



# HHS Public Access

Author manuscript

*Nat Biotechnol.* Author manuscript; available in PMC 2020 September 16.

Published in final edited form as:

*Nat Biotechnol.* 2020 June ; 38(6): 722–727. doi:10.1038/s41587-020-0456-9.

## Massively parallel Cas13 screens reveal principles for guide RNA design

Hans-Hermann Wessels<sup>1,2,†</sup>, Alejandro Méndez-Mancilla<sup>1,2,†</sup>, Xinyi Guo<sup>1,2</sup>, Mateusz Legut<sup>1,2</sup>, Zharko Daniloski<sup>1,2</sup>, Neville E. Sanjana<sup>1,2,\*</sup>

<sup>1</sup>New York Genome Center, New York, NY, USA.

<sup>2</sup>Department of Biology, New York University, New York, NY, USA.

### Abstract

Type VI CRISPR enzymes are RNA-targeting proteins with nuclease activity that enable specific and robust target gene knock-down without altering the genome. To define rules for the design of Cas13d guide RNAs, we conducted massively-parallel screens targeting mRNAs of a green fluorescent protein (GFP) transgene and CD46, CD55 and CD71 cell surface proteins in human cells. In total, we measured the activity of 24,460 guide RNAs with and without mismatches relative to the target sequences. Knock-down efficacy is driven by guide RNA-specific features and target site context. Single mismatches generally reduce knock-down to a modest degree, but spacer nucleotides 15 – 21 are largely intolerant to target site mismatches. We developed a computational model to identify optimal guide RNAs and confirm its generalizability testing 3,979 guides targeting mRNAs of 48 endogenous genes. We show that Cas13 can be used in forward transcriptomic pooled screens and, using our model, predict optimized Cas13 guide RNAs for all protein-coding transcripts in the human genome.

### Editorial summary

Users may view, print, copy, and download text and data-mine the content in such documents, for the purposes of academic research, subject always to the full Conditions of use: [http://www.nature.com/authors/editorial\\_policies/license.html#terms](http://www.nature.com/authors/editorial_policies/license.html#terms)

\*Corresponding author. [neville@sanjanalab.org](mailto:neville@sanjanalab.org).

†These authors contributed equally.

#### Authors contributions

H.H.W and N.E.S conceived the project. H.H.W., N.E.S and A.M.-M. designed the experiments. A.M.-M. and H.H.W. performed the experiments. A.M.-M. and H.H.W analyzed experiments. H.H.W. analyzed the screen data, built the guide RNA prediction software and the online repository. X.G., M.L. and Z.D. helped with post-screen validation experiments. N.E.S. supervised the work. H.H.W. and N.E.S. wrote the manuscript with input from all of the authors.

#### Competing interests

The New York Genome Center and New York University have applied for patents relating to the work in this article. N.E.S. is an advisor to Vertex.

#### Data availability statement

Screen data are being deposited to GEO (<https://www.ncbi.nlm.nih.gov/geo/>) with the accession number GSE142675. All code and software to reproduce our entire analyses are available on our gitlab repository (<https://gitlab.com/sanjanalab/cas13>). Moreover, we provide pre-computed guide RNA predictions targeting all protein-coding transcripts in the human genome on our web-based repository (<https://cas13design.nygenome.org>). Other data and materials that support the findings of this research are available from the corresponding author upon reasonable request.

#### Code availability statement

The predictive on-target model as well as all code for the analyses presented in the paper is available on our gitlab repository (<https://gitlab.com/sanjanalab/cas13>).

Design of optimal guide RNAs for Cas13d-mediated RNA knock-down is enabled by computational modeling of large-scale screening data

Type VI CRISPR enzymes have recently been identified as programmable RNA-guided, RNA-targeting Cas proteins with nuclease activity that allow for target gene knock-down without altering the genome. In addition to target RNA knock-down<sup>1-10</sup>, Cas13 proteins have been used for viral RNA detection<sup>7,9,11,12</sup>, site-directed RNA editing<sup>13</sup>, demethylation of m<sup>6</sup>A-modified transcripts<sup>14</sup>, RNA live-imaging<sup>15,16</sup>, and modulation of splice site choice as well as cleavage and polyadenylation site usage<sup>5,17,18</sup>. Cas13 proteins are guided to their target RNAs by a single CRISPR RNA (crRNA) composed of a direct repeat (DR) stem loop and a spacer sequence (guide RNA or gRNA) that mediates target recognition by RNA-RNA hybridization. Although Cas13 enzymes exert some non-specific collateral nuclease activity upon activation<sup>4-6,11,19</sup>, they have greatly reduced off-target activity in cultured cells compared to RNA interference<sup>2,5,13</sup>. Previous studies have shown that Cas13 gRNAs have minimal Protospacer Flanking Sequence (PFS) constraints<sup>1,4,13,20</sup> and that RNA target sites should be accessible for Cas13 binding<sup>1,2,4</sup>. Beyond these basic parameters, we currently lack information about optimal Cas13 crRNA designs for effective target RNA knock-down.

To date, three Cas13 effector proteins (*Pgt*Cas13b, *Psp*Cas13b, *Rfx*Cas13d) have been reported to show high RNA knock-down efficacy with minimal off-target activity<sup>5,13</sup>. We compared the ability of these Cas13 enzymes to knock-down GFP mRNA when directed to either the cytosol or the nucleus. *Rfx*Cas13d (CasRx) consistently showed the strongest target knock-down, especially when fused to a nuclear localization sequence (NLS) (Supplementary Fig. 1a – c). Using Cas13d-NLS, we varied the gRNA length while maintaining a constant gRNA 5' end or 3' end relative to a 30 nt reference gRNA, and found that 23 – 30 nt gRNAs confer the most pronounced target knock-down (Supplementary Fig. 1d).

To systematically assess the *Rfx*Cas13d target knock-down efficacy of thousands of gRNAs, we established a monoclonal HEK293 cell line expressing destabilized GFP and doxycycline-inducible Cas13d-NLS nuclease. We lentivirally delivered a library of 7,500 crRNAs that target the GFP coding sequence, containing perfect match and mismatch gRNAs (Fig. 1a). We performed fluorescence activated cell sorting (FACS) to gate cells in four bins based on their GFP intensity (Supplementary Fig. 2a). Guide RNA counts showed high concordance between bins across three independent transductions with clear separation of bin 1, which contained cells with the lowest GFP expression (Supplementary Fig. 2b–d).

We calculated the log<sub>2</sub> fold-change (log<sub>2</sub>FC) gRNA enrichment between all bins and the unsorted input gRNA distribution (Supplementary Data 1). Perfect match gRNAs were enriched in bin 1, while increasing numbers of mismatches led to a gradual decrease in gRNA enrichment (Fig. 1b, Supplementary Fig. 3a–c). This was true for the whole gRNA population as well as for individual gRNAs and their corresponding gRNAs with 1 – 3 mismatches (Fig. 1b–c, Supplementary Fig. 3d). As a control, the library also contained 537 non-targeting crRNAs and they were effectively depleted from bin 1 (Fig. 1b, Supplementary Fig. 3a–c). As expected, gRNA abundances in bin 1 were negatively

correlated to those in bins 2 to 4, which contained cells with higher GFP intensities (Supplementary Fig. 3e, f). Taken together, this suggests that the enrichments of gRNAs in bin 1 accurately reflect target mRNA knock-down.

We noticed considerable heterogeneity of gRNA enrichment within each gRNA class (Fig. 1b–c). For perfect match gRNAs targeting different regions of the target mRNA, we observed position-dependent effects, suggesting an influence of the target sequence context on the gRNA efficacy (Fig. 1d). We selected 6 gRNAs along the GFP target transcript with either high or low enrichment and validated their relative target knock-down efficacies by transfection of individual gRNAs followed by flow cytometry (Fig. 1e).

To examine if Cas13 can tolerate mismatches between the gRNA and the target RNA, we calculated the relative  $\log_2$  fold change ( $\log_2\text{FC}$ ) for each mismatch gRNA by subtracting the  $\log_2\text{FC}$  from the reference (perfect match) gRNA (Fig. 1f). We found a critical (“seed”) region for Cas13d knock-down efficacy between gRNA nucleotides 15 to 21 with its center at nucleotide 18 relative to the gRNA 5’ end. Although seed regions have been shown for Cas13a<sup>1,21,22</sup>, one group reported no clear seed region for Cas13d<sup>23</sup> while another group showed position-dependent mismatch sensitivity for Cas13d in a cell-free assay<sup>24</sup>. Within the seed region, single mismatches led to diminished gRNA enrichment, while mismatches outside the seed region were better tolerated (Fig. 1f). The critical region was present irrespective of the mismatch identity (Fig. 1g). Similarly, consecutive double and triple mismatches indicated the presence of the critical region (Fig. 1g, Supplementary Fig. 4a). For randomly distributed double mismatches, the largest change in enrichment was observed in cases where both mismatches are in the seed region (Supplementary Fig. 4b). Increasing the number of mismatches to three largely abrogated target knock-down (Supplementary Fig. 4a). For this reason, the critical region may have been masked in previous studies on *EsCas13d* which tested four consecutive mismatches<sup>23</sup>.

Given the heterogeneity in enrichment for gRNAs with mismatches in the seed region, we sought to assess the effect of surrounding nucleotide context (Supplementary Fig. 5a). Controlling for the reference gRNA efficacy, mismatches in a ‘U’-context at the target site negatively impacted Cas13d activity, whereas mismatches in a ‘GC’-context were better tolerated (Supplementary Fig. 5b). We confirmed the presence of the seed region in transfection experiments using gRNAs with single or double nucleotide mismatches to the GFP mRNA (Fig. 1h). While a perfect match gRNA decreased the percentage of GFP-positive cells to ~29%, a single mismatch at gRNA position 18 resulted in 75% GFP-positive cells and a double mismatch at positions 17 and 18 resulted in ~79% GFP positive cells (Fig. 1h).

