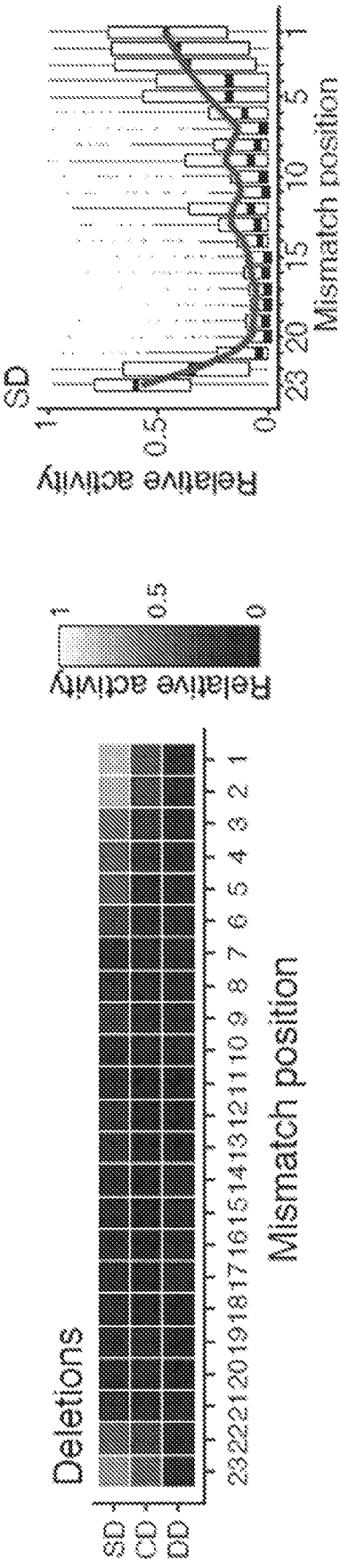


FIG. 2E

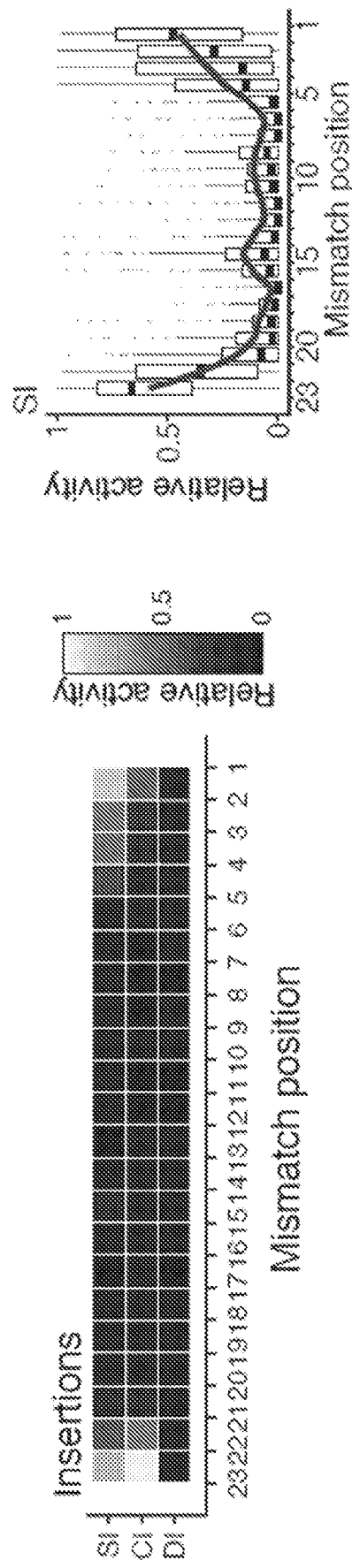


Fig. 2G

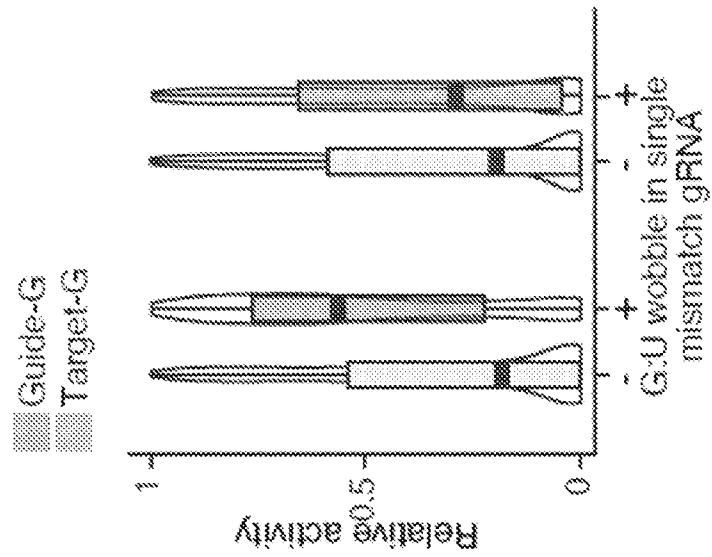


FIG. 2F

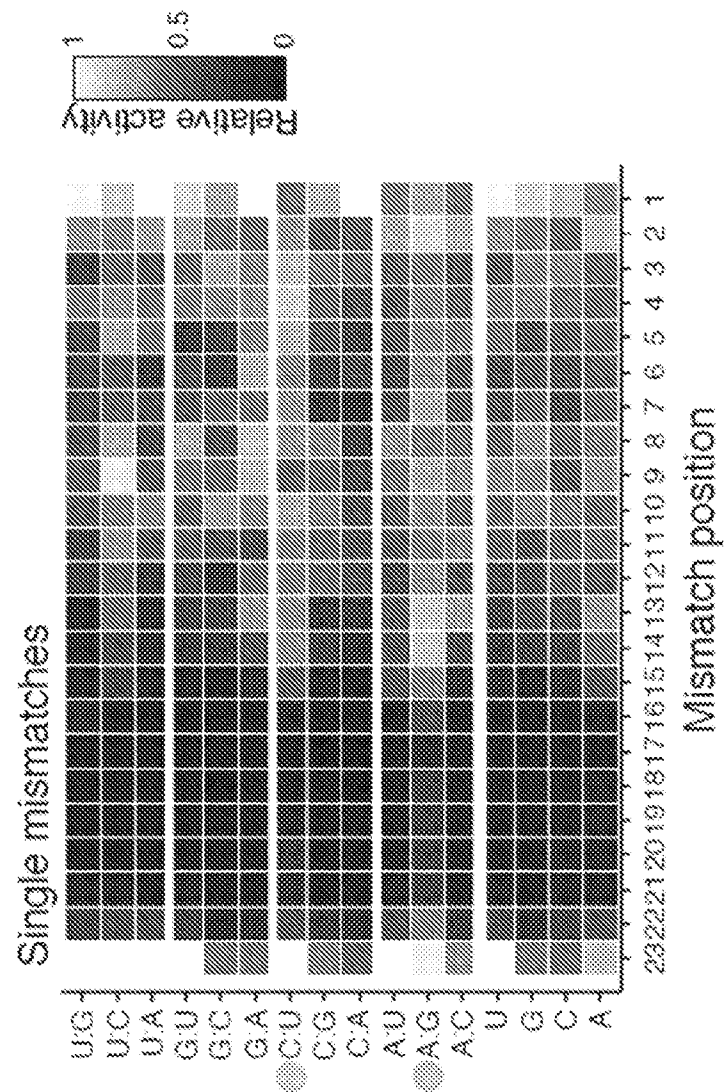


FIG. 3A

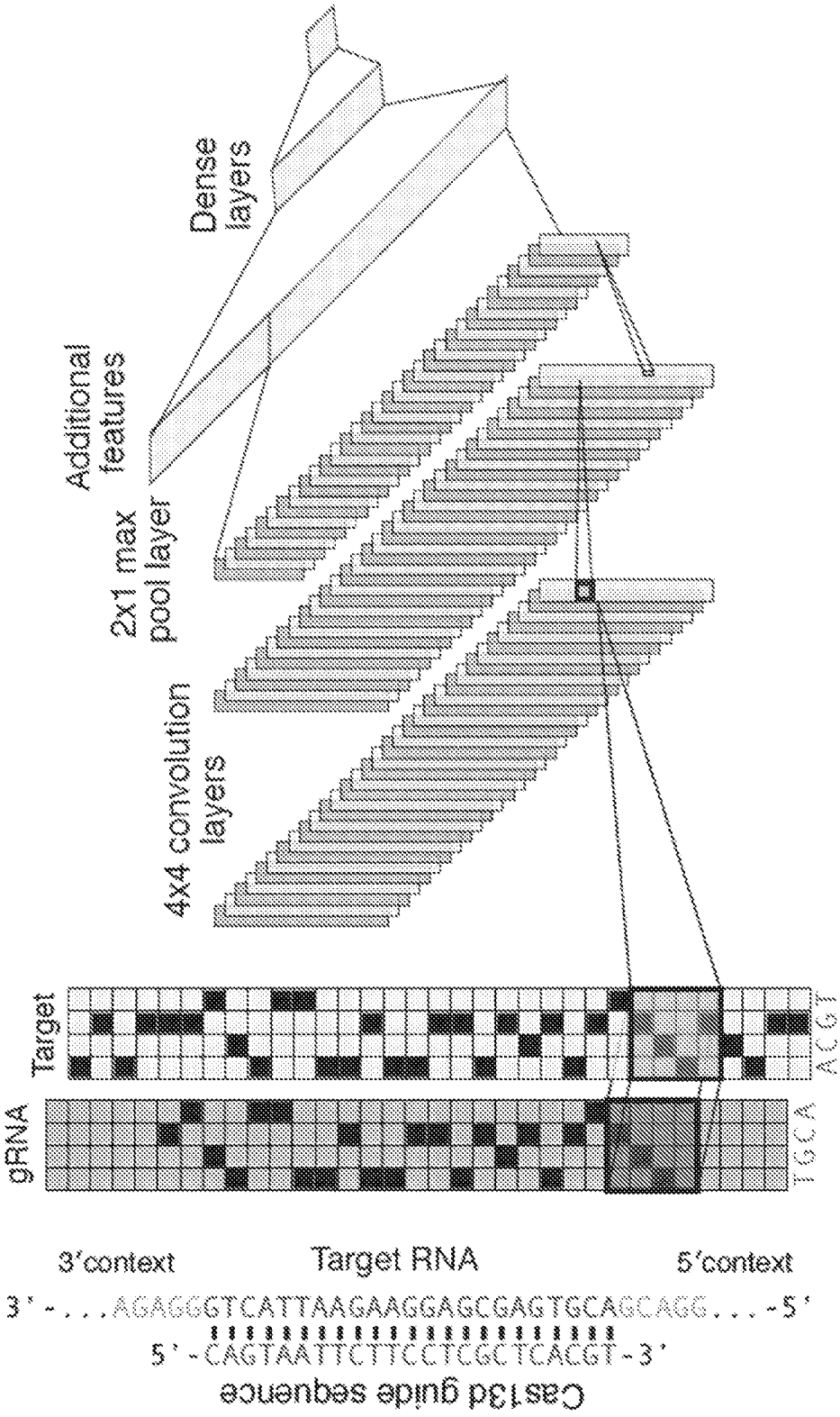


FIG. 3B

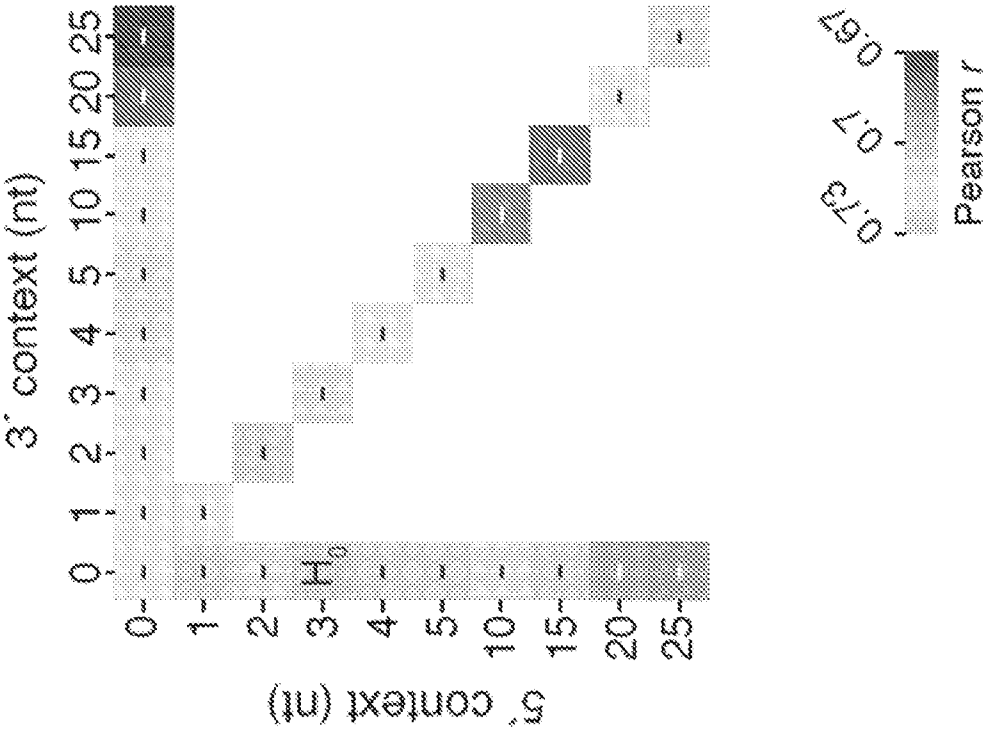


FIG. 3C

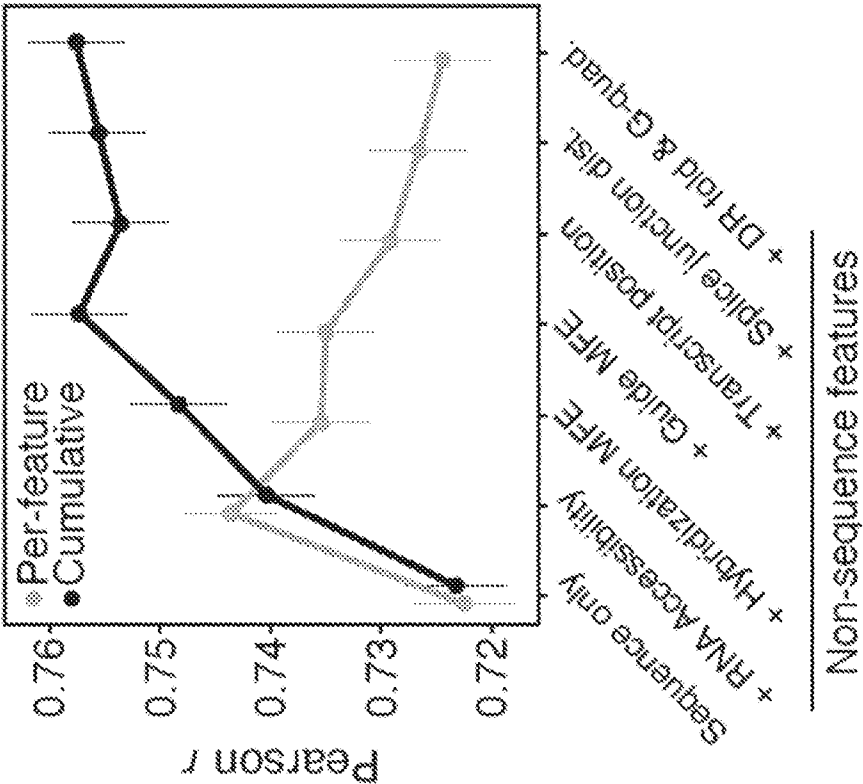


FIG. 3D

Essential genes dropout pooled screen
(this study - gene holdout)

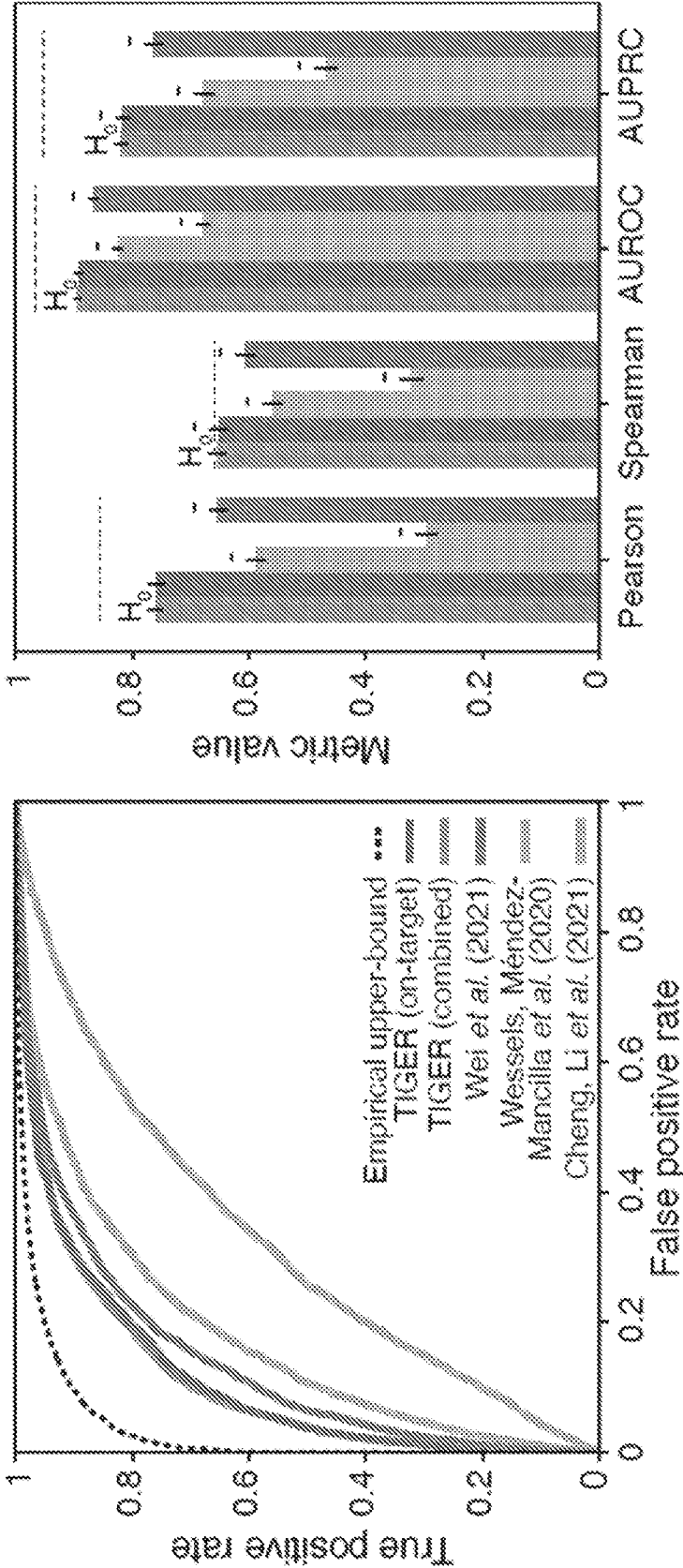


FIG. 3E

Cell surface marker pooled screen
(Wessels, Méndez-Mancilla *et al.*, 2020)