Next, we sought to assess the features that may affect knock-down efficacy for perfect match gRNAs (see Supplementary Note 1 for details). One of the features impacting the observed gRNA enrichments in the GFP tiling screen was crRNA folding: Predicted secondary structures and corresponding minimum free energy (MFE) of perfect match crRNAs showed a positive correlation between the MFE and gRNA efficacy (Supplementary Fig. 6a). In particular, ‘G’-dependent structures, such as predicted G-quadruplexes, showed diminished target knock-down. Given that the crRNA folding is critical for effective target knock-down,

we sought to further stabilize and improve the DR by repairing a predicted bulge in the DR, by varying the length of the stem loop or by disrupting bases in the proximal DR stem (Supplementary Fig. 6b). Analysis of the crystal structure of *EsCas13d* and *UfCas13d* together with its crRNA suggested that the terminal loop in the DR may not be embedded within the protein and thus may allow extension (and further stabilization) of the stem loop<sup>23,24</sup> similar to those previously found to enhance Cas9 activity<sup>25,26</sup>. We observed that any change in stem length abrogated target knock-down completely (Supplementary Fig. 6c). Also, repairing the bulged nucleotide within the stem decreased target knock-down. However, disrupting the first base pair within the proximal stem further increased Cas13d targeting efficacy, leading to a novel *RfxCas13d* DR with improved knock-down. We tested the modified DR on 6 additional gRNAs targeting GFP and found that the modified DR improved target knock-down especially for gRNAs with low knock-down efficacy (Supplementary Fig. 6d).

We defined 15 crRNA and target-RNA features based on their correlation with observed gRNA enrichment (Supplementary Table 1, Supplementary Note 1). With these features, we sought to derive a generalizable ‘on-target’ model to predict Cas13d target knock-down. We compared the ability of machine learning approaches to predict gRNA efficacy (see Methods) and found that a Random Forest (RF) model had the best prediction accuracy (Supplementary Fig. 7a), weighting the crRNA folding energy, the local target ‘C’-context, and the upstream target ‘U’-context as the most important features (Supplementary Fig. 7b). Other learning approaches frequently chose similar features, suggesting that these features are the main drivers of Cas13d GFP knock-down (Supplementary Fig. 7c). To identify key predictor of gRNA efficacy, we iteratively reduced the number of features, monitoring the model performance and derived a minimal model that explained about 37% of the variance ( $r^2$ ) with a Spearman correlation ( $r_s$ ) of  $\sim 0.58$  to the held-out data (Fig. 2a, Supplementary Fig. 7d–f). In comparison, a support vector machine (SVM) regression model with a similar structure to a Cas9 gRNA prediction algorithm<sup>27</sup> performed worse when applied to this data ( $r^2 = 0.21$ ,  $r_s = 0.44$ ) (Fig. 2a).

To show that our model is generalizable, we predicted gRNAs to target the endogenous transcripts of *CD46* and *CD71*, which encode cell surface proteins, and measured the gRNA knock-down efficacy by flow cytometry. For each gene, we chose 3 gRNAs predicted to have high knock-down efficacy (top 2 quartiles) and 3 gRNAs predicted to have low knock-down efficacy (lower 2 quartiles). On an individual gRNA level, we found that the majority of gRNAs with higher predicted guide scores suppressed CD46 and CD71 protein expression more robustly than gRNAs with lower guide scores (Fig. 2b). Comparing the observed knock-down across all three high-scoring to all three low-scoring gRNAs, we found a significant improvement for CD71, while for CD46 we observed considerable variance. To increase throughput and test gRNA efficacy predictions for more genes, we first generated a small crRNA library targeting 10 essential and 10 control genes with both 3 high-scoring and 3 low-scoring gRNAs and monitored their depletion in a gene essentiality screen over time. Essential genes were chosen from genes that were strongly depleted in previous RNAi screens<sup>28</sup> (Supplementary Fig. 8a). Most high-scoring gRNAs targeting essential genes were progressively depleted over time, while low-scoring gRNAs showed largely no depletion (Fig. 2c, Supplementary Fig. 8b).

In addition, we performed a second targeted essentiality screen in A375 cells targeting 35 essential and 65 control genes with both 20 high-scoring and 20 low-scoring gRNAs. Similar to the HEK293 screen above, we found that high-scoring gRNAs that target essential genes were progressively depleted over time (Fig. 2d, Supplementary Fig. 8c). Although high-scoring gRNAs were generally more depleted than low-scoring gRNAs on a per gene level, we noticed that not all predicted essential genes showed depletion upon Cas13d targeting (Supplementary Fig. 8c–d), suggesting that RNAi-screen derived essentiality scores may not be directly comparable to Cas13-derived essentiality.

We calculated a significance score of gene depletion based on the gRNA rank consistency for the 20 high-scoring gRNAs and found strong enrichment of defined essential genes at the top of the list (Fig. 2e). The gRNA depletion scores correlated better with the DEMETER2 RNAi<sup>28</sup> scores used to define the set of essential genes to be tested (up to  $r_s = 0.71$  using the best gRNA) than with the Cas9 STARS scores<sup>29</sup> (up to  $r_s = 0.61$ ) (Fig. 2f). Taken together, this suggests that the crRNA and target RNA features derived from the GFP tiling screen can generalize to predict Cas13d gRNA efficacy for novel targets, and that these gRNA predictions can be used in pooled CRISPR-Cas13 screens.

Our predictive on-target model based on the GFP tiling screen was largely able to separate gRNAs with low knock-down efficacy from those with high efficacy. However, given that we observed remaining heterogeneity among the predicted high-scoring gRNAs, we sought to improve our on-target model by expanding our training dataset. Therefore, we performed three additional Cas13d tiling screens targeting the main transcript isoforms of the cell surface proteins CD46, CD55 and CD71 in HEK293 cells coupled with FACS to select cells with decreased surface protein expression (Fig. 3a–c; Supplementary Fig. 9a–c). In addition to perfect match gRNAs, we added several additional gRNA classes (Supplementary Fig. 9a). For each screen, perfect match gRNAs showed the strongest gRNA enrichment relative to the unsorted input samples, while reverse complement negative control gRNAs and non-targeting gRNAs were depleted (Supplementary Fig. 9d). In the new screens we reduced the overall gRNA length to 23 bases and included a set of gRNA length variants ranging in length from 15 to 36 nucleotides. Starting from 23 nucleotide length, gRNAs exerted full knock-down efficacy, while longer gRNA 3' ends did not have any deleterious effects (Supplementary Fig. 9e).

Perfect match gRNAs targeting coding regions (CDS) were more enriched compared to gRNAs targeting untranslated regions (UTRs) or introns (Supplementary Fig. 9f). UTR-targeting gRNAs may show lower enrichments as each target gene may be represented by multiple transcript isoforms with alternative UTR usage. Hence, gRNAs targeting coding regions have a higher likelihood to find the cognate target site while, for example, 3' UTR-targeting gRNAs find their target site only in a fraction of the expressed transcripts isoforms. Accordingly, the low enrichment for intron-targeting gRNAs may be explained by the short-lived nature of introns. For these gRNAs, the intronic target site is present only for a short period of time, which likely enables the transcript to evade Cas13 targeting. For this reason, gRNA knock-down efficacy may not be directly comparable between CDS-targeting gRNAs and UTR-or intron-targeting gRNAs. Across all 39 introns present, we found that intron-targeting gRNAs were only mildly enriched. In these introns, we observed a slight decrease

in gRNA efficacy immediately downstream of the 5'-splice-site and within -50 to 0 nucleotides upstream of the 3'-splice-site (Supplementary Fig. 9g). These sites are typically bound by the spliceosome<sup>30</sup>, suggesting that gRNAs targeting these regions may compete with the splicing machinery and other splice factors for target sequences. As transcript maturation in the nucleus seemingly influences the gRNA targeting efficacy, we wondered if the exon-junction-complex (EJC) would affect knock-down of the matured transcript in the same way. The EJC typically binds ~20–24 nucleotides 5' upstream to the exon-exon-junction upon splicing<sup>31,32</sup>. Indeed, we observed a depletion of high-scoring gRNAs within a window of -20 to 0 nucleotides 5' of the exon junction (Supplementary Fig. 9h).

To improve our on-target model, we focused on perfect match gRNAs that target CDS-regions and increased the number of high-confidence model input observations from ~400 to nearly 3000. Similar to the initial GFP-screen, gRNAs efficacies were distributed along the coding region in a non-random manner (Fig. 3a–c). We repeated the assessment of features that may affect knock-down efficacy (see Supplementary Note 2 for details). Notably, the increased number of observations uncovered positional nucleotide preferences (Supplementary Fig. 10a–b). Guide RNA enrichments correlated positively with G- and C-base probabilities in the seed region around gRNA position 18. Surrounding this region, U- and A-base probabilities correlated positively with the target knock-down. We derived an updated on-target model using 2,918 CDS-targeting gRNAs across all four tiling screens and selected 35 out of 644 evaluated features in a similar fashion as before (see methods) (Supplementary Table 2, Supplementary Note 2, Supplementary Data 2).

The combined Random Forest model (RF<sub>combined</sub>) displayed improved prediction accuracy compared to the initial RF<sub>minimal</sub> model (from here on referred to as RF<sub>GFP</sub>) explaining ~47% of the variance ( $r^2$ ) with a Spearman correlation ( $r_s$ ) of ~0.67 to the held-out data (Fig. 3d, Supplementary Fig. 10c). Using 10-fold cross-validation, the model effectively separated low-scoring gRNAs from high-scoring gRNAs, assigning 63% of the gRNAs correctly to the highest efficacy-quartile (Fig. 3e). Similarly, the predicted guide scores of the top- or bottom-ranked gRNAs (ranked by the observed knock-down efficacy) separate gRNAs that performed well from those that performed poorly more than expected by chance (Supplementary Fig. 10d). Further, we performed leave-one-out cross-validation training on three data sets while predicting guide scores for the held-out fourth screen. The RF<sub>combined</sub> model generalized well for endogenous genes (mean  $\pm$  sd:  $r_s = 0.63 \pm 0.01$ ) but was less predictive for the GFP transgene ( $r_s = 0.33$ ) (Supplementary Fig. 10e).

Finally, we compared the ability of both models, RF<sub>GFP</sub> and RF<sub>combined</sub>, to correctly predict the knockdown efficacies for the two essentiality screens. Both screens were designed based on gRNA predictions made by the RF<sub>GFP</sub> model. In both cases, the RF<sub>combined</sub> was in better agreement with the observed knock-down efficacies across all genes (Fig. 3f). Likewise, we found that the RF<sub>combined</sub> showed improved agreement with the observed gRNA depletion also on a gene level for the 10 most depleted genes in the A375 fitness screen (RF<sub>GFP</sub>:  $r_s = 0.46 \pm 0.16$ , RF<sub>combined</sub>:  $0.58 \pm 0.14$ ). Taken together, we show that our updated RF<sub>combined</sub> on-target model is able to predict Cas13d gRNA target knock-down efficacies, separating poorly performing gRNAs from gRNAs with high efficacy and generalizing across numerous targets.

We applied our model and predicted gRNAs for all protein-coding transcripts in the human genome (GENCODE v19). We made these predictions available through a user-friendly, web-based application (<https://cas13design.nygenome.org>). In addition, we report the 10 highest-scoring crRNAs for the 5' UTR, CDS and 3' UTR of each transcript (Supplementary Fig. 11a, Supplementary Data 3). We partitioned the predicted gRNAs according to the efficacy quartiles in our four screens. Only 15.2% of all possible gRNAs fall into the highest scoring (best knock-down) quartile (Q4) (Supplementary Fig. 11b). A large fraction of gRNAs are predicted to have lower efficacy (36.8% of all gRNAs are in Q1 or Q2), which emphasizes the value of optimal gRNA selection for high knock-down efficacy. However, almost all transcripts contain top-scoring gRNAs (Supplementary Fig. 11c).

Taken together, we performed a set of pooled screens for CRISPR-Cas13d and defined targeting rules for optimal gRNA design. We show that crRNA features and target RNA-context constrain target knock-down efficacy and, using this data, we developed a model to predict gRNAs with high efficacy. We validated this model using pooled Cas13d screens and compared the ability of Cas13 perturbations to identify a set of essential genes with prior RNAi and Cas9 screens. While all three perturbations methods broadly agree, it is important to note that a comprehensive genome-wide comparison is pending. An important distinction between RNA-targeting approaches is that, while RNAi is restricted to the cytosol, Cas13 allows for compartmentalized targeting (nucleus, cytosol, and other sub-cellular compartments) and sophisticated transcriptome engineering with catalytically dead (dCas13) effector fusions. Overall, our study provides a detailed characterization of Cas13 targeting and a predictive model for high activity gRNAs, yielding a valuable platform for the design of massively-parallel RNA-targeting screens.

## Online Methods

### Cloning of Cas13 nuclease, guide RNAs and destabilized EGFP plasmids

Using Gibson cloning, we modified the EF1a-short (EFS) promoter-driven lentiCRISPRv2 (Addgene 52961) or lentiCas9-Blast (Addgene 52962) plasmids with several different transgenes<sup>33</sup>. For the destabilized EGFP construct, we introduced a PEST sequence and nuclear localization tag on EGFP to create EFS-EGFPd2PEST-2A-Hygro (pLentiEGFPdestabilized) from lentiCas9-Blast. To test the upstream U-content, we introduced a multiple cloning site (MCS) into pLentiEGFPdestabilized right after the stop codon, and used the MCS to introduce oligonucleotide sequences with variable U-content.

For the CRISPR Type-VI orthologs, we cloned effector proteins (*PguCas13b*: Addgene 103861, *PspCas13b*: Addgene 103862, *RfxCas13d*: Addgene 109049) and their direct repeat (DR) sequences (*PguCas13b*: Addgene 103853, *PspCas13b*: Addgene 103854, *RfxCas13d*: Addgene 109053) into lentiCRISPRv2. In this manner, we created pLentiRNACRISPR constructs: hU6-[Cas13 DR]-EFS-[Cas13 ortholog]-[NLS/NES]-2A-Puro-WPRE, where [Cas13 ortholog] was one of *PguCas13b*, *PspCas13b*, or *RfxCas13d* and [NLS/NES] was either a nuclear localization signal or nuclear export signal. To generate doxycycline-inducible Cas13d cell lines, we cloned NLS-*RfxCas13d*-NLS (Addgene 109049) into TetO-[Cas13]-WPRE-EFS-rtTA3-2A-Blast. For the screens, we changed the DR in the lentiGuide-Puro vector (Addgene 52963) to contain the *RfxCas13d* DR using Gibson

cloning to create lentiRfxGuide-Puro (pLentiRNAGuide)<sup>33</sup>. All plasmids will be made available on Addgene.

Guide cloning was done as described previously<sup>33</sup>. All constructs were confirmed by Sanger sequencing. All primers used for molecular cloning and guide sequences are shown in Supplementary Data 4. The main plasmids used in this work and described above have been submit to Addgene.

### Cell culture and monoclonal cell line generation

HEK293FT cells were acquired from Thermo Fisher Scientific (R70007) and A375 cells were acquired from ATCC (CRL-1619). HEK293FT and A375 cells were maintained at 37°C with 5% CO<sub>2</sub> in D10 media: DMEM with high glucose and stabilized L-glutamine (Caisson DML23) supplemented with 10% fetal bovine serum (Serum Plus II Sigma-Aldrich 14009C) and no antibiotics.

To generate doxycycline-inducible *RfxCas13d*-NLS HEK293FT and A375 cells, we transduced cells with a *RfxCas13d*-expressing lentivirus at low MOI (<0.1) and selected with 5µg/mL Blasticidin S (ThermoFisher A1113903). Single cell colonies were picked after by sparse plating. Clones were screened for Cas13d expression by western blot using mouse anti-FLAG M2 antibody (Sigma F1804).

For the GFP tiling screen *RfxCas13d*-expressing cells were transduced with pLentiEGFPdestabilized lentivirus at low MOI (<0.1) and selected with 100µg/ml Hygromycin B (ThermoFisher 10687010) for 2 days. Single-cell colonies were grown by sparse plating. Resistant and GFP-positive clonal cells were expanded and screened for homogenous GFP expression by flow cytometry.

### Transfection and flow cytometry

For all transfection experiments, we seeded 2×10<sup>5</sup> HEK293FT cells per well of a 24-well plate prior to transfection (12 – 18 hours) and used 500 or 750 ng plasmid together with a 5-to-1 ratio of Lipofectamine 2000 (ThermoFisher 11668019) or 1 mg/mL polyethylenimine (Polysciences 23966) to DNA (e.g. 2.5µl Lipofectamine2000 or PEI mixed with 0.5µg plasmid DNA). Flow cytometry or fluorescence-assisted cell sorting (FACS) was performed at 48 hrs post-transfection. All transfection experiments were performed in biological triplicates.

For the CRISPR Type-VI ortholog comparison (Supplementary Fig. 1a–c), we cloned the effector proteins (*PguCas13b*: Addgene 103861, *PspCas13b*: Addgene 103862, *RfxCas13d*: Addgene 109049) and their direct repeat sequences (*PguCas13b*: Addgene 103853, *PspCas13b*: Addgene 103854, *RfxCas13d*: Addgene 109053) as described above. We co-transfected the pLentiRNACRISPR constructs together with a GFP expression plasmid (pLentiEGFPdestabilized) in a 2:1 molar ratio. The guide RNA length comparison (Supplementary Fig. 1d) was done using previously published *RfxCas13d* constructs (Addgene 109049 and 109053), except that we removed the GFP cassette from the *RfxCas13d* plasmid. The modified *RfxCas13d* construct and guide plasmids were co-transfected together with pLentiEGFPdestabilized in a 2:2:1 molar ratio. For the DR



modification experiment (Supplementary Fig. 6c) we transfected *RfxCas13d* expressing cells, starting doxycycline-induction (1µg/ml) at the time of cell plating. The guide plasmid and GFP expression plasmid were co-transfected at a 1:1 molar ratio.

For the model validation flow cytometry (Fig. 2b) and CD46 screen validation (Supplementary Fig. 9c) we transfected *RfxCas13d*-expressing cells with a guide RNA expressing plasmid. 48 hours post transfection, the cells were stained for the respective cell surface protein for 30 min at 4°C and measured by FACS. (BioLegend: CD46 #352405 clone TRA-2-10, CD71 (TFRC) #334105 clone CYIG4).

For the screen result validation (Fig. 1e) and seed validation experiments (Fig. 1h) we co-transfected *RfxCas13d*-expressing cells with a guide RNA expressing plasmid and pLentiEGFPdestabilized at a 1:1 molar ratio. At 48 hours post-transfection, the cells were analyzed by flow cytometry.

To assess the upstream U-context (Supplementary Note 1), we transfected upstream-U context modified pLentiEGFPdestabilized-MCS plasmid together with either a crRNA plasmid into *RfxCas13d*-expressing in a 2:1 molar ratio. Each GFP-upstreamU-context plasmid was co-transfected with both a targeting or a non-targeting guide RNA used for calculating the knock-down, as a change in 3'UTR uridine content could attract RNA-binding proteins that may affect RNA stability independent of Cas13. We selected the zero-uridine oligonucleotide from a set of 10000 *in silico* randomized 52mers with {A<sub>24</sub>,C<sub>14</sub>,G<sub>14</sub>} with minimal predicted RNA-secondary structure as determined by RNAfold<sup>34</sup> with default setting.

For flow cytometry analysis, cells were gated by forward and side scatter and signal intensity to remove potential multiplets. If present, cells were additionally gated with a live-dead staining (LIVE/DEAD Fixable Violet Dead Cell Stain Kit, Thermo Fisher L34963). For each sample we analyzed at least 5000 cells. If cell numbers varied, we randomly sampled all samples to the same number of cells before calculating the mean fluorescence intensity (MFI). For GFP co-transfection experiments, we only considered the percentage of transfected cells with the highest GFP expression determined by comparing the non-targeting control to wild-type control cells. For the upstream U-context co-transfection experiments, we considered the whole cell populations.

For knock-down experiments of endogenous genes (Fig. 2b and Supplementary Fig. 9c), we determined the percentage of transfected cells with lower target gene signal than the non-targeting control in the condition with the highest observed knock-down. For all conditions, we analyzed the same bottom percentage of cells. For the selected cells, we compared the MFI of targeting guides relative to non-targeting guides to determine the percent knock-down. To directly compare relative rank of individual guides as done in Fig. 2b, we normalized the effect size by setting the most effective guide to 100%. For the seed validation (Fig. 1f), we determined the percentage of transfected (GFP-positive) cells with GFP signal higher than Lipofectamine vehicle treated control cells. The percentage of transfected cells was normalized to percentage of GFP-positive cells in the non-targeting guide control.

## Screen library design and pooled oligo cloning

To design the *Rfx*Cas13d guide RNA library for GFP, we selected the 714 bp coding sequence (without start codon) to be targeted. *In silico*, we generated all perfectly matching 27mer guide RNAs with minimal constraints (T-homopolymer < 4, V-homopolymer < 5, 0.1 < GC-content < 0.9) and selected 400 by random sampling. From these, we sampled 100 guide RNAs and introduced one random nucleotide conversion at each position (SM set  $n = 2,700$ ). From these 100, we randomly sampled 17 guide RNAs and introduced 26 or 25 consecutive double (CD set  $n = 442$ ) and triple (CT set  $n = 425$ ) mismatches, respectively. We sampled an additional 13 guide RNAs from the SM set (in total, 30 guide RNAs) and introduced 100 random double mismatches at any position for each guide RNA if not present already in the set of 17 consecutive double mismatches (RD set  $n = 3,000$ ). In total, we designed 6,967 GFP targeting guides and added 533 non-targeting guides (NT set) of the same length from randomly generated sequences that did not align to the human genome (hg19) with less than 3 mismatches.