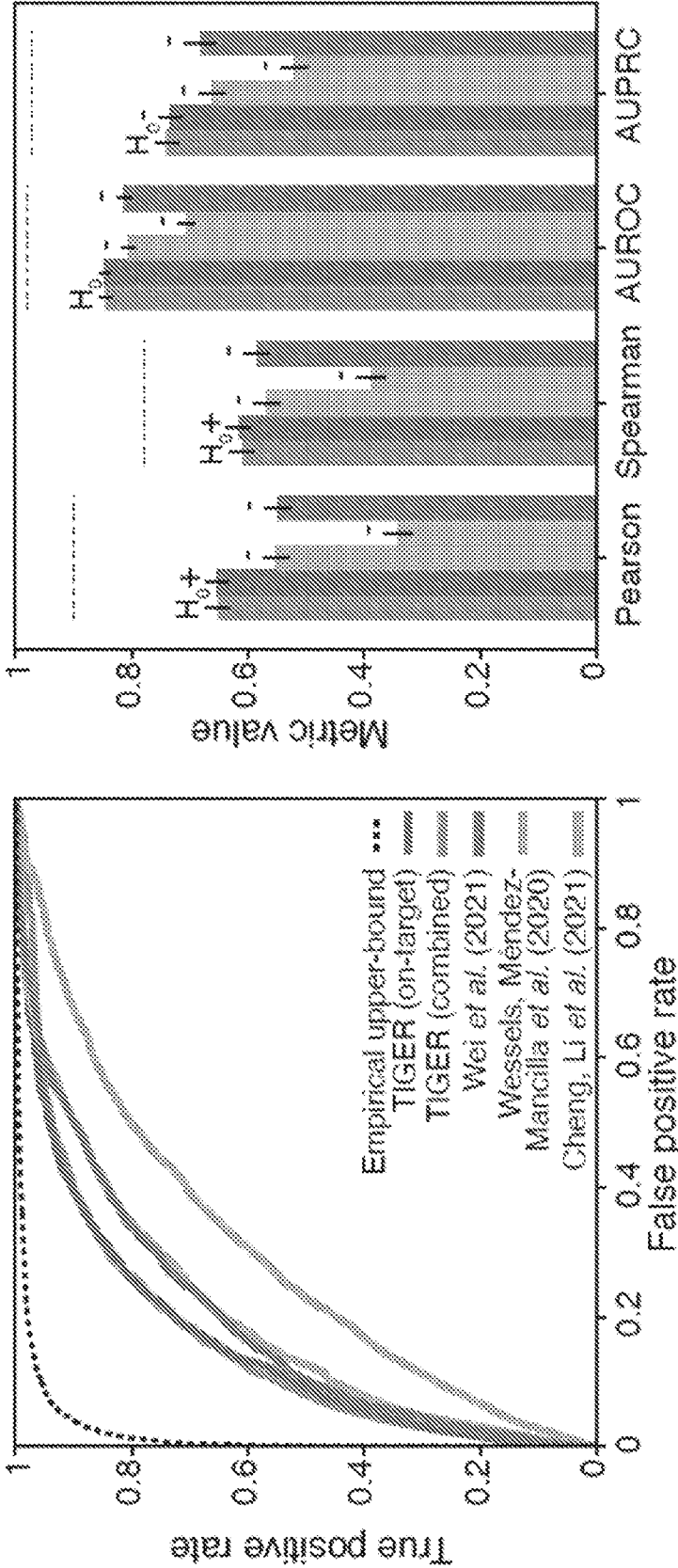


FIG. 3F

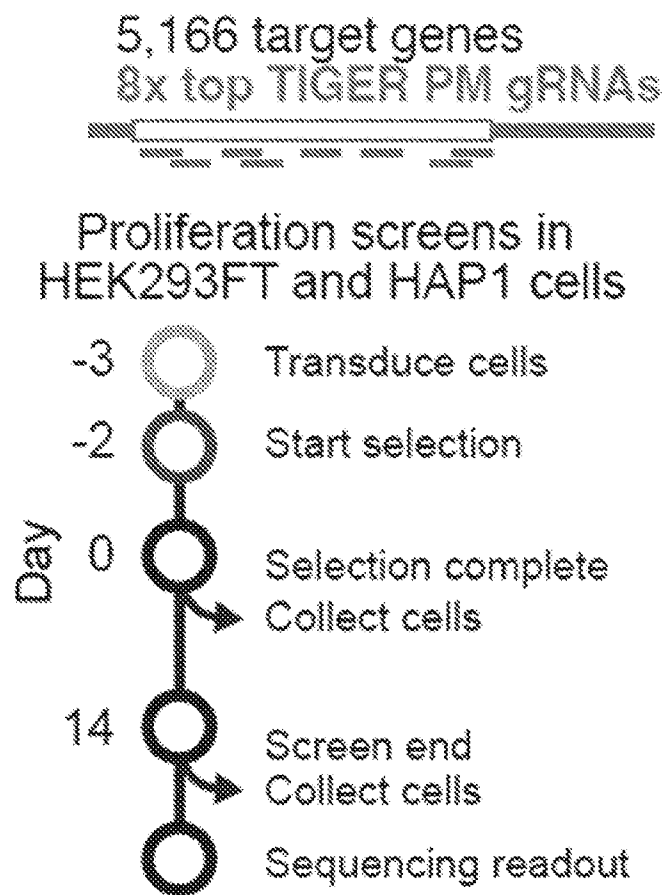


FIG. 3G

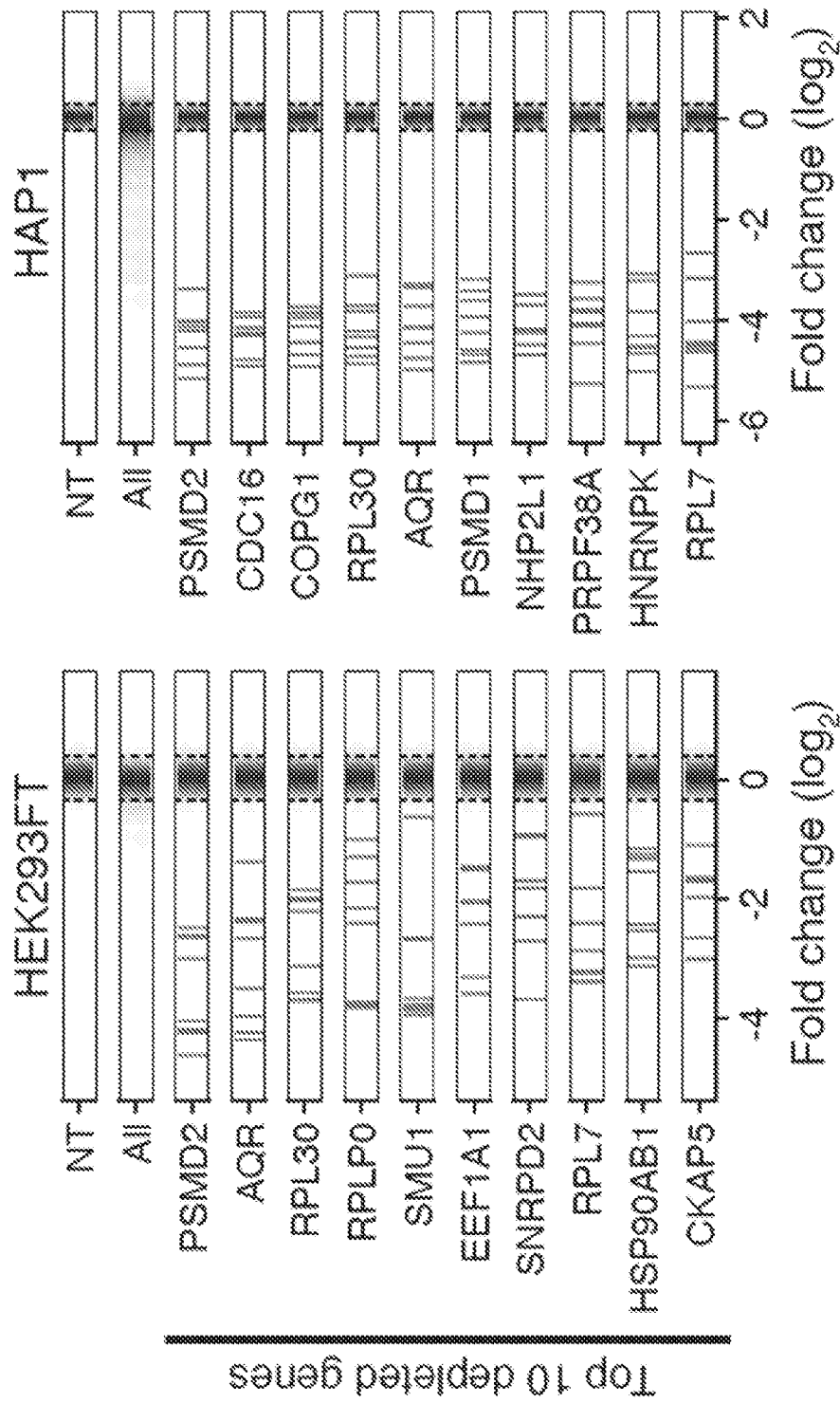


FIG. 3H

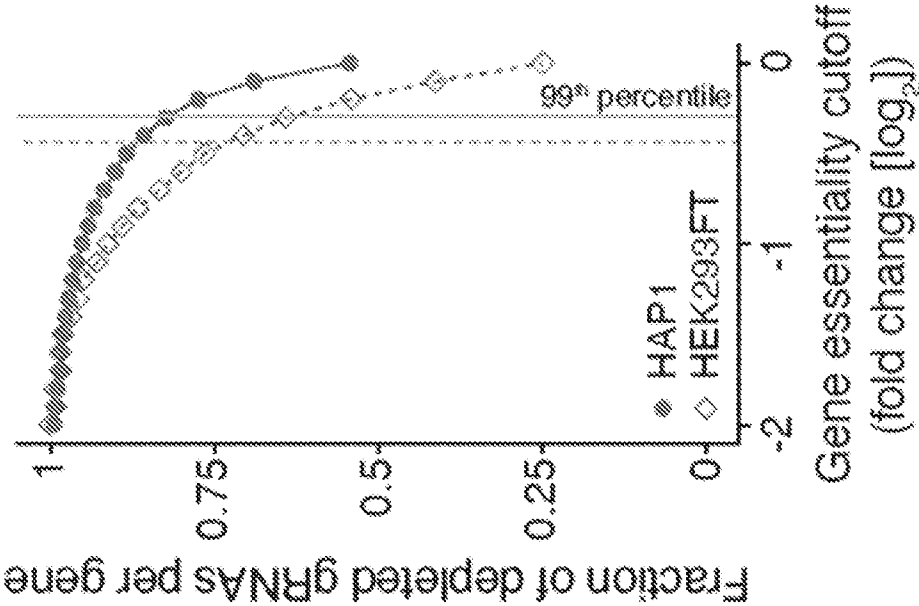


FIG. 3I

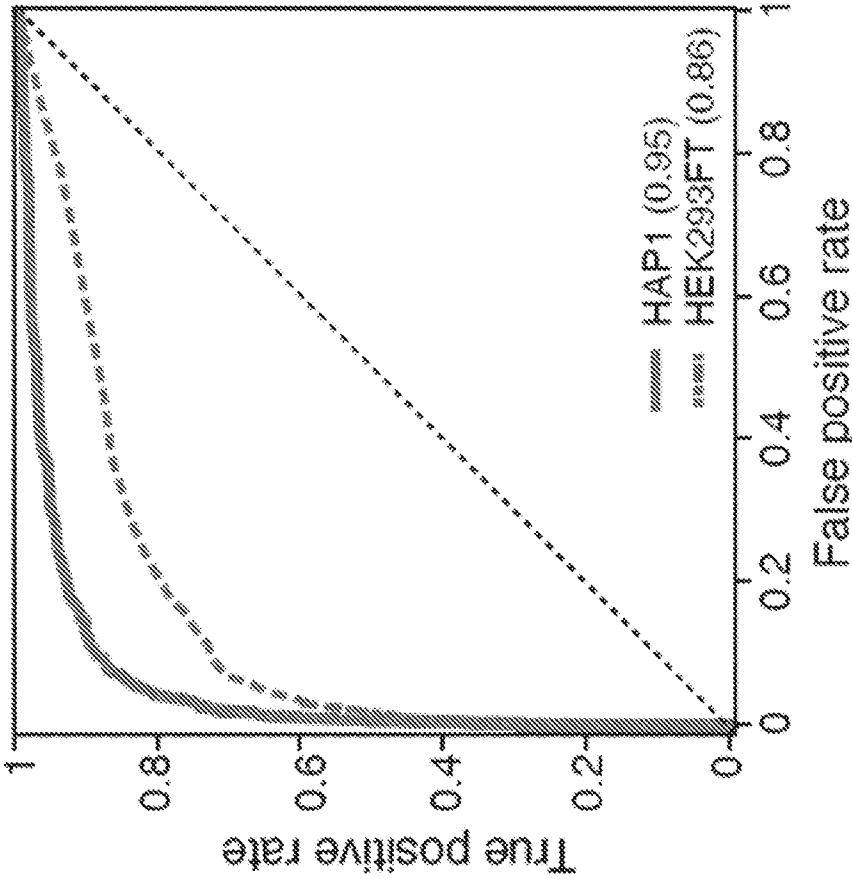


FIG. 4A

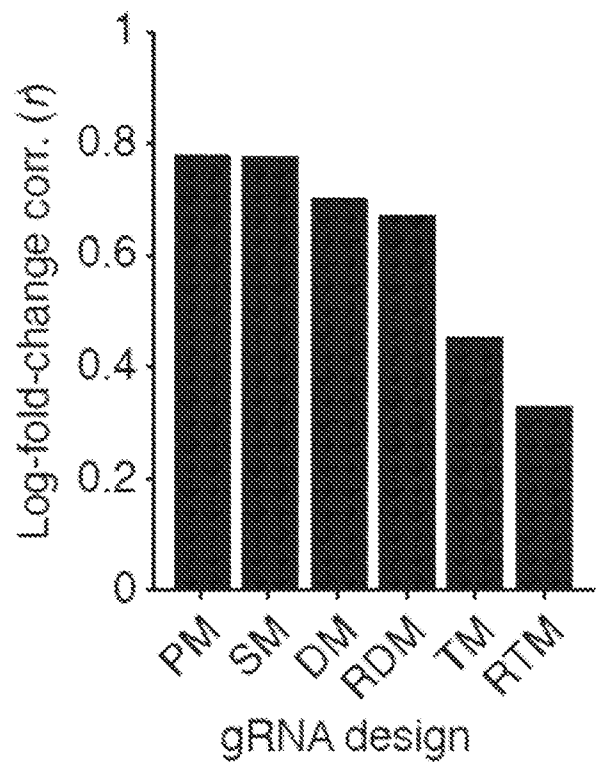


FIG. 4B

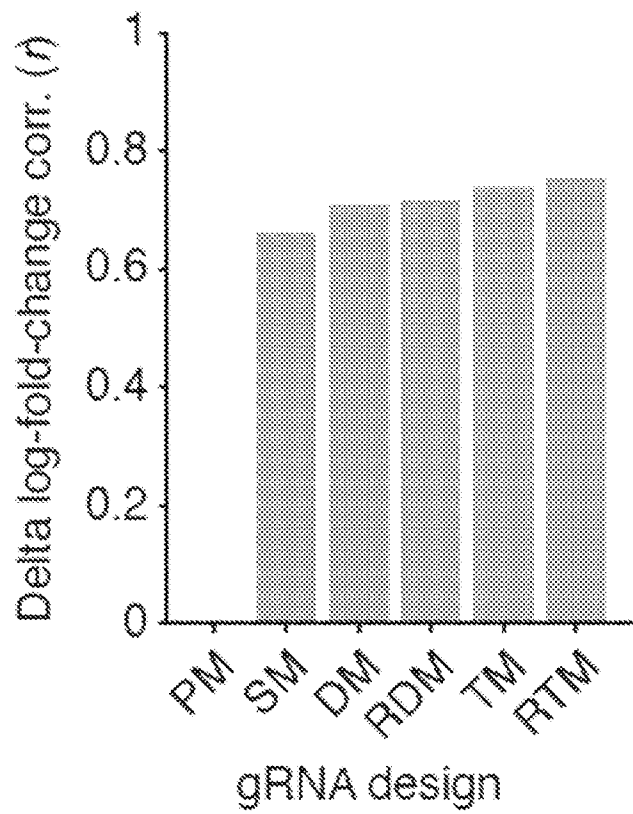


FIG. 4C

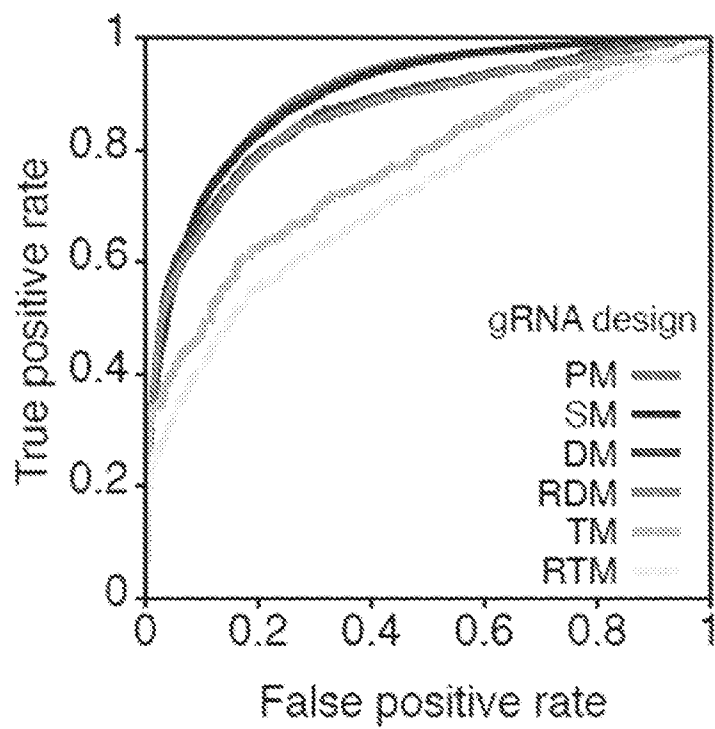


FIG. 4D

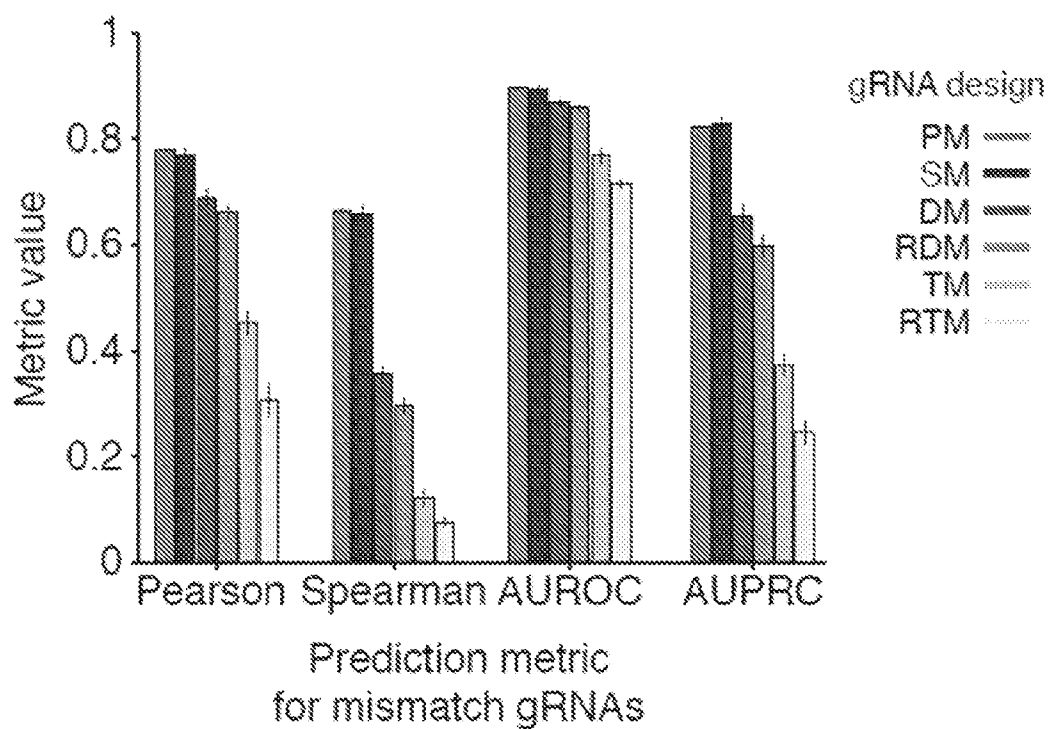


FIG. 4F

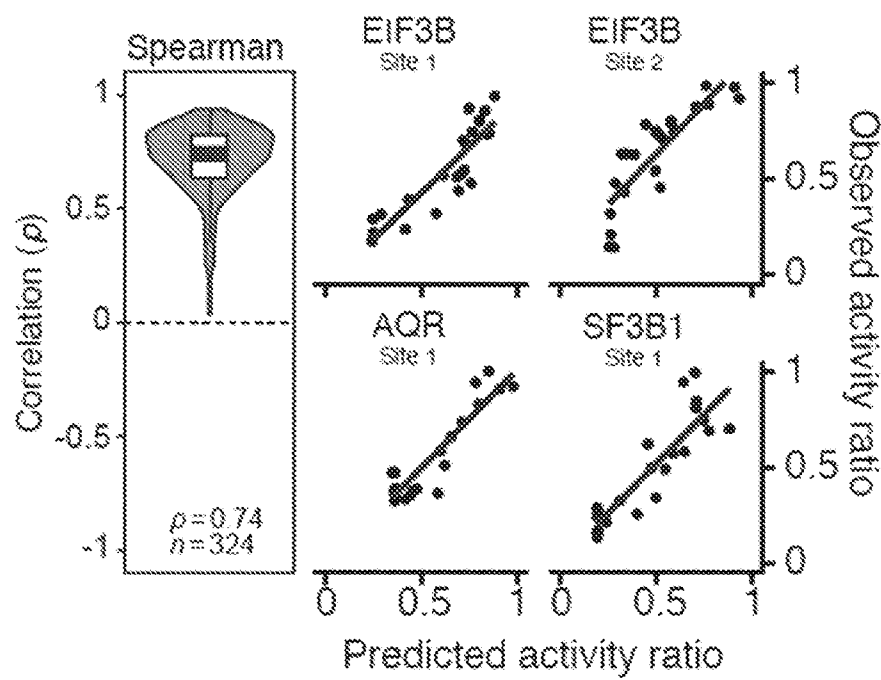


FIG. 4G

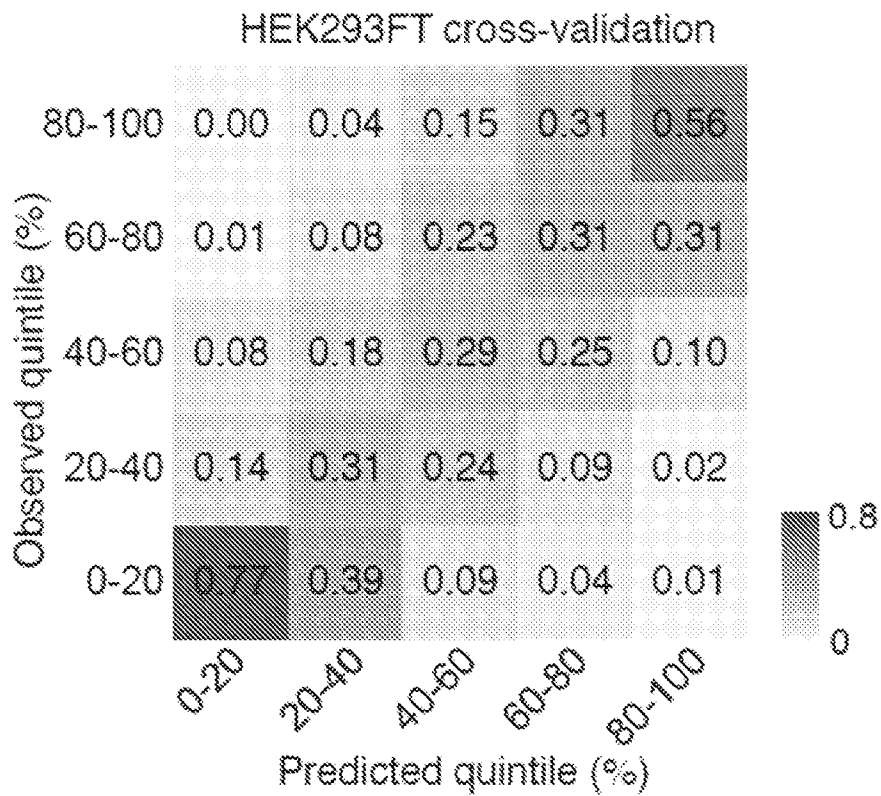


FIG. 4H

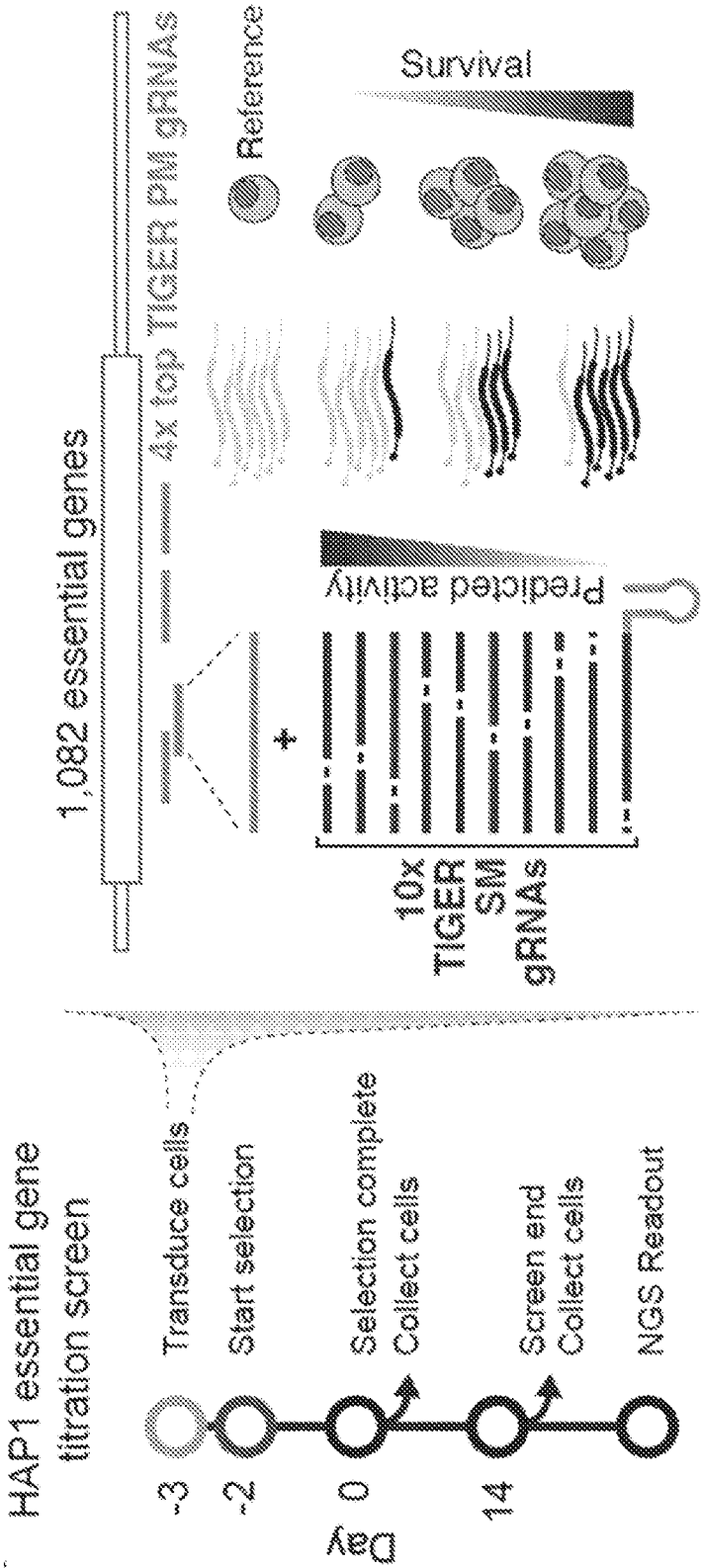


FIG. 4I

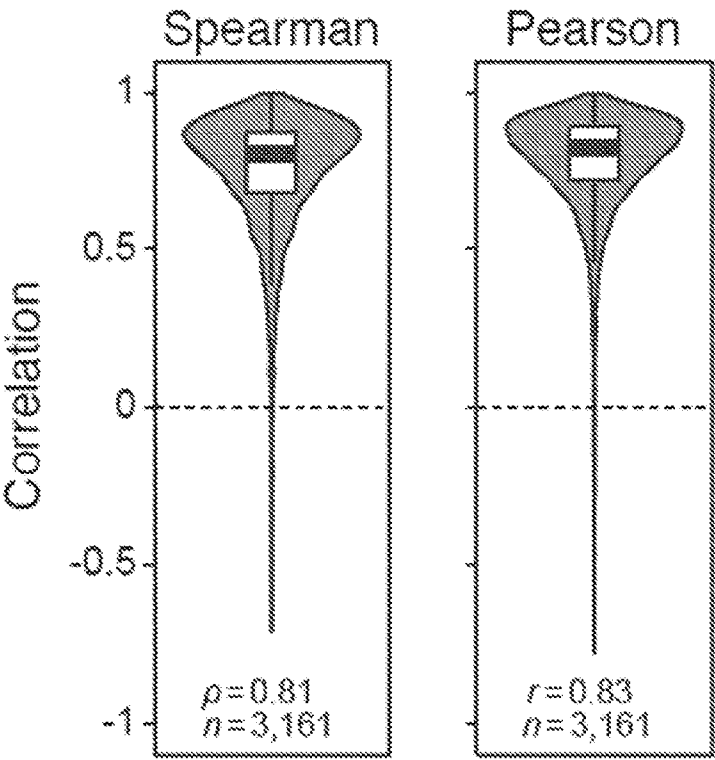


FIG. 4J

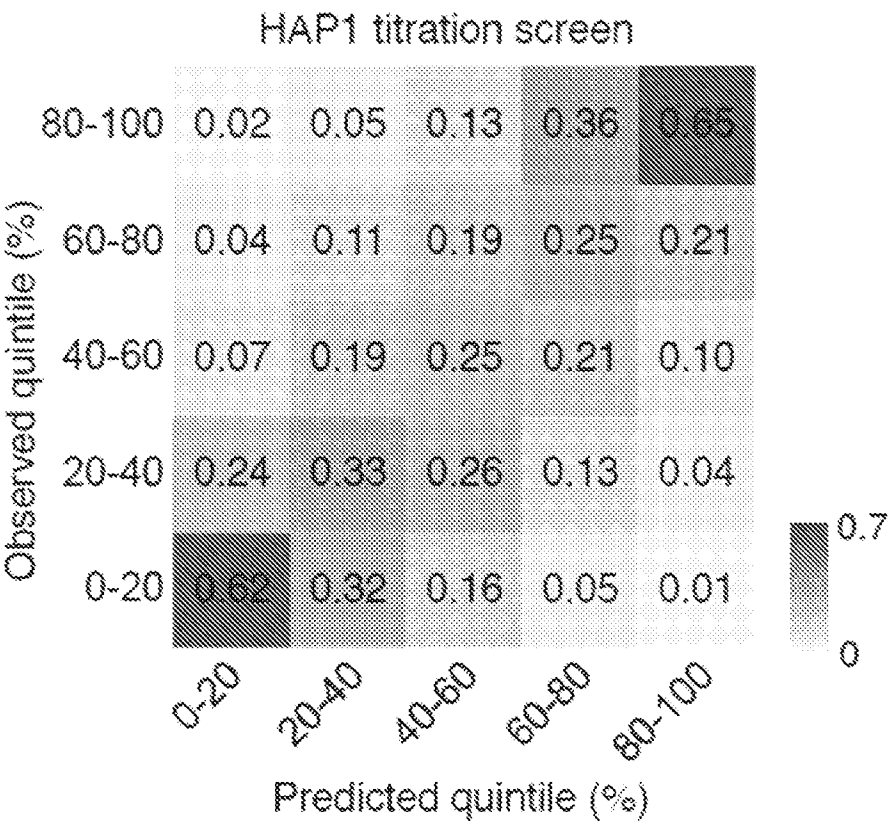


FIG. 5

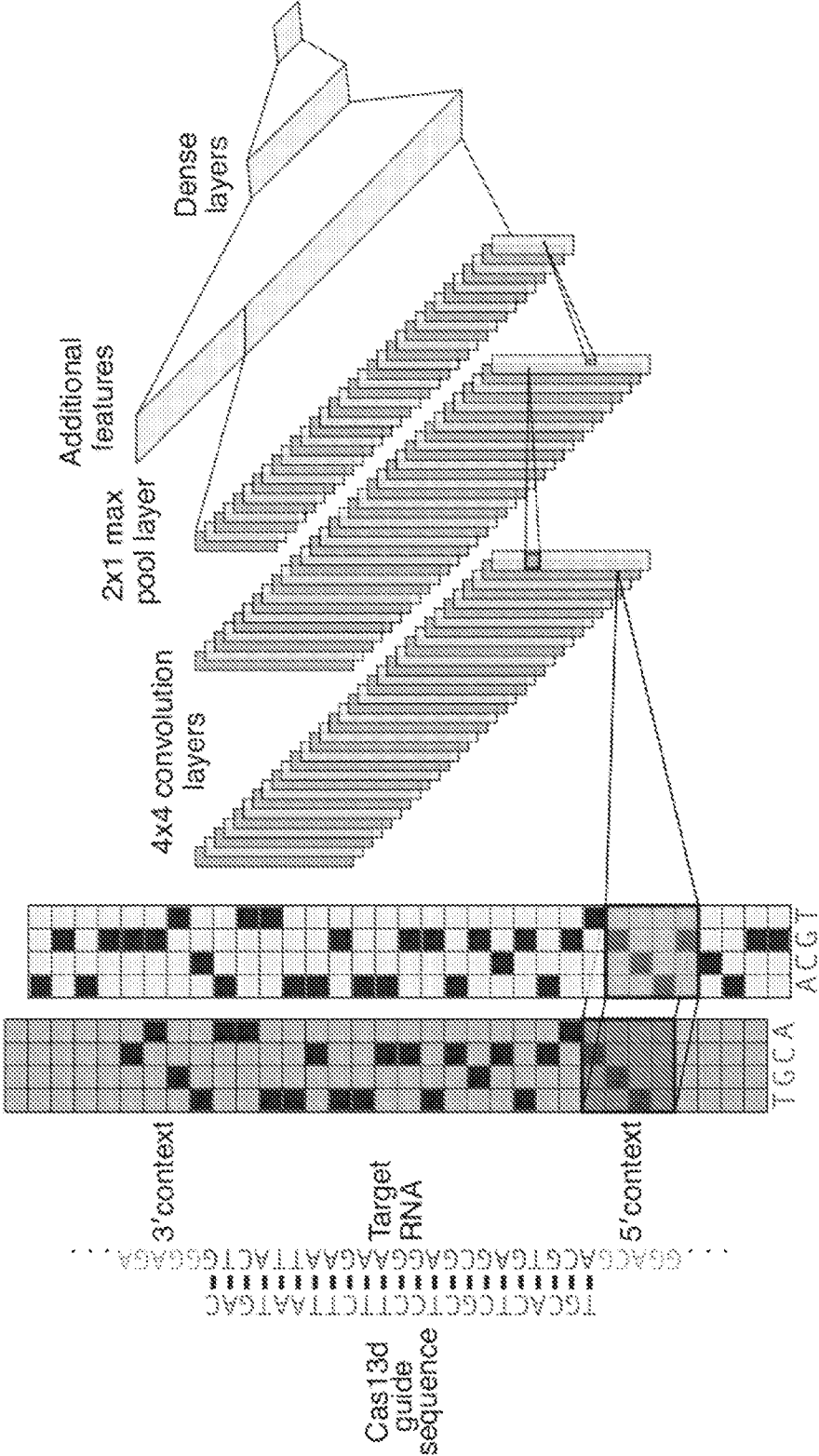