For CD46, CD55 and CD71 library design, we selected the transcript isoform with highest isoform expression in HEK-TE samples (determined by Cancer Cell Line Encyclopedia CCLE; GENCODE v19) and longest 3'UTR isoform (CD46: ENST00000367042.1, CD55: ENST00000367064.3, CD71: ENST00000360110.4). As described above, we generated all perfectly matching 23mers, and selected ~2000 evenly spaced guide RNAs per target. In addition to PM, SM, RD and NT sets as described above, we included for each target a set of guide length variants (LV set  $n = 450$ ), guide RNAs targeting intronic sequences near splice-donor and splice-acceptor sites across all 39 annotated introns (I set  $n = 2,122$ ) and an additional negative control set of reverse complementary perfect match sequences (RC set  $n = 300$ ). Further details are in Supplementary Data 5.

For both targeted essentiality screens, we used the DEMETER2 v5<sup>28</sup> data set from the Cancer Dependency Map portal (DepMap) to determine essential and control genes. Specifically, we selected essential genes with low  $\log_2$  fold-change (FC) enrichments across all cell lines and in the respective assay cell line (Supplementary Fig. 8a,c). For our HEK293FT cells, we considered data for HEK-TE cells. Furthermore, we selected genes with one transcript isoform constituting more than 75% of the gene expression with expression level less than ~150 transcripts per million (TPM). We predicted guide RNA efficiencies using the minimal  $RF_{GFP}$  model and removed all guides with matches or partial matches elsewhere in the transcriptome. We allowed up to 3 mismatches when looking for potential off-targets. From the set of remaining perfect match guide RNA predictions, we manually selected three high-scoring and three low-scoring guides for the HEK293FT cell line screen to ensure that each guide fell into non-overlapping regions of the target transcripts. For the A375 cell line targets, we selected the top 20 high-scoring guide RNAs. For the set of 20 low-scoring guides, we chose among the bottom 60 to reduce the overlap of guide RNAs that fall into the same region. In this way, we assayed 20 genes in HEK293FT cells targeting 10 essential and 10 control genes with three low-scoring and three high-scoring guides, as well as three non-targeting guides ( $n = 123$ ). For the A375 screen, we targeted 100 genes (35 essential and 65 control genes) with 40 guides each (20 high-and 20 low-scoring) and included 680 non-targeting sequences ( $n = 4,680$ ).

All large-scaled pooled crRNA libraries were synthesized as single-stranded oligonucleotides (Twist Biosciences), PCR amplified using NEBNext High-Fidelity 2X PCR Master Mix (M0541S) and Gibson cloned into pLentiRfxGuide-Puro. The guides for the HEK293FT essentiality screen were ordered from IDT, array cloned, confirmed by Sanger sequencing, and subsequently pooled using equal amounts. Complete library representation with minimal bias (90<sup>th</sup> percentile/10<sup>th</sup> percentile crRNA read ratio: 1.68 – 2.17) were verified by Illumina sequencing (MiSeq).

### Pooled lentiviral production and screening

Lentivirus was produced via transfection of library plasmid with appropriate packaging plasmids (psPAX2: Addgene 12260; pMD2.G: Addgene 12259) using polyethylenimine (PEI) reagent in HEK293FT. At 3 days post-transfection, viral supernatant was collected and passed through a 0.45  $\mu$ m filter and stored at  $-80^{\circ}\text{C}$  until use.

Doxycycline-inducible *RfxCas13d*-NLS human HEK293FT, double-transgenic HEK293FT-GFP or A375 cells were transduced with the respective library pooled lentiviruses in separate infection replicates ensuring at least 1000x guide representation in the selected cell pool per infection replicate using a standard spinfection protocol. We generated either 2 or 3 independent replicate experiments. After 24 hours, *RfxCas13d* expression was induced by addition of 1  $\mu$ g/ml doxycycline (Sigma D9891) and cells were selected with 1  $\mu$ g/ml puromycin (ThermoFisher A1113803), resulting in  $\sim$ 30% cell survival. Puromycin-selection was complete  $\sim$ 48 post puromycin-addition. Assuming independent infection events (Poisson), we determined that  $\sim$ 83% of surviving cells received a single sgRNA construct. Cells were passaged every two days maintaining at least the initial cell representation and supplemented with fresh doxycycline.

The tiling screens were terminated after 5 to 10 days. For all targets we noted maximal knock-down after 2–4 days (data not shown). For cell surface proteins, cells were stained in batches of  $1 \times 10^7$  cells for 30 min at  $4^{\circ}\text{C}$  (BioLegend: CD46 clone TRA-2-10 #352405 – 3  $\mu$ l per  $1 \times 10^6$  cells; CD55 clone JS11 #311311 – 1.5  $\mu$ g per  $1 \times 10^6$  cells; CD71 clone CYIG4 #334105 – 4  $\mu$ l per  $1 \times 10^6$  cells). We collected unsorted samples for input guide RNA representation of approximately 1000x coverage for each sample and sorted at least another 1000x representation into the assigned bins based on their signal intensities (GFP: lowest 20%, 20%, 20% and remaining highest 40%, Supplementary Fig. 2a; CD proteins lowest 20% and highest 20%, Supplementary Fig. 9b; Supplementary Data 5). Cells were PBS-washed and frozen at  $-80^{\circ}\text{C}$  until sequencing library preparation. In each case, the bin containing the lowest 20% represented the strongest target knock-down.

The essentiality screens were started (Day 0) upon complete puromycin selection, which was at 5 days after transduction. Cells were passaged every two to three days maintaining at least the initial cell representation and supplemented with fresh doxycycline. At Day 0 (=Input) and every 7 days, we collected a  $>1000x$  representation from each sample. The HEK293FT cell screen was conducted in triplicate and cultured for 4 weeks. The A375 cell screen was conducted in duplicate and cultured for 2 weeks.

## Screen readout and read analysis

For each sample, genomic DNA was isolated from sorted cell pellets using the GeneJET Genomic DNA Purification Kit (ThermoFisher K0722) using  $2 \times 10^6$  cells or less per column. The crRNA readout was performed using two rounds of PCR<sup>35</sup>. For the first PCR step, a region containing the crRNA cassette in the lentiviral genomic integrant was amplified from extracted genomic DNA using the PCR1 primers in Supplementary Data 4.

For each sample, we performed PCR1 reactions as follows: 20  $\mu$ l volume with 2  $\mu$ g of gDNA in each reaction limited by the amount of extracted gDNA (total gDNA ranged from 8  $\mu$ g to 50  $\mu$ g per sample with an estimated representation of  $10^6$  diploid cells per  $\sim 6.6$   $\mu$ g gDNA. PCR1: 4  $\mu$ l 5x Q5 buffer, 0.02U/ $\mu$ l Q5 enzyme (M0491L), 0.5uM forward and reverse primers and 100ng gDNA/ $\mu$ l. PCR conditions: 98°C/30s, 24x[98°C/10s, 55°C/30s, 72°C/45s], 72°C/5min).

We pooled the unpurified PCR1 products and used the mixture for a single second PCR reaction per sample. This second PCR adds on Illumina sequencing adaptors, barcodes and stagger sequences to prevent monotemplate sequencing issues. Complete sequences of the 5 forward and 3 reverse Illumina PCR2 readout primers used are shown in Supplementary Data 4. (PCR2: 50  $\mu$ l 2x Q5 master mix (NEB #M0492S), 10  $\mu$ l PCR1-product, 0.5uM forward and reverse PCR2-primers in 100  $\mu$ l. PCR conditions: 98°C/30s, 17x[98°C/10s, 63°C/30s, 72°C/45s], 72°C/5min).

Amplicons from the second PCR were pooled by screen experiment (e.g. all GFP-screen samples) in equimolar ratios (by gel-based band densitometry quantification) and then purified using a QiaQuick PCR Purification kit (Qiagen 28104). Purified products were loaded onto a 2% E-gel and gel extracted using a QiaQuick Gel Extraction kit (Qiagen 28704). The molarity of the gel-extracted PCR product was quantified using KAPA library quant (KK4824) and sequenced on an Illumina NextSeq 500 -II MidOutput 1  $\times$  150 v2.5.

Reads were demultiplexed based on Illumina i7 barcodes present in PCR2 reverse primers using bcl2fastq and by their custom in-read i5 barcode using a custom python script. Reads were trimmed to the expected guide RNA length by searching for known anchor sequences relative to the guide sequence using a custom python script. For the tiling screens, pre-processed reads were either aligned to the designed crRNA reference using bowtie<sup>36</sup> (v.1.1.2) with parameters -v 0 -m 1 or collapsed (FASTX-Toolkit) to count perfect duplicates followed by string-match intersection with the reference to retain only perfectly matching and unique alignments. Pre-processed guide RNA sequences from the essentiality screens were aligned allowing for up to 1 mismatch (-v 1 -m 1). Alignment statistics are available in Supplementary Data 6. The raw guide RNA counts (Supplementary Data 7) were normalized separated by screen dataset using a median of ratios method like in DESeq2<sup>37</sup> and underwent batch-correction using combat implemented in the SVA R package<sup>38</sup>. Non-reproducible technical outliers were removed by applying pair-wise linear regression for each sample after normalization and batch-correction, collecting the residuals and taking the median value for each guide RNA across all sample-centric comparisons. We removed all crRNA counts within the top  $X\%$  residuals across all samples (GFP: 2%, CD proteins: 0.5%, Essentiality screen: no outlier removal). For the GFP screen, we only remove outliers on a

per-sample basis as needed (but not the entire guide RNA). For CD46, CD55 and CD71 screens, since the number of outliers was small, we decided to remove the entire guide RNA from the analysis. The table below indicates all filtering applied:

Screen	not detected*	<N reads in input samples*	0 reads in any sample*	masked outlier	filtered	total	remaining
GFP	0	not applied	not applied	4**	4	7500	7496
CD46	19	427 (<50)	22	77	545	5605	5060
CD55	23	88 (<50)	0	79	190	5356	5166
CD71	3	48 (<50)	0	75	123	5999	5876
HEK293	0	0 (<50)	0	0	0	123	123
A375	2	10 (<100)	0	0	12	4680	4668

\* Removed before normalization

\*\* filtered for Bin1 guide RNAs

Processed crRNA counts are available in Supplementary Data 8. Guide RNA enrichments were calculated building the count ratios between a bin or timepoint and the corresponding input sample and  $\log_2$ -transformation ( $\log_2FC$ ). Consistency between replicates was estimated using robust rank aggregation (RRA)<sup>39</sup>. Delta  $\log_2FC$  for mismatching guides was calculated by subtracting the  $\log_2FC$  of the perfectly matching reference guide. For the tiling screens, all plots and analyses were performed using the mean guide RNA enrichments of bin 1 (= bottom 20%) across replicates, unless indicated otherwise. Similarly, we used the mean guide RNA enrichments relative to Day 0 across replicates for the essentiality screen. Guide RNA enrichment scores ( $\log_2FC$ ) are available in Supplementary Data 1. In all combined analyses across all four tiling screens, we scaled the observed  $\log_2FC$  separately to improve comparability. For the generation of a the combined on-target model, we normalized the  $n = 2,918$  selected CDS-targeting guides RNA across the four tiling screens to the same scale prior to training and testing the model. To do so, for each dataset  $D$ , we computed the upper and lower quartiles of the guide  $\log_2FC$  ( $UQ_D$  and  $LQ_D$ , respectively) as well as the corresponding quartiles for the  $\log_2FC$  among all datasets pooled together ( $UQ_P$  and  $LQ_P$ ). We then updated each fold change  $x$  as follows:  $x^{\wedge} = [(x - LQ_D) / (UQ_D - LQ_D)] * (UQ_P - LQ_P) + LQ_P$ . By centering on quartiles, this procedure normalized the fold-change distributions in a way that was less susceptible to the influence of outliers of a single screen.