FIG. 6

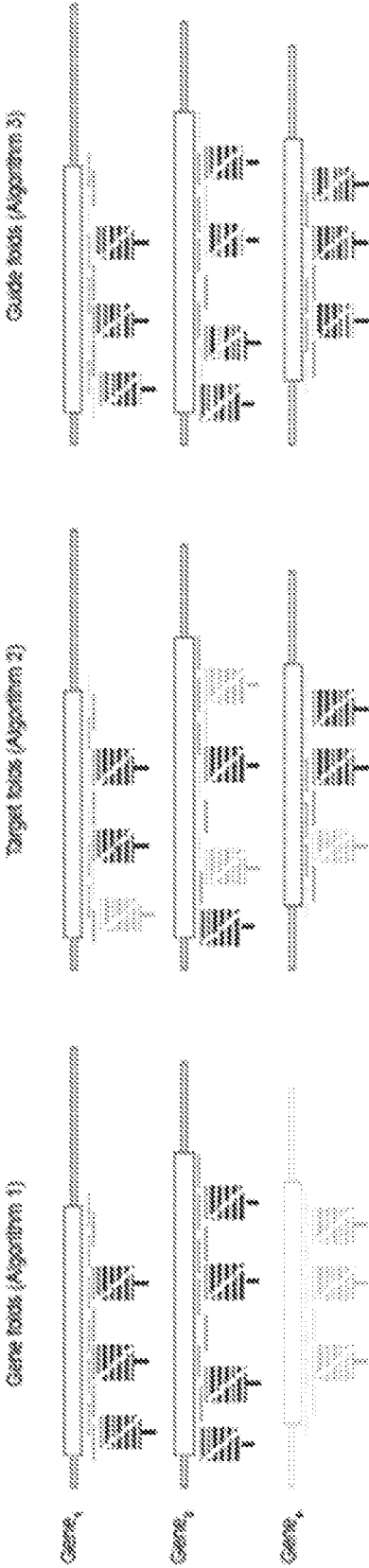


FIG. 7

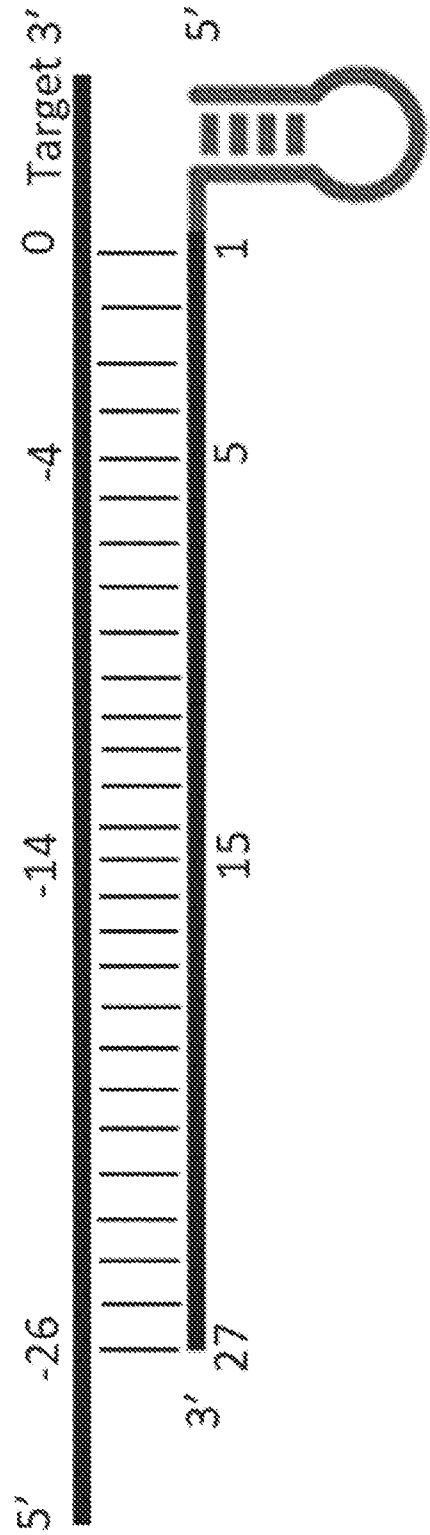


FIG. 8A

Corellation of target site accessibility and guide RNA efficacy

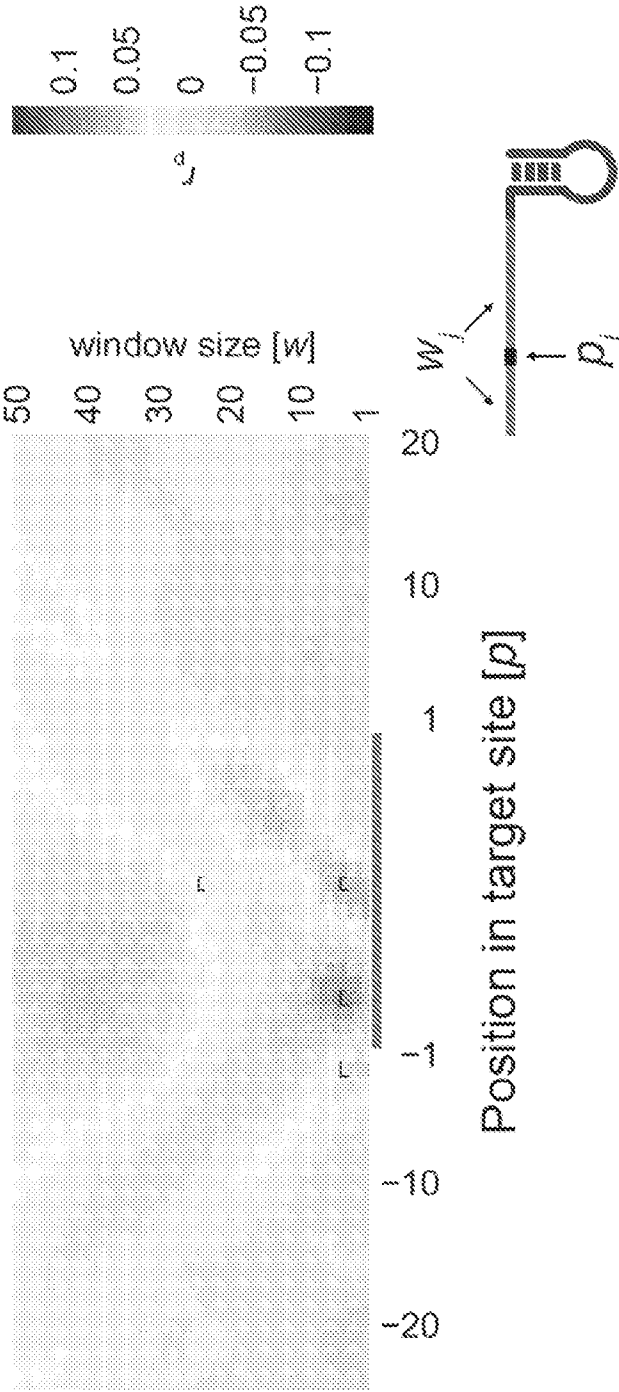


FIG. 8B
Corellation of target site hybridization and guide RNA efficacy

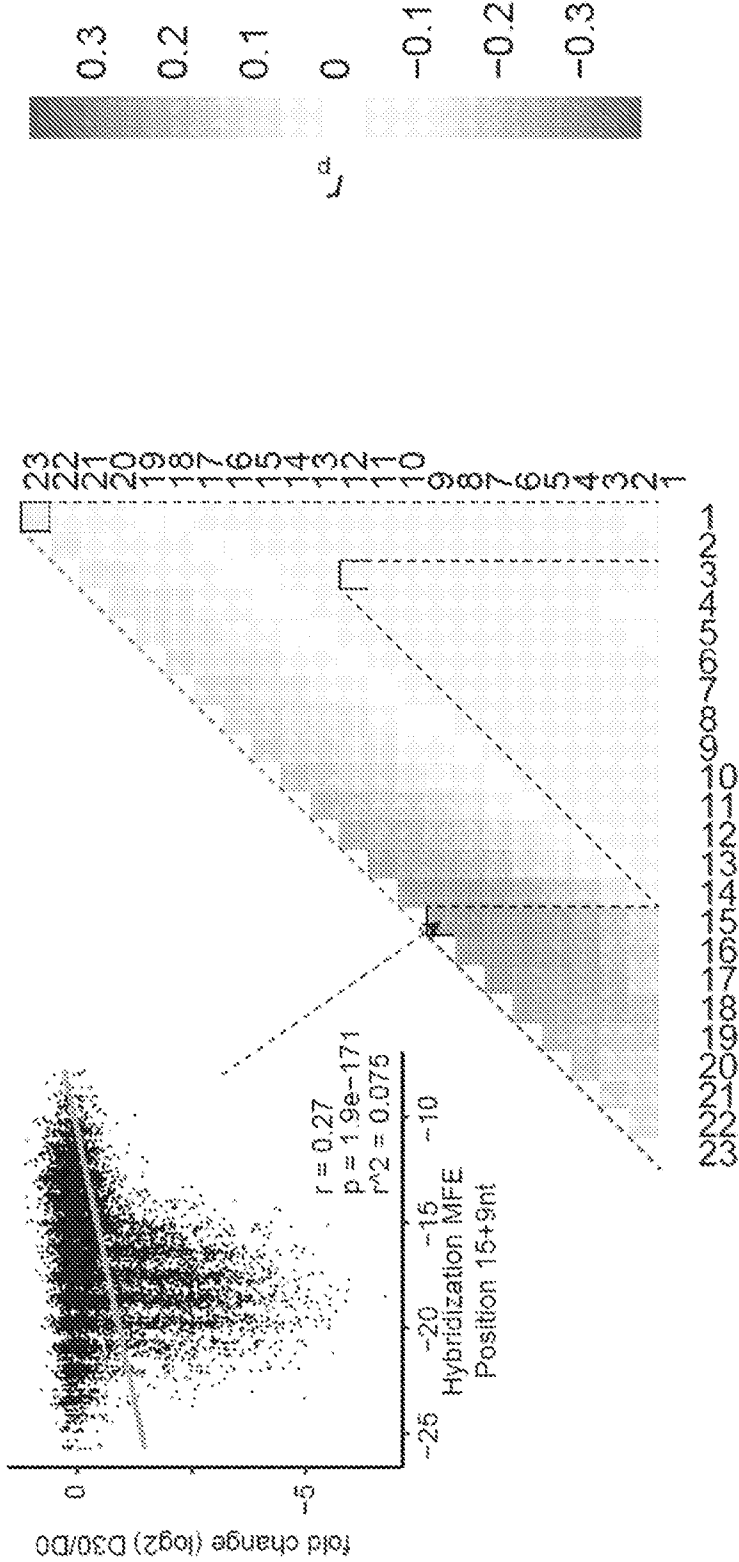


FIG. 9A

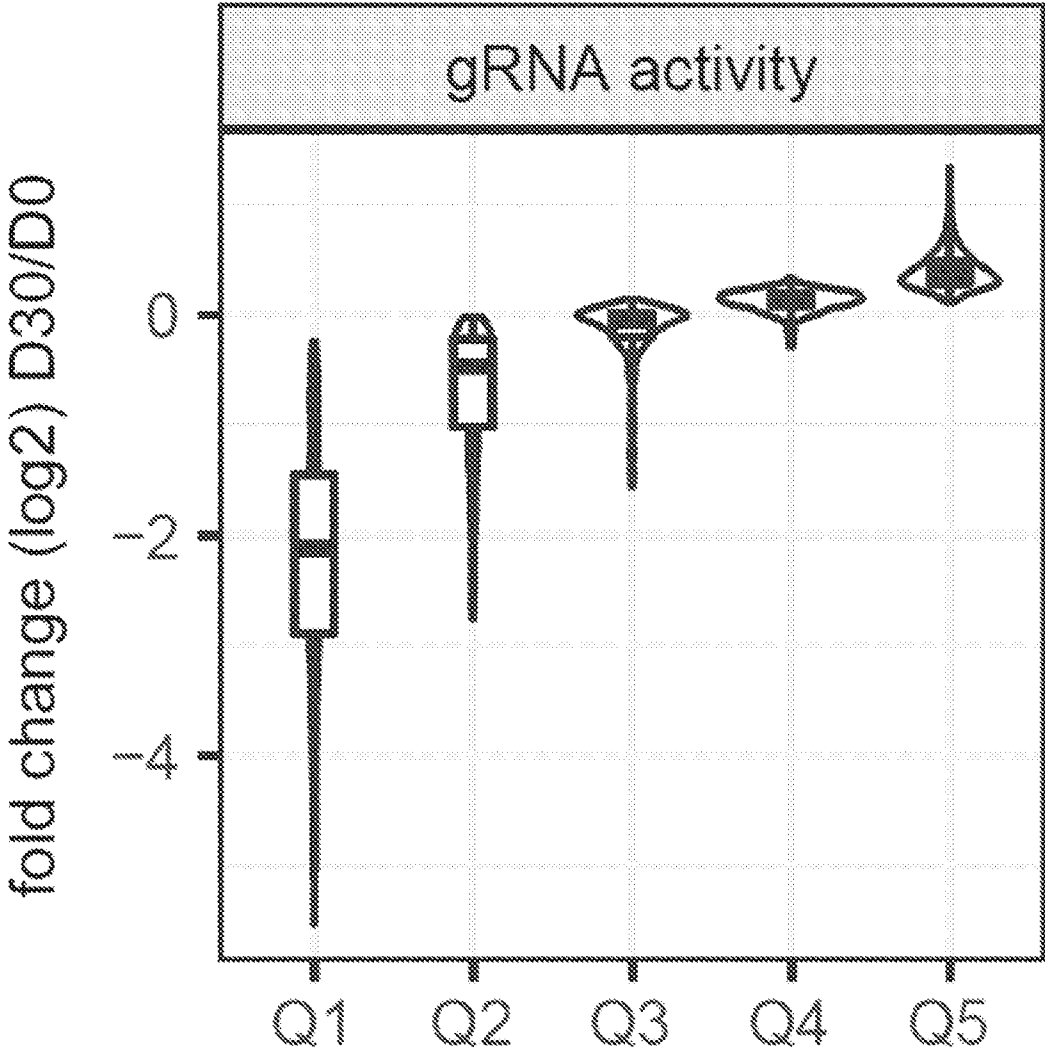


FIG. 9B

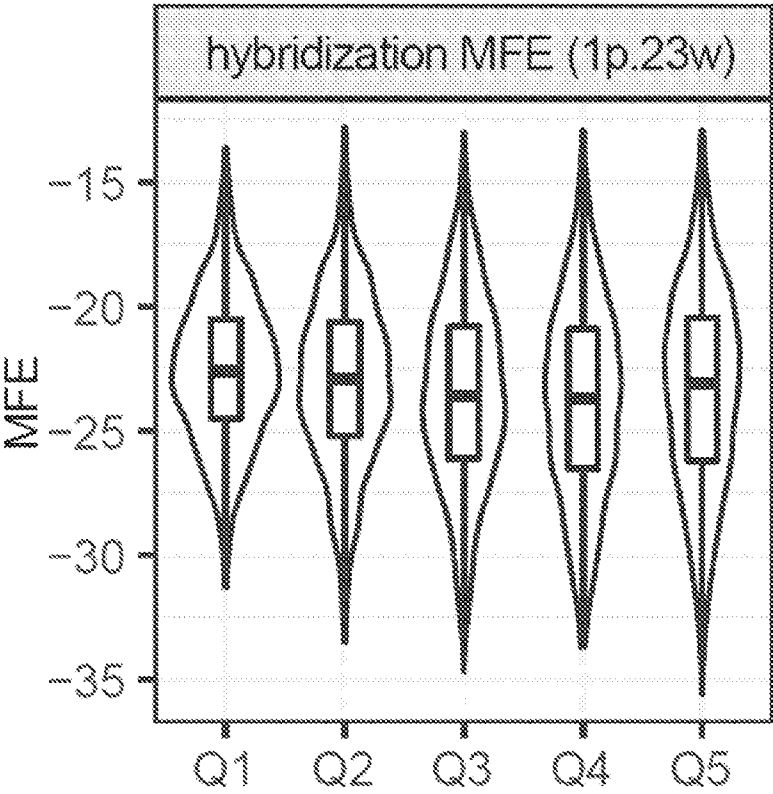


FIG. 9C

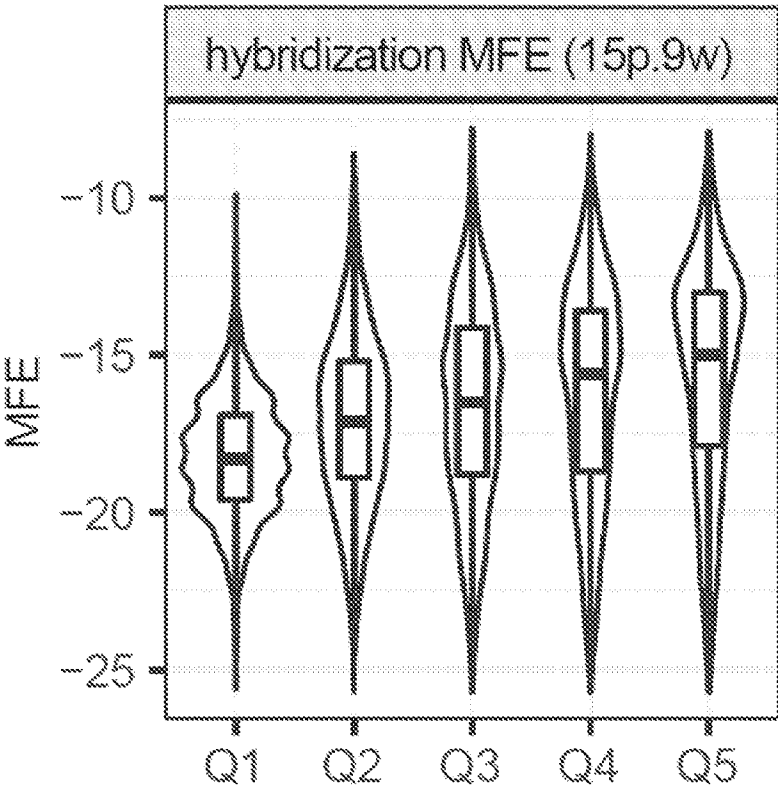


FIG. 9D

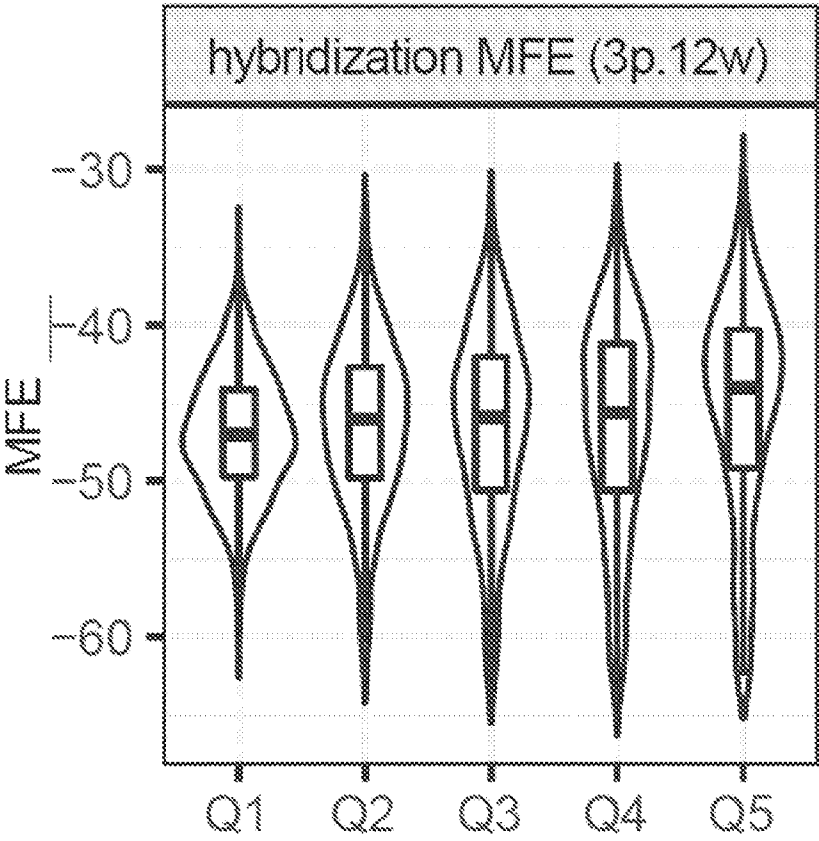


FIG. 9E

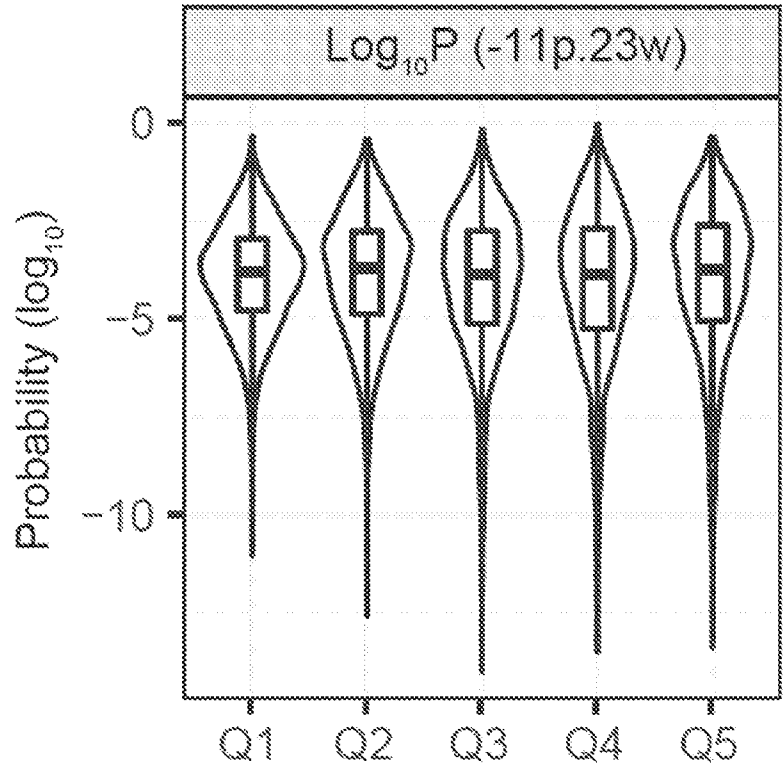


FIG. 9F

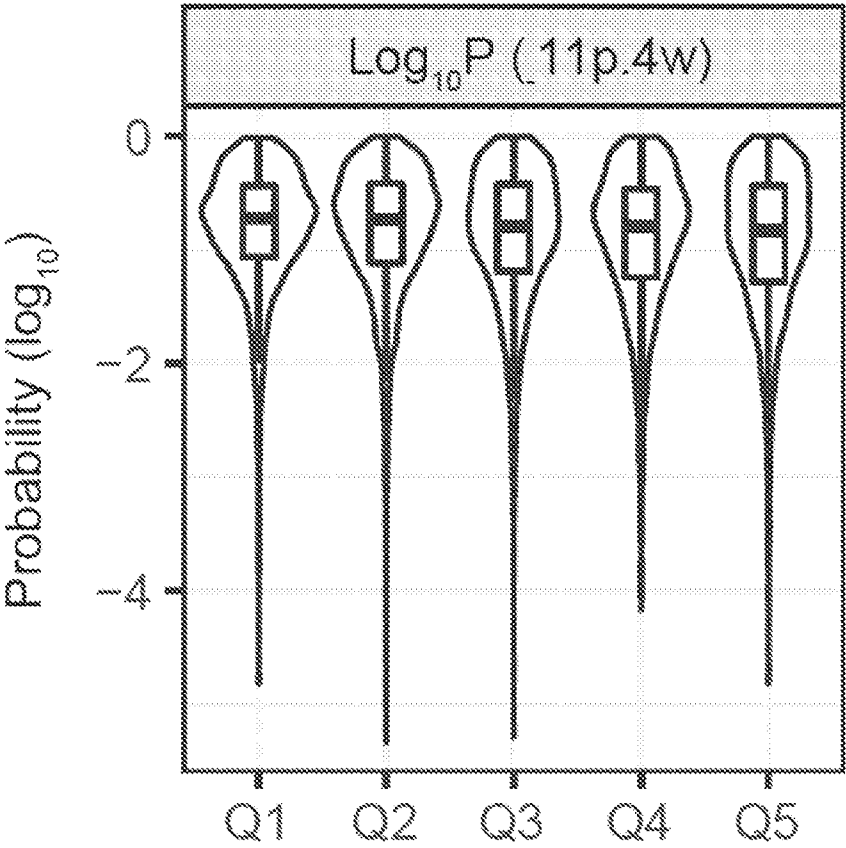


FIG. 9G

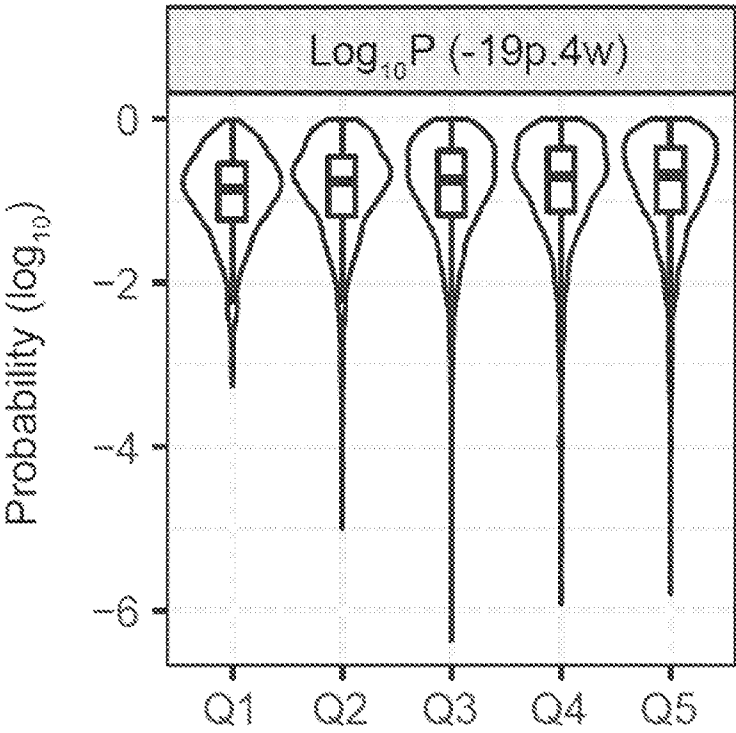


FIG. 9H

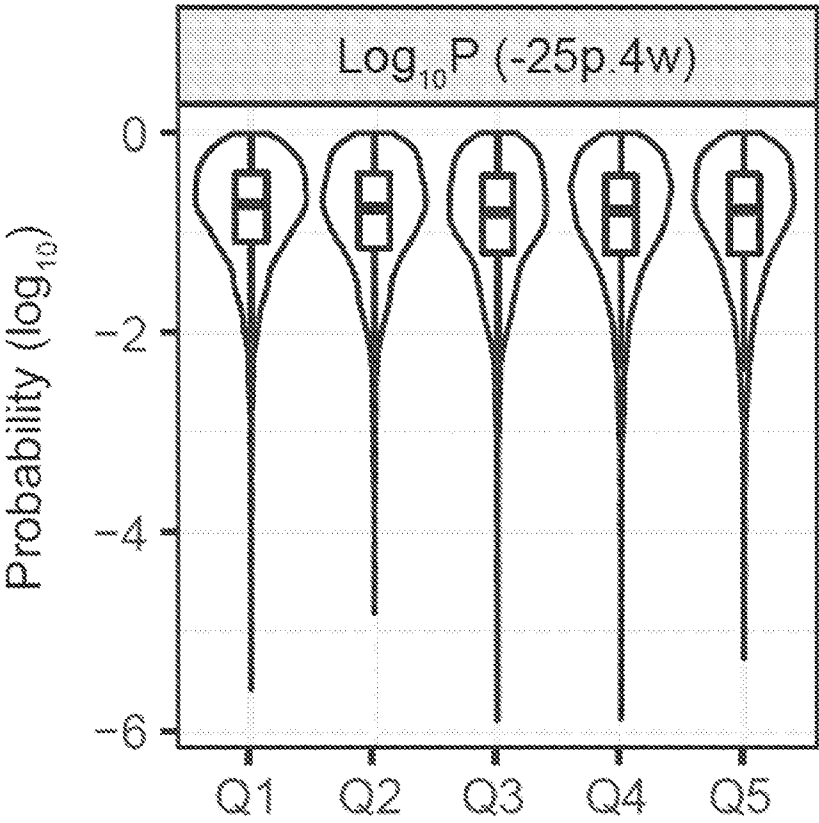


FIG. 9I

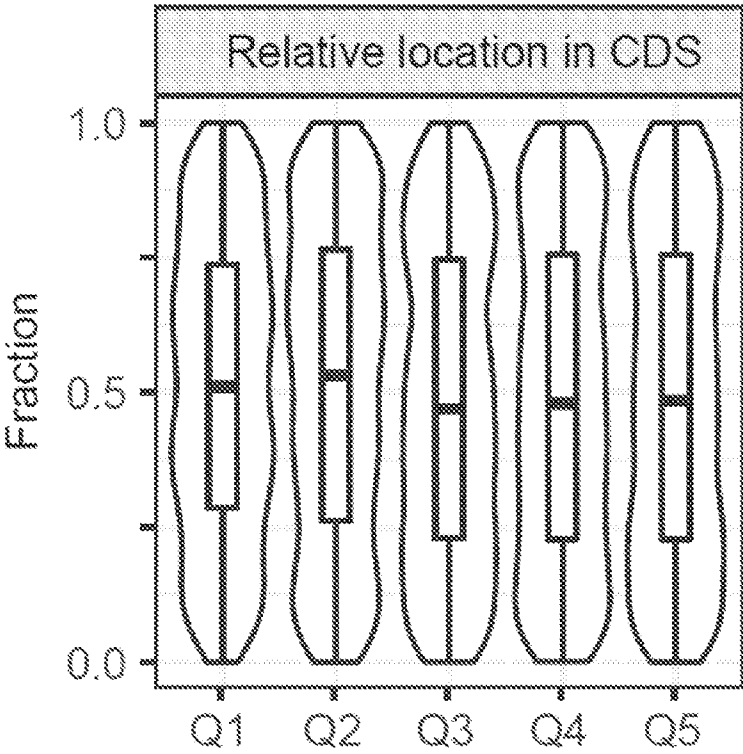


FIG. 9J

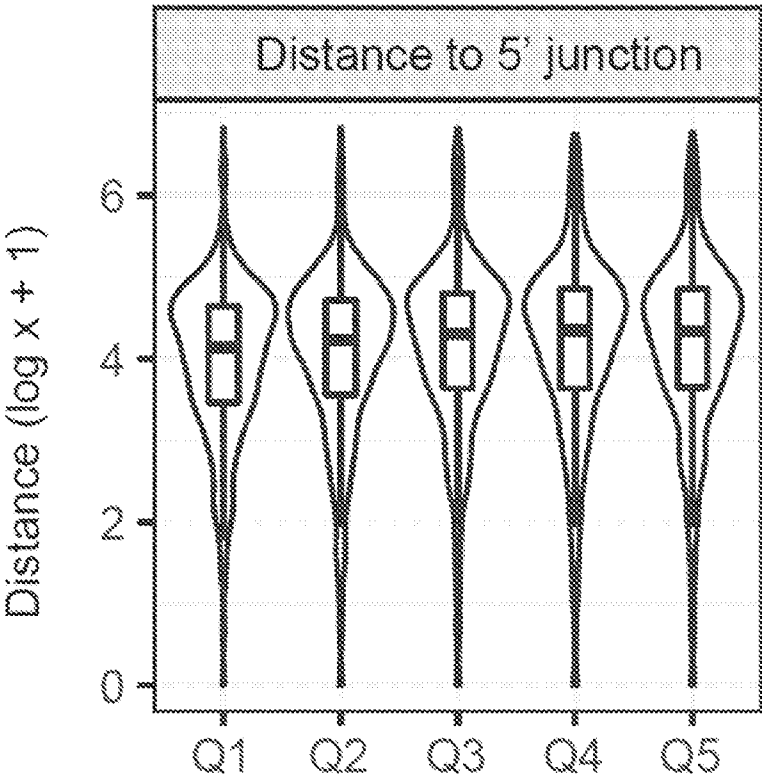


FIG. 9K

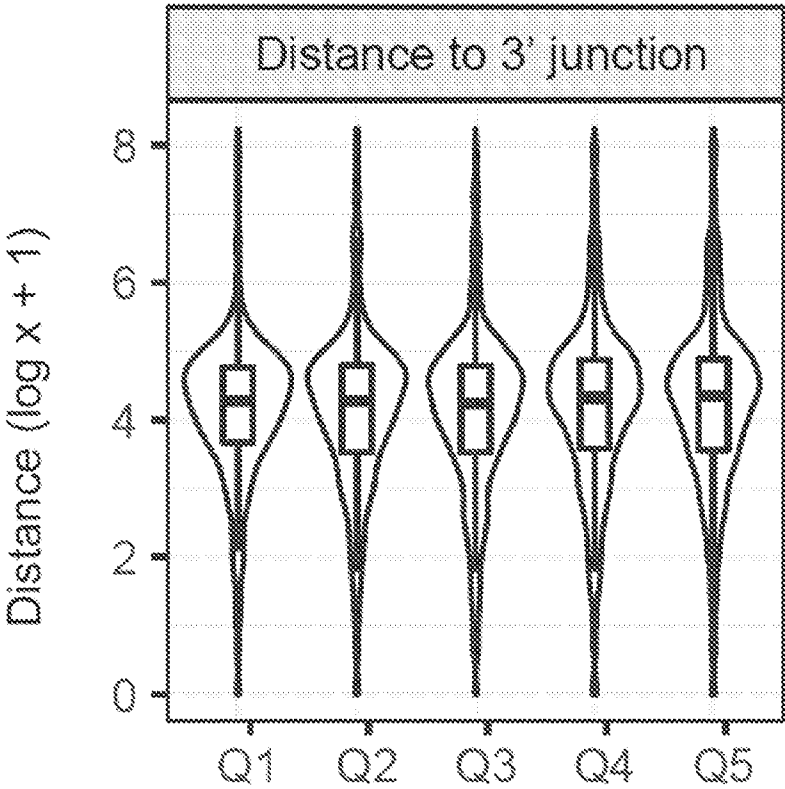


FIG. 9L

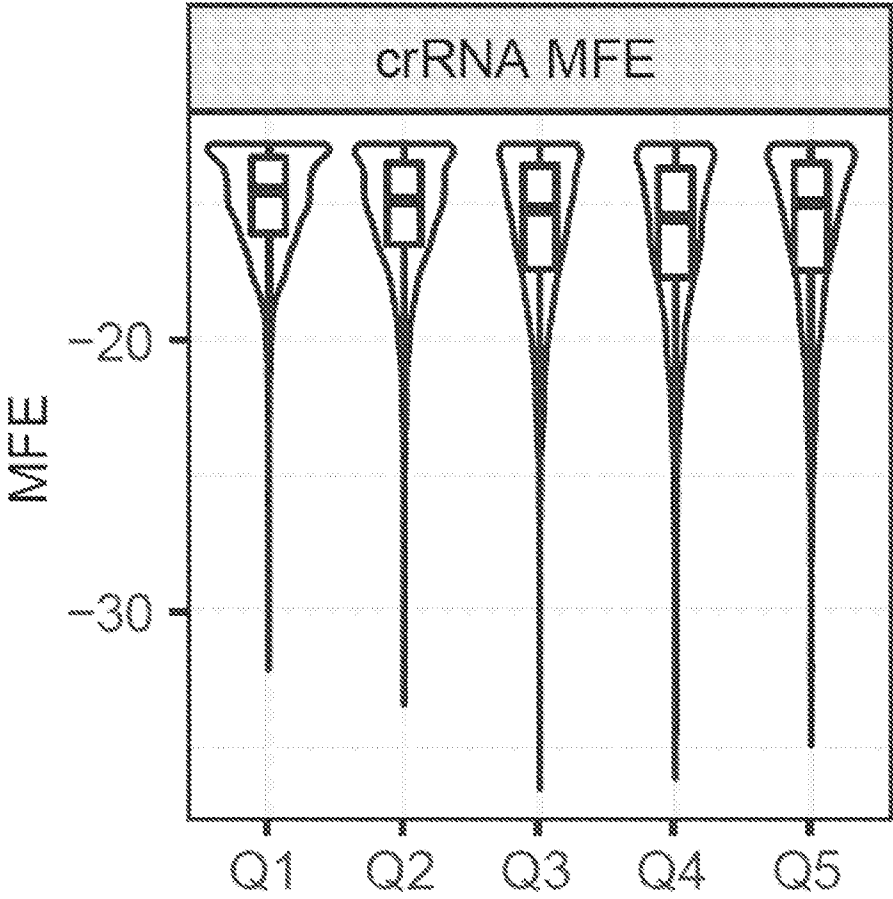


FIG. 9M

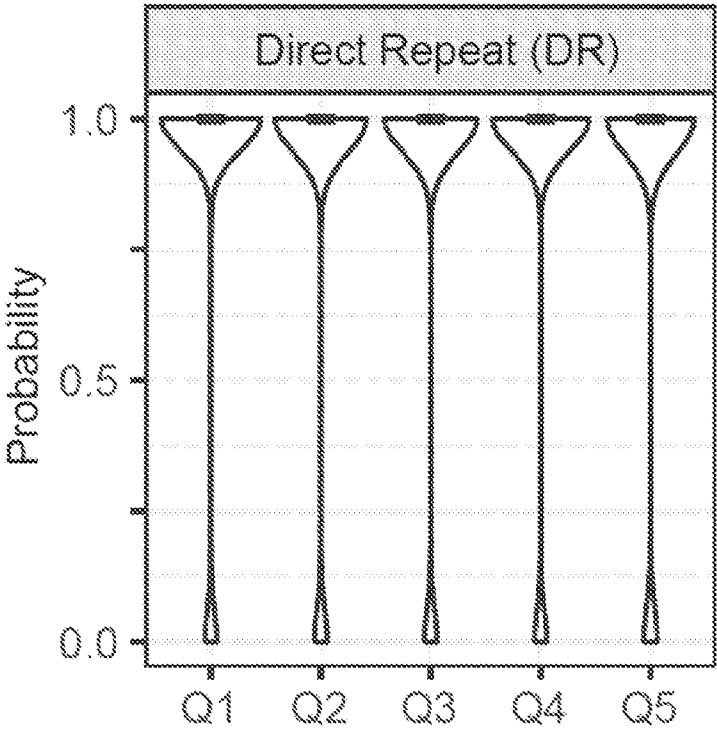


FIG. 9N

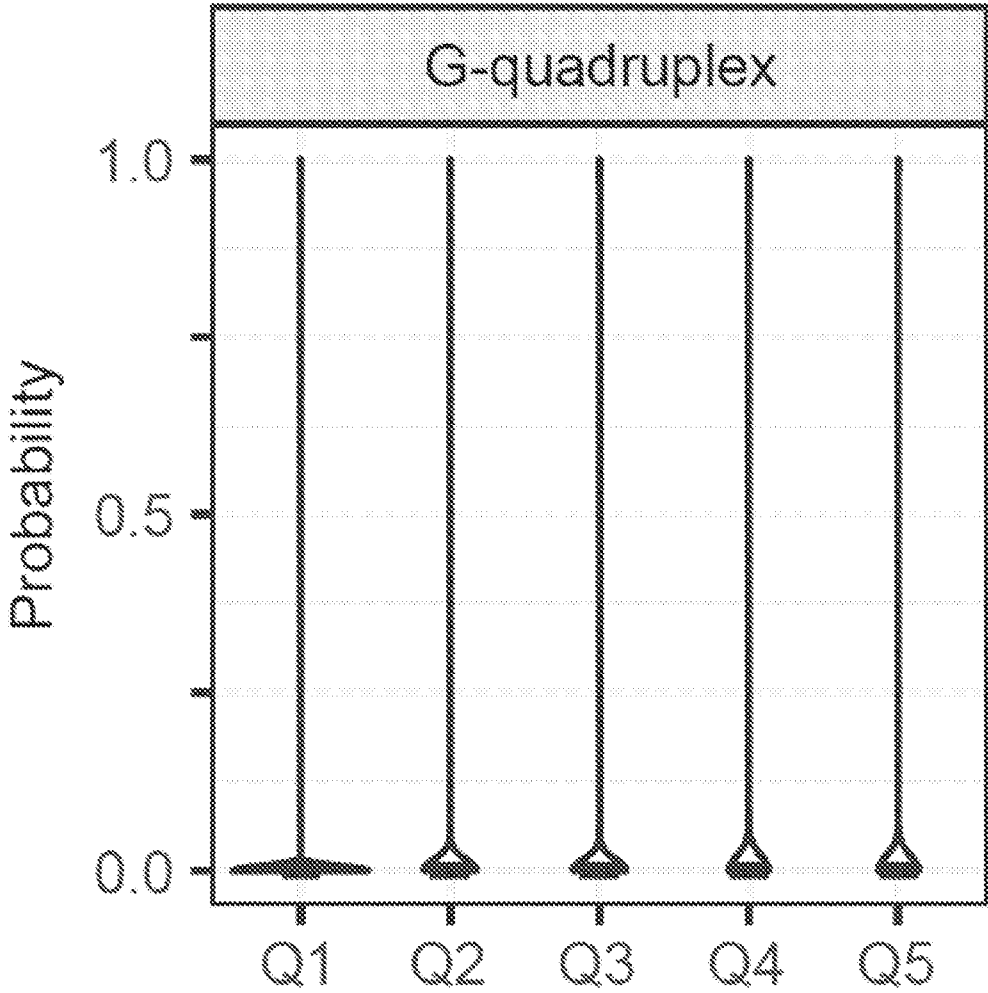


FIG. 10A

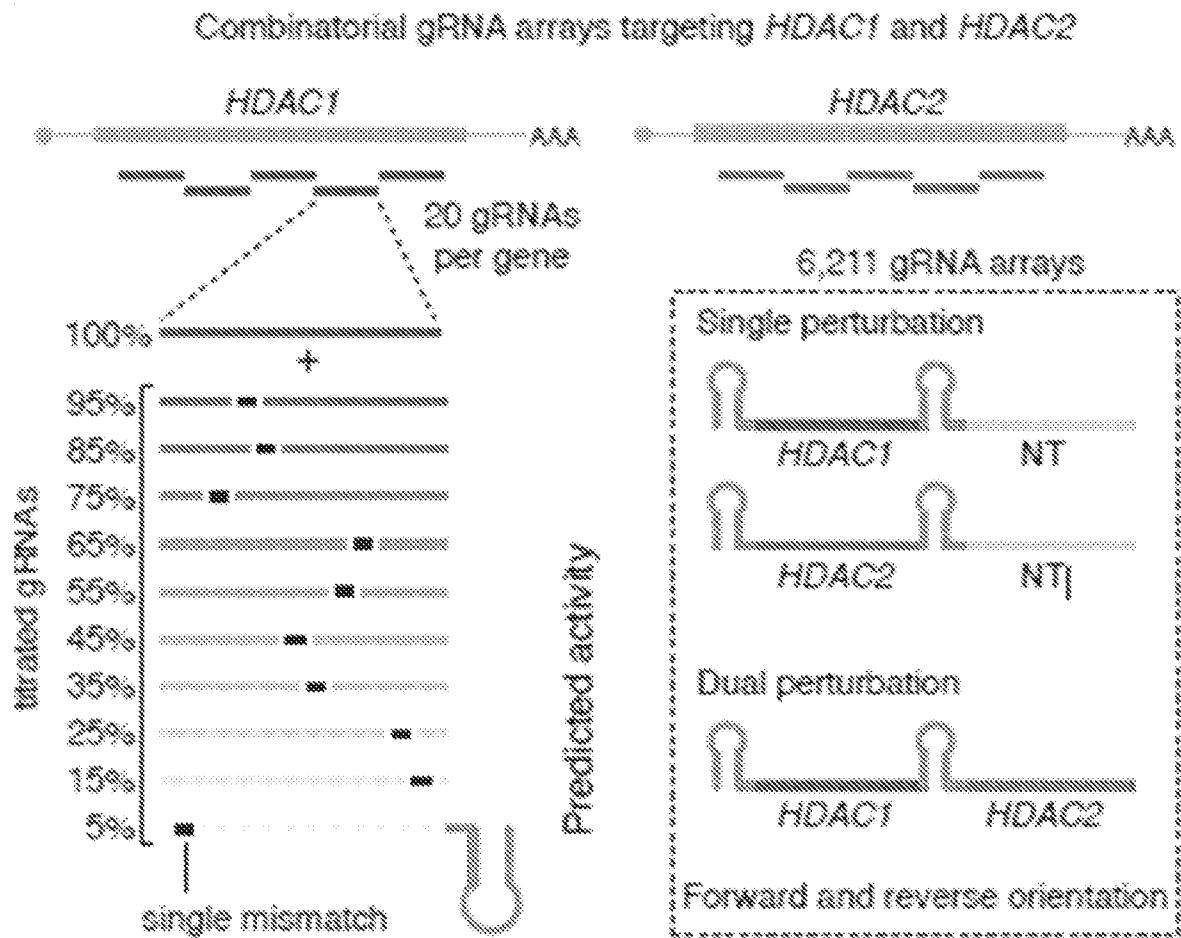
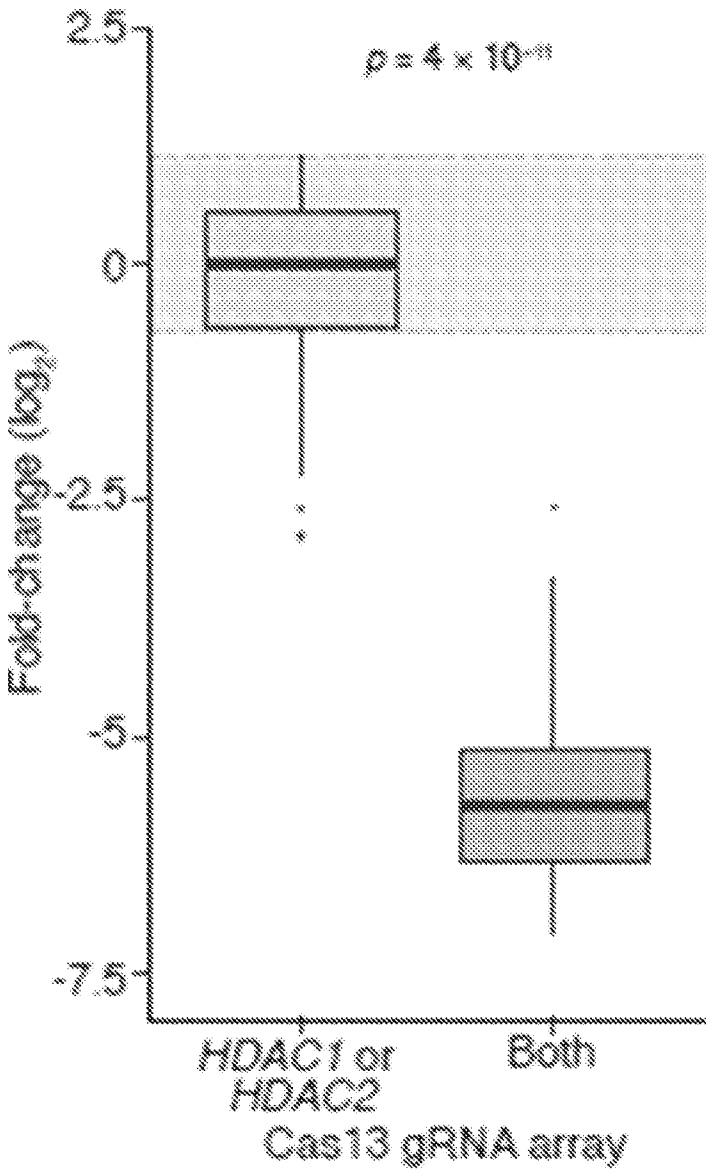


FIG. 10B



INTERNATIONAL SEARCH REPORT

International application No.

PCT/US2024/033813

A. CLASSIFICATION OF SUBJECT MATTER

IPC: **G16B 35/20** (2024.01); **G16B 40/00** (2024.01); **G16B 20/00** (2024.01); **G16B 25/10** (2024.01); **C12N 15/113** (2024.01); **C12N 15/85** (2024.01); **C12Q 1/6813** (2024.01); **C12Q 1/6876** (2024.01); **C40B 30/06** (2024.01); **C12N 9/22** (2024.01); **A61K 48/00** (2024.01)