### Predicting RNA secondary structures and RNA-RNA hybridization energies

crRNA secondary structure and minimum free energies (MFEs) was derived using RNAfold [ --gquad ] on the full-length crRNA (DR + guide) sequence<sup>34</sup>. For building the combined on-target model and for testing the RF<sub>GFP</sub> model on the combined data set, we assumed 23mer guide RNAs for all guides in the GFP tiling screen to prevent length dependent differences in the crRNA MFE. Target RNA unpaired probability (accessibility) was calculated using RNAplfold [ -L 40 -W 80 -u 50 ] as described before<sup>40</sup>. 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 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 guide RNA  $\log_2$ FC for each point and window relative to the guide RNA. RNA-RNA-hybridization between the guide RNA and its target site was calculated using RNAhybrid [ -s -c ]<sup>41</sup>. For the hybridization calculation, we did not include the direct repeat of the crRNA. 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 crRNA  $\log_2$ FC or using partial correlation to account for the crRNA folding MFE. In each case, we computed the Pearson correlation.

### Assessing guide RNA nucleotide composition

Guide RNA composition was derived by calculating the nucleotide probability within the respective guide RNA sequence length. To assess the presence of sequence constraints similar to a previously described anti-tag<sup>20</sup> or 5' and 3' Protospacer Flanking Sequences (PFS), we ranked all perfectly matching guide RNAs by their  $\log_2$ FC enrichment within each screen separately. We selected the top and bottom 20% enriched/depleted guide RNAs and calculated the positional nucleotide probability for the four nucleotides upstream and downstream relative to the guide RNA match. To assess nucleotide preferences at any guide RNA match position in addition to upstream and downstream nucleotides, we selected the top 20% of the  $\log_2$ FC-ranked perfectly matching guides as described above and calculated nucleotide preferences as described before<sup>27</sup>. In brief, we calculated the probability of each nucleotide at each position for the top guide RNAs and all guide RNAs. The effect size is the difference of nucleotide probability by subtracting the values from all guides from the top guides (delta  $\log_2$ FC).  $p$ -values were calculated from the binomial distribution with a baseline probability estimated from the full-length GFP mRNA target sequence for all perfectly matching crRNAs.  $p$ -values were adjusted using a Bonferroni multiple hypothesis testing correction.

### 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 crRNA). 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  $\log_2$ FC crRNA enrichments for all perfect match crRNAs, either directly or using partial correlation accounting for crRNA folding MFE. In each case we used Pearson correlation.

The RNA context around single nucleotide mismatches was assessed accordingly with a slight modification. Here, the nucleotide context was assessed relative to mismatch position summarizing the nucleotide probability in a window of 1 to 15 nucleotides to either side (e.g.  $p = 18$  with  $w = 5$  summarizes the nucleotide content in a window of 11 nucleotides from 23 to 13). For more details on  $p$  and  $w$ , please see the diagram in Supplementary Figure 5b. We used all 2,700 single nucleotide mismatch guides in the GFP tiling screen

(100 guide RNAs x 27 mismatched positions per guide). The nucleotide context of each position and each window size was then correlated with the observed delta log<sub>2</sub>FC relative to the perfectly matching reference guide RNA, either directly or using partial correlation accounting for crRNA folding MFE. In each case, we used Pearson correlation.

### On-target model selection

An explanation for all selected features for the RF<sub>GFP</sub> and RF<sub>combined</sub> model can be found in Supplementary Table 1 and Supplementary Table 2, respectively. The RF<sub>combined</sub> model feature input values can be found in Supplementary Data 2. All continuous feature scores were scaled to the [0, 1] interval limited to the 5<sup>th</sup> and 95<sup>th</sup> percentile, with a mean set to the 5<sup>th</sup> percentile. Scaled values exceeding the [0, 1] interval were set to 0 or 1, respectively.

Scaling parameters used to normalize data to the [0, 1] interval for the Random Forest Models

Model	Feature	5 <sup>th</sup> percentile	95 <sup>th</sup> percentile
RF <sub>GFP</sub>	crRNA MFE	-23.4000	-14.5000
	Local A probability	0.0000	0.4286
	Local C	0.2273	0.5000
	Local G probability	0.1429	0.4286
	Local U probability	0.0556	0.2778
	upstream U probability	0.0667	0.2000
RF <sub>combined</sub>	crRNA MFE	-20.3000	-12.8000
	Log <sub>10</sub> Unpaired probability	-7.5546	-1.6244
	hybridization MFE nt 3-15	-29.4000	-17.9000
	hybridization MFE nt 15-23	-21.8000	-12.3000
	Local A <sub>max</sub> probability	0.0000	0.5000
	Local C <sub>max</sub> probability	0.0000	0.5000
	Local G <sub>max</sub> probability	0.0000	0.6667
	Local U <sub>max</sub> probability	0.0833	0.5000
	Local AU <sub>max</sub> probability	0.2727	0.7272
	Local GC <sub>max</sub> probability	0.2222	0.7778
	Local A <sub>min</sub> probability	0.0000	0.5714
	Local C <sub>min</sub> probability	0.0000	0.5556
	Local G <sub>min</sub> probability	0.0000	0.4444
	Local U <sub>min</sub> probability	0.0000	0.5000
	Local AU <sub>min</sub> probability	0.2222	0.7778
	Guide A nt probability	0.0870	0.4348
	Guide C nt probability	0.0870	0.3913
	Guide G nt probability	0.0870	0.4348
	Guide AA di-nt probability	0.0000	0.1818
	Guide AC di-nt probability	0.0000	0.1364
Guide AG di-nt probability	0.0000	0.1818	

Model	Feature	5 <sup>th</sup> percentile	95 <sup>th</sup> percentile
	Guide AU di-nt probability	0.0000	0.1364
	Guide CA di-nt probability	0.0000	0.1818
	Guide CC di-nt probability	0.0000	0.1364
	Guide CG di-nt probability	0.0000	0.1364
	Guide CU di-nt probability	0.0000	0.1364
	Guide GA di-nt probability	0.0000	0.1364
	Guide GC di-nt probability	0.0000	0.1818
	Guide GG di-nt probability	0.0000	0.1818
	Guide GU di-nt probability	0.0000	0.1364
	Guide UA di-nt probability	0.0000	0.1364
	Guide UC di-nt probability	0.0000	0.1818
	Guide UG di-nt probability	0.0000	0.1818

To evaluate and compare model performances, we randomly sampled 1,000 bootstrap datasets from the data of perfect match guide RNA  $\log_2$ FC response values and selected features. We used 399 data points for the initial RF<sub>GFP</sub> model and 2918 data points for all CDS-annotating perfect match guides across the four tiling screens. For the RF<sub>combined</sub> model we normalized the observed  $\log_2$ FC values data prior to training and testing as described earlier. Normalized response values showed better generalizability compared to unnormalized or scaled  $\log_2$ FC. For each bootstrap sample, 70% of the data was used for training and the remaining 30% of the data was held out for testing, ensuring a 70/30 split for each screen dataset when testing the RF<sub>combined</sub> model. Linear dependencies between features were identified using the function `findLinearCombos` from the R package `caret` and removed. The model performance was evaluated by calculating the Spearman correlation coefficient  $r_s$  and Pearson  $r^2$  to the held-out data. We compared a variety of different methods<sup>40</sup> (see Table below).

#	Name	Function	Parameter	R package
1	all subsets regression, maximizing the Bayesian information criterion (BIC)	<i>regsubsets</i>	<code>nvmax=15, nbest=1, method="forward", really.big=T</code>	leaps
2	stepwise regression, maximizing the BIC	<i>stepAIC</i>	–	MASS
3	stepwise regression, maximizing the Akaike information criterion (AIC)	<i>stepAIC</i>	–	MASS
4	Lasso regression	<code>cv.glmnet</code>	<code>family="gaussian", nfolds=10, alpha = 1</code>	glmnet
5	multivariate adaptive regression splines (MARS)	<code>earth</code>	<code>degree = 1, trace = 0, nk = 500</code>	earth
6	Random Forest	<code>randomForest</code>	–	randomForest
7	principal component regression (PCR)	<code>pcr</code>	<code>ncomp = 5 (during prediction)</code>	pls
8	Partial Least Squares (PLSR)	<code>pls</code>	<code>ncomp = 5 (during prediction)</code>	pls



#	Name	Function	Parameter	R package
9	Support Vector Machine w/ L1 loss function (SMV+L1)	tune	method=svm, ranges = list(epsilon=seq(0,1,0.025), cost=2^(2:8)), kernel="radial"	e1071

For both models, we tested a variety of feature combinations including crRNA folding energies, RNA-RNA hybridization energies, target site accessibility, overall and positional (di-)nucleotide probabilities, and one-hot encoding for single and di-nucleotide of the guide target-sites and their upstream and downstream flanking four nucleotides. Together, these represented 644 features for the combined on-target model. A full set of features for the combined on-target model can be found in Supplementary Data 2. For the initial on target model based on the GFP screen data, we evaluate a set of 15 defined features (Supplementary Table 1) along-side with one-hot encoded positional nucleotide information and GC content. These 15 features were defined based on their positive or negative correlation to the observed response value during the data exploration (see also Supplementary Note 1). We iteratively reduced the numbers of features from 15 to 6 for the RF<sub>GFP</sub> model and monitored the model performance as described above. At each iteration, the Random Forest model performed slightly better than any other learning approach. Reducing the features to fewer than the selected 6 features (RF<sub>minimal</sub> = RF<sub>GFP</sub>) reduced the model performance. For the combined on-target model, we did not iteratively reduce the set of 35 selected features. We compared the RF<sub>GFP</sub> model to an SVM+L1 model similar to one of the first CRISPR-Cas9 on-target model. Specifically, we used one-hot encoding for all 35 (31) nucleotide positions considered (27 guide RNA positions in case of the GFP screen and 23 for the combined model and 8 additional positions with 4 upstream and 4 downstream nucleotides). Considering all positions, the feature space contained 140 single nucleotide features, 544 di-nucleotide features and the GC-content (685 non-all-zero features). Here, we used tuning (see table below for parameters) to increase model performance for SVM +L1 specifically. Here, but also for the combined model, one-hot encoded features did not lead to high Spearman correlation coefficient  $r_s$  to the held-out data.

For further evaluation of the random forest models we used 10-fold cross-validation by randomly partitioning the data into 10 equally sized partitions ensuring even contribution from each screen to each partition. We trained the model 10 times on 90% of the data and predicted the held-out 10%. For each data point, we assigned the known guide RNA efficacy quartile based on the  $\log_2$ FC enrichment and compared it the predicted efficacy quartiles in the held-out data. For the RF<sub>GFP</sub> model we found that the model could readily separate poor performing from effective guide RNAs. Accordingly, 46% of the guides present in the highest efficacy-quartile are predicted to reside in the best performing quartile. Conversely, 64% of guides present in the lowest efficacy-quartile are predicted to reside in the poorest performing quartile (Supplementary Fig. 7e). We also assessed the predicted guide score by calculating the median predicted guide score for the top and bottom ranked crRNAs in the 10% held-out data based on the known  $\log_2$ FC-rank for all 10 cross-validation folds (top/bottom  $n = 2, 4, 8, 16, 32, 64, 128$  or 256 guide RNAs). To compute the null distribution, we calculated the median predicted guides scores of randomly selected guide RNAs across 1,000 samplings for each  $n$ . For the leave-one-out cross-validation we trained on all data

from three tiling screens and performed Spearman rank correlation of the predicted the guide efficiency of the held-out fourth screen to the observed  $\log_2FC$  enrichments.

To make the guide score more interpretable, we standardized the guide score to a [0, 1] interval preserving the distribution between 5<sup>th</sup> and 95<sup>th</sup> percentile. Normalized values exceeding the [0, 1] interval were set to 0 or 1, respectively. The final RF<sub>GFP</sub> model was trained on all data points for perfect match guides using the six selected features with 1500 regression trees. The model explains 36.9% of the observed variance with a mean of squared residuals of 0.139. The table below shows the feature contribution for the RF<sub>GFP</sub> model.

Feature	%IncMSE	IncNodePurify
crRNA MFE	57.989	22.617
Local A probability	30.542	7.529
Local C probability	46.255	13.683
Local G probability	38.557	9.256
Local U probability	29.953	6.555
upstream U probability	31.559	8.629

Similarly, final RF<sub>combined</sub> was trained on 2,918 data points using 35 selected features. Tuning the number of trees (ntree) and number of splitting variables per node (mtry) led to insignificant performance improvements compared to default settings. The model (mtry = 12, ntree = 2000) explains 47.16% of the observed variance, a mean of squared residuals of 0.168, and the feature contribution as indicated below ranked by importance:

Feature	%IncMSE	IncNodePurify
Local G <sub>max</sub> probability	86.6814	47.9987
crRNA MFE	72.5480	69.6068
Unpaired probability	72.1078	57.5635
Local C <sub>min</sub> probability	56.9153	37.2137
hybridization MFE nt 15–23	54.5793	94.9857
Guide G nt probability	47.7324	29.2452
Guide CA di-nt probability	47.6241	21.4664
hybridization MFE nt 3–15	47.3083	46.7462
Guide CG di-nt probability	44.0711	11.7069
Guide AU di-nt probability	42.1128	15.6205
Guide CU di-nt probability	40.6967	16.9172
Guide A nt probability	39.7297	25.0736
Local GC <sub>max</sub> probability	39.7236	59.0805
Local AU <sub>min</sub> probability	38.8318	56.3599
Local U <sub>max</sub> probability	38.7535	24.4657
Guide GG di-nt probability	38.5896	22.2621
Local G <sub>min</sub> probability	36.5244	19.6159

Feature	%IncMSE	IncNodePurity
Guide UC di-nt probability	36.3043	15.2075
Guide AC di-nt probability	36.2901	14.2361
Guide C nt probability	36.0901	16.8637
Local $U_{\min}$ probability	35.6886	16.2571
Guide AG di-nt probability	34.9534	14.6260
Local $A_{\min}$ probability	34.6632	17.4928
Local $A_{\max}$ probability	33.6557	16.1888
Guide AA di-nt probability	33.2106	13.3830
Guide GC di-nt probability	33.0670	13.3209
Guide CC di-nt probability	32.2834	12.9797
Guide UG di-nt probability	31.9188	13.1659
Guide GU di-nt probability	31.8413	12.8934
Guide GA di-nt probability	31.3837	13.1399
Guide UA di-nt probability	30.6930	12.5052
Local $C_{\max}$ probability	29.8186	14.2189
Local $AU_{\max}$ probability	22.4842	14.2585
Predicted correctly folded DR	8.2669	4.0679
Predicted G-quadruplex	-3.0274	0.0219

### RfxCas13d guide scoring

We created a user-friendly R script that readily predicts *RfxCas13d* on-target guide scores. The only user-provided argument is a single-entry FASTA file input of minimally 30nt that represents the target sequence, such as a transcript isoform sequence. The software first generates all possible 23mer guide RNAs and collects all required features and predicts guide RNA efficacies. The only filter applied removes guide RNAs with homopolymers of 5 or more Ts and 6 or more Vs ( $V = A, C, G$ ). Such guide RNAs may trigger early transcript termination for PolIII transcription or cause difficulties during oligo synthesis. The software returns a FASTA file with guide RNA sequences ranked by the predicted standardized guide score. In addition, a csv file is created following providing additional information. Optionally, the script can be used to plot the guide score distribution along the provided target sequence for visualization. The software is available here: <https://gitlab.com/sanjanalab/cas13>.

We used this software to predict guide scores for all transcripts (including all biotypes: protein\_coding, nonsense\_mediated\_decay, non\_stop\_decay, IG\*\_gene, TR\*\_gene, polymorphic\_pseudogene) of protein coding genes annotated in GENCODE v19 (GRCh37) ( $n = 94,873$  transcripts) and provide the top 10 ranked 5'UTR, coding sequence and 3'UTR annotating guide RNA sequences (Supplementary Data 3). We have made all guide score predictions available online (<https://cas13design.nygenome.org>).

## RfxCas13d guide scoring validation

To validate our that our initial RF<sub>GFP</sub> model can readily separate between poorly and well-performing crRNAs, we performed several experiments.

First, we chose two genes that encode for cell surface proteins that allow for quantitative assessment of their expression levels by FACS. For each gene we predicted crRNAs for the highest expressed transcript isoform in HEK293FT cells (CD46: ENST00000367042.1, CD71 [TFRC]: ENST00000360110.4). For each gene, we selected 3 guides present in the low scoring quartiles (Q1 and Q2) and 3 guides in the high scoring quartiles (Q3 and Q4). We selected the guides to be non-overlapping and to reside in 3 different regions of the target transcript.

Then, we performed two essentiality screens with a dropout growth phenotype readout in HEK293FT and A375 cells, respectively. We designed two crRNA libraries targeting essential and control genes with a number of predicted low-scoring and high-scoring guide RNAs as described above (see Screen library design and pooled oligo cloning). For the HEK293FT cell screen, we compared the guide depletion of four groups of 30 guides (Essential gene targeted by high-scoring guide or by low-scoring guide, and control genes targeted by high-scoring guide or by low-scoring guide). We expected the greatest depletion for the 30 high-scoring guide RNAs targeting essential genes. Similarly, we compared the relative guide depletion of the same four groups of guide RNAs in the A375 screen, with the expectation that the 20 high-scoring guides per essential gene would be the most depleted.

For gene ranking based on guide depletion, we used robust rank aggregation (RRA)<sup>39</sup> to assign a *p*-value based on the consistency of log<sub>2</sub>FC-based rank-consistency of the most depleted *N* guide RNAs per gene (with *n* in {1, 5, 20}) across the two A375 screen replicates. The -log<sub>10</sub> transformed *p*-values were then compared to other growth screens (RNAi and Cas9) using Spearman rank correlation. Specifically, we compared the RRA-derived log<sub>10</sub> *p*-value to the log<sub>2</sub>FC from an RNAi-based DEMETER2 v5 repository<sup>28</sup> and the merged STARS scores from a Cas9-based approach<sup>29</sup>. For the correlation we only used genes with value present in all scores (All essential genes: *n* = 35; Control genes: *n* = 15).

Furthermore, we used the log<sub>2</sub>FC guide depletion values to compare the predictive value of the RF<sub>GFP</sub> and RF<sub>combined</sub> models. Specifically, for both essentiality screens we used 10 essential genes (all in HEK293FT and the 10 most depleted in A375 cells) and correlated the predicted guide scores from both models to the observed log<sub>2</sub>FC guide depletion scores (normalized to 0–100% per gene) of all detected guide RNAs (HEK293FT: *n* = 60 with 6 guides per gene; A375: *n* = 398 with up to 40 guides per gene). We made the same comparison on a per-gene level using all 40 guide RNAs per gene in the A375 screen.

## Data representation

In all boxplots, boxes indicate the median and interquartile ranges, with whiskers indicating either 1.5 times the interquartile range, or the most extreme data point outside the 1.5-fold interquartile. All transfection experiments show the mean of three replicate experiments with individual replicates plotted as points.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgements

We thank the entire Sanjana laboratory for support and advice. We thank R. Satija for critical feedback. We also thank D. Knowles for discussion about model building and M. Zaran for assistance with the web-tool server. N.E.S. is supported by NYU and NYGC startup funds, NIH/NHGRI (R00HG008171, DP2HG010099), NIH/NCI (R01CA218668), DARPA (D18AP00053), the Sidney Kimmel Foundation, the Melanoma Research Alliance, and the Brain and Behavior Foundation. A.M.-M. is supported by CONACyT-Mexico Fellowship (#412653).