CPC: **G16B 35/20**; **C12N 15/111**; **G16B 40/00**; **C12N 15/113**; **C12N 15/85**; **G16B 25/10**; **G16B 20/00**; **C12N 15/1089**; **C12Q 1/6813**; **C12Q 1/6876**; **A61K 48/00**; **C12N 2310/20**; **C12N 2330/31**; **C12N 2320/11**; **C12N 9/22**; **C40B 30/06**

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

See Search History Document

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

See Search History Document

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

See Search History Document

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2021/0308171 A1 (THE BROAD INSTITUTE INC. et al.) 07 October 2021 (07.10.2021) entire document	1-4, 7-11, 18, 19
A	US 2022/0259593 A1 (THE REGENTS OF THE UNIVERSITY OF CALIFORNIA) 18 August 2022 (18.08.2022) entire document	1-4, 7-11, 18, 19
A	US 2022/0001030 A1 (FONDAZIONE TELETHON) 06 January 2022 (06.01.2022) entire document	1-4, 7-11, 18, 19
Y	WESSELS et al., Massively parallel Cas13 screens reveal principles for guide RNA design, Nature biotechnology, Vol. 38, 16 March 2020, Pgs. 722-727, [Retrieved on 04.09.2024]. Retrieved from the internet:<URL: http://www.sanjanalab.org/reprints/WesselsMen dez_NatureBiotech_2020.pdf >. entire document	8-11



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

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

“D” document cited by the applicant in the international application

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

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

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

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

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

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

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

“&” document member of the same patent family

Date of the actual completion of the international search

07 September 2024 (07.09.2024)

Date of mailing of the international search report

10 October 2024 (10.10.2024)

Name and mailing address of the ISA/US

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

Authorized officer

**MATOS
TAINA**

Facsimile No. **571-273-8300**

Telephone No. **571-272-4300**

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US2024/033813

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	BODAI et al., Targeting double-strand break indel byproducts with secondary guide RNAs improves Cas9 HDR-mediated genome editing efficiencies. Nat Commun, Vol. 13:2351, 09 May 2022, Pgs. 1-15, [Retrieved on 05.09.2024]. Retrieved from the internet: <URL: https://www.nature.com/articles/s41467-022-29989-9 >. entire document	8-11

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US2024/033813

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

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

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☒ Claims Nos.: **5, 6, 12-17**
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).