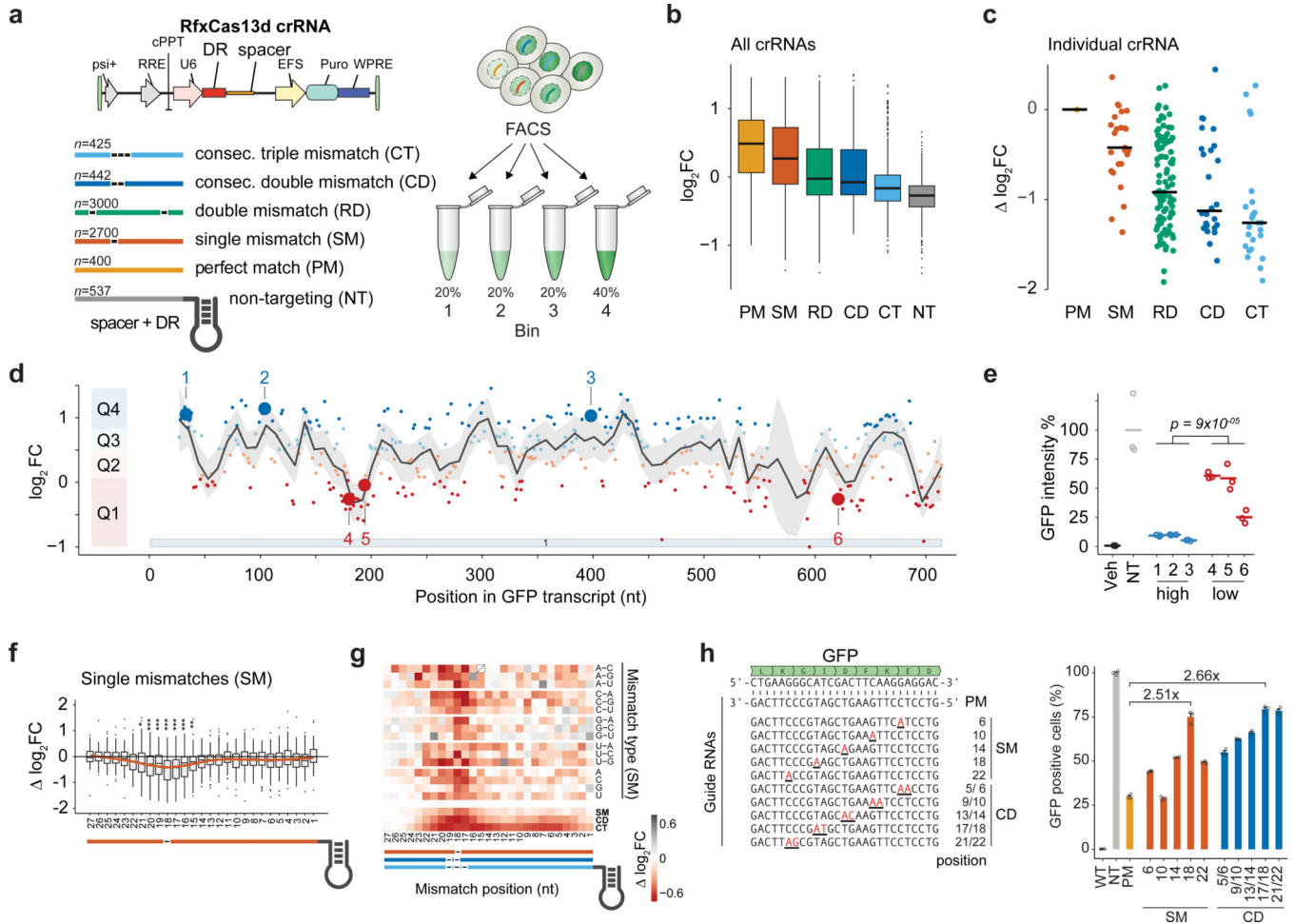
## References

1. Abudayyeh OO et al. C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science* 353, (2016).
2. Abudayyeh OO et al. RNA targeting with CRISPR-Cas13. *Nature* 550, 280–284 (2017). [PubMed: 28976959]
3. East-Seletsky A. et al. Two distinct RNase activities of CRISPR-C2c2 enable guide-RNA processing and RNA detection. *Nature* 538, 270–273 (2016). [PubMed: 27669025]
4. Smargon AA 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). [PubMed: 28065598]
5. Konermann S. et al. Transcriptome Engineering with RNA-Targeting Article Transcriptome Engineering with RNA-Targeting. *Cell* 173, 1–12 (2018). [PubMed: 29570990]
6. Yan WX 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). [PubMed: 29551514]
7. Freije CA et al. Programmable Inhibition and Detection of RNA Viruses Using Cas13. *Mol. Cell* 76, 1–12 (2019). [PubMed: 31585100]
8. Poosala P, Lindley SR, Anderson KM & Anderson DM Targeting Toxic Nuclear RNA Foci with CRISPR-Cas13 to Treat Myotonic Dystrophy. *bioRxiv* 1–29 (2019). doi:doi.org/10.1101/716514
9. Mahas A, Aman R. & Mahfouz M. CRISPR-Cas13d mediates robust RNA virus interference in plants. *Genome Biol.* 20, 1–16 (2019). [PubMed: 30606230]
10. Kushawah G. et al. CRISPR-Cas13d induces efficient mRNA knock-down in animal embryos. *bioRxiv* 2020.01.13.904763 (2020). doi:10.1101/2020.01.13.904763
11. Gootenberg JS et al. Nucleic acid detection with CRISPR-Cas13a/C2c2. *Science* 356, 438–442 (2017). [PubMed: 28408723]
12. Gootenberg JS et al. Multiplexed and portable nucleic acid detection platform with Cas13, Cas12a, and Csm6. *Science* 360, 439–444 (2018). [PubMed: 29449508]
13. Cox DBT et al. RNA editing with CRISPR-Cas13. *Science* 358, 1019–1027 (2017). [PubMed: 29070703]
14. Li J. et al. Targeted mRNA demethylation using an engineered dCas13b-ALKBH5 fusion protein. *bioRxiv* (2019). doi:10.1101/614859
15. Wang H. et al. CRISPR-mediated live imaging of genome editing and transcription. *Science* 365, 2–6 (2019).
16. Yang L-Z et al. Dynamic Imaging of RNA in Living Cells by CRISPR-Technology. *Mol. Cell* 76, 1–17 (2019). [PubMed: 31585100]
17. Jillette N. & Cheng AW CRISPR Artificial Splicing Factors. *bioRxiv* (2018). doi:10.1101/431064
18. Anderson KM, Poosala P, Lindley SR & Anderson DM Targeted Cleavage and Polyadenylation of RNA by CRISPR-Cas13. *bioRxiv* (2019). doi:10.7143/jhep.46.175
19. Meeske AJ, Nakandakari-Higa S. & Marraffini LA Cas13-induced cellular dormancy prevents the rise of CRISPR-resistant bacteriophage. *Nature* 570, 241–245 (2019). [PubMed: 31142834]

20. Meeske AJ & Marraffini LA RNA Guide Complementarity Prevents Self-Targeting in Type VI CRISPR Systems. *Mol. Cell* 71, 791–801 (2018). [PubMed: 30122537]
21. Liu L. et al. The Molecular Architecture for RNA-Guided RNA Cleavage by Cas13a. *Cell* 170, 714–720 (2017). [PubMed: 28757251]
22. Tambe A, East-seletsky A, Knott GJ, Connell MRO & Doudna JA RNA Binding and HEPN-Nuclease Activation Are Decoupled in CRISPR-Cas13a. *Cell Rep.* 24, 1025–1036 (2018). [PubMed: 30044970]
23. Zhang C. et al. Structural Basis for the RNA-Guided Ribonuclease Activity of CRISPR-Cas13d. *Cell* 175, 212–223 (2018). [PubMed: 30241607]
24. Zhang B. et al. Two HEPN domains dictate CRISPR RNA maturation and target cleavage in Cas13d. *Nat. Commun* 10, (2019).
25. Konermann S. et al. Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature* 517, 583–588 (2015). [PubMed: 25494202]
26. Replogle JM et al. Direct capture of CRISPR guides enables scalable, multiplexed, and multi-omic Perturb-seq. *bioRxiv* (2018). doi:10.1101/503367
27. Doench JG et al. Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat. Biotechnol* 32, 1262–1267 (2014). [PubMed: 25184501]
28. McFarland JM et al. Improved estimation of cancer dependencies from large-scale RNAi screens using model-based normalization and data integration. *Nat. Commun* 9, 1–13 (2018). [PubMed: 29317637]
29. Doench JG et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol* 34, 184–191 (2016). [PubMed: 26780180]
30. Briesse M. et al. A systems view of spliceosomal assembly and branchpoints with iCLIP. *Nat. Struct. Mol. Biol* 26, 930–940 (2019). [PubMed: 31570875]
31. Saulière J. et al. CLIP-seq of eIF4AIII reveals transcriptome-wide mapping of the human exon junction complex. *Nat. Struct. Mol. Biol* 19, 1124–1131 (2012). [PubMed: 23085716]
32. Hauer C. et al. Exon Junction Complexes Show a Distributional Bias toward Alternatively Spliced mRNAs and against mRNAs Coding for Ribosomal Proteins. *Cell Rep.* 16, 1588–1603 (2016). [PubMed: 27475226]

## Method References

33. Sanjana NE, Shalem O. & Zhang F. Improved vectors and genome-wide libraries for CRISPR screening. *Nat. Methods* 11, 783–784 (2014). [PubMed: 25075903]
34. Lorenz R. et al. ViennaRNA Package 2.0. *Algorithms Mol. Biol* 6, 26 (2011). [PubMed: 22115189]
35. Shalem O. et al. Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells. *Science* (80-. ). 343, 84–88 (2014).
36. Langmead B, Trapnell C, Pop M. & Salzberg SL Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10, (2009).
37. Love MI, Huber W. & Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 550 (2014). [PubMed: 25516281]
38. Leek JT, Johnson WE, Parker HS, Jaffe AE & Storey JD The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 28, 882–883 (2012). [PubMed: 22257669]
39. Kolde R, Laur S, Adler P. & Vilo J. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics* 28, 573–580 (2012). [PubMed: 22247279]
40. Agarwal V, Subtelny AO, Thiru P, Ulitsky I. & Bartel DP Predicting microRNA targeting efficacy in *Drosophila*. *Genome Biol.* 19, 1–23 (2018). [PubMed: 29301551]
41. Krueger J. & Rehmsmeier M. RNAhybrid : microRNA target prediction easy, fast and flexible. *Nucleic Acids Res.* 34, 451–454 (2006). [PubMed: 16421272]

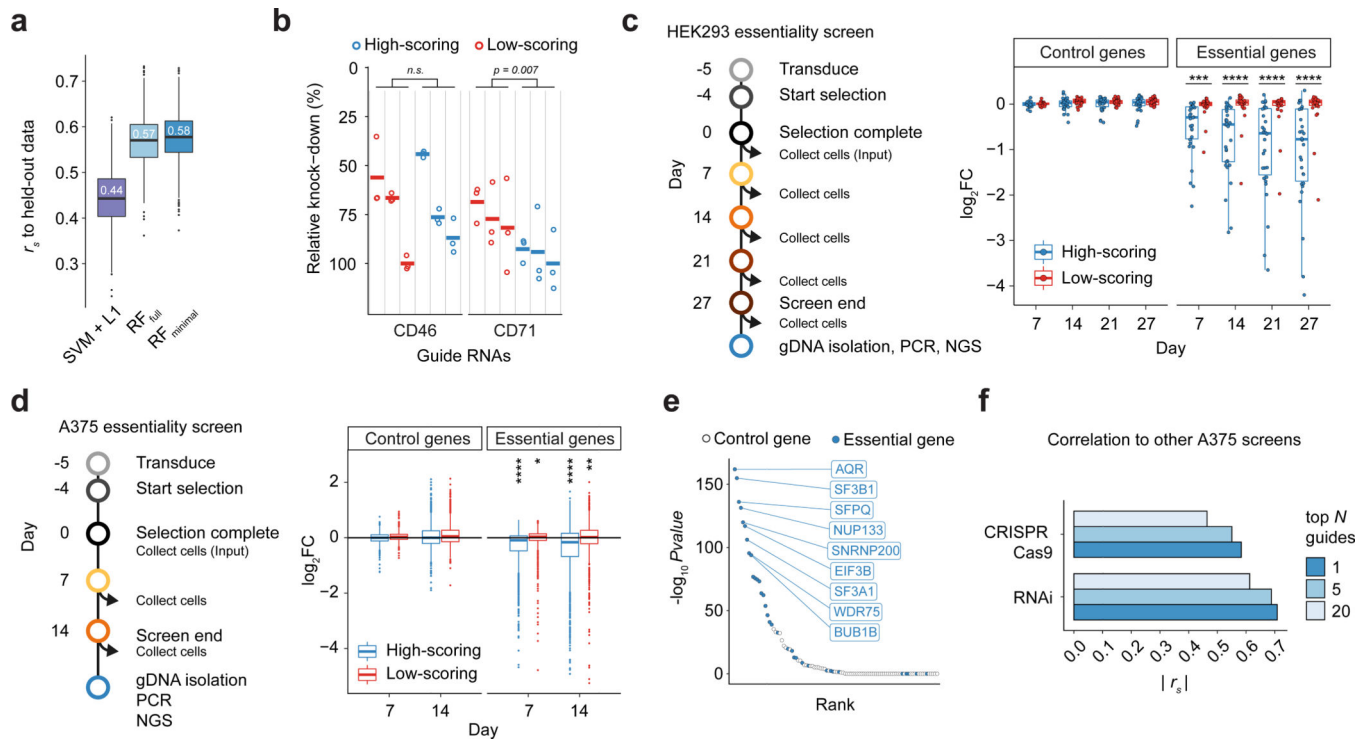


**Figure 1. CRISPR Type VI-D RfxCas13d GFP knock-down pooled tiling screen.**

(a) The GFP-targeting guide RNA (gRNA) library ( $n = 7,500$ ) was lentivirally transduced into TetO-RfxCas13d and GFPd2PEST HEK293 cells in  $n = 3$  transduction replicates. After selection, cells are sorted by GFP intensities into 4 bins. (b, c)  $\log_2$  fold-change ( $\log_2FC$ ) enrichment scores of gRNAs comparing gRNA counts of the lowest fluorescence (Bin 1) to the input (unsorted) cell population. Scores are demarcated by gRNA type as given by the list in a. (b) All gRNAs. (c) A single perfect match (PM) gRNA and corresponding derivative gRNAs with mismatches. Guide RNA  $\log_2FC$  enrichments are calculated relative to the perfect match reference gRNA ( $\log_2FC$ ). Black lines denote medians. (d) Distribution of perfect match gRNAs along the GFP mRNA and their  $\log_2FC$  enrichment ( $n = 399$ ). Guide RNAs are separated into targeting efficacy quartiles Q1-Q4 with Q4 containing guides with the best knock-down efficacy. Line indicates LOESS fit with 95% confidence interval shading. (e) GFP knock-down validation for 6 guides (3 with high efficacy and 3 with low efficacy) highlighted in d. (Veh = vehicle transfection, NT = non-targeting gRNAs; lines indicate mean of  $n = 3$  biological replicates). (f) Relative targeting efficacy ( $\log_2FC$ ) of gRNAs with single nucleotide mismatches (SM) at the indicated position relative to their cognate perfect match gRNAs ( $n = 100$ ). Significance: \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$  from a two-tailed t-test. (g) (top) Change in targeting efficacy by

gRNA nucleotide identity or mismatch type. (bottom) Change in targeting efficacy for SM, CD, or consecutive triple mismatch (CT) by position. **(h)** Validation of RfxCas13d seed region. (left) Individual perfect match and mismatch gRNAs relative to GFP target mRNA. (right) Percent GFP negative cells after co-transfection of specific GFP-targeting gRNAs normalized to the non-targeting control. (Veh = vehicle transfection, NT = non-targeting gRNAs, PM = perfect match guide RNA; bars indicate mean of n = 3 biological replicates). Boxes in b and f indicate the median and interquartile ranges, with whiskers indicating 1.5 times the interquartile range, or the most extreme data point outside the 1.5-fold interquartile.

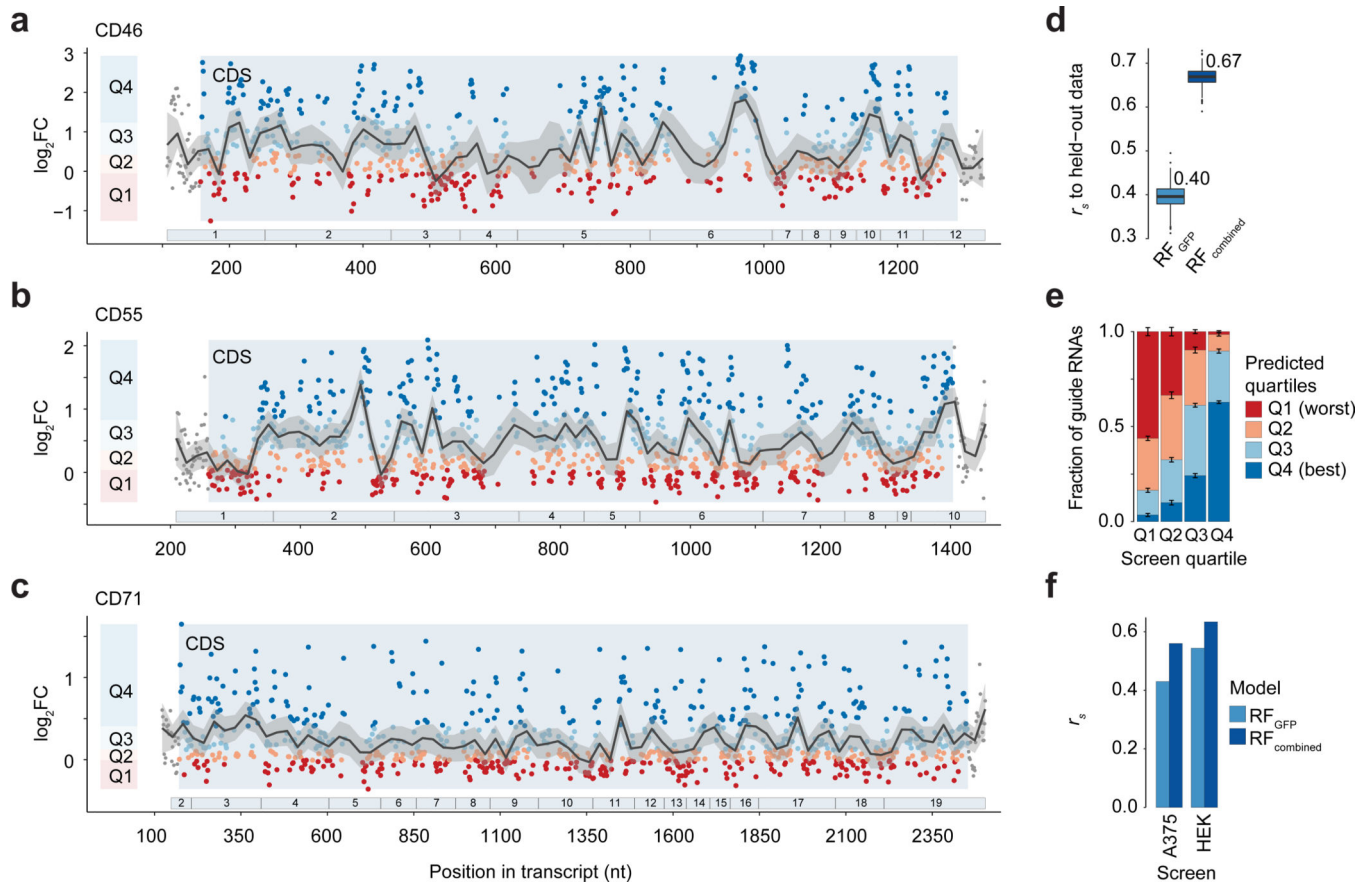




**Figure 2. Rfx-Cas13d on-target guide RNA prediction model.**

(a) Spearman correlation  $r_s$  of predictions from a Random Forest (RF) regression model (either with all features or a minimal set of the most predictive features) and a support vector machine with L1 regression to held-out screen data ( $n = 1,000$  bootstraps). (b) Validation of on-target model testing 3 high-scoring and 3 low-scoring guide RNAs (gRNAs) via targeting of cell-surface proteins and antibody labeling to measure target knock-down by flow cytometry. Relative knock-down indicates the percent reduction (relative to non-targeting gRNAs) in the mean fluorescence intensity (lines indicate mean of  $n = 3$  biological replicates). (c) Validation of on-target model assaying 3 high-scoring and 3 low-scoring gRNAs per gene in a gene essentiality screen in HEK293 cells with growth dropout phenotype testing 10 essential genes and 10 control genes. Each point represents one gRNA as a mean of three replicate experiments. The  $y$ -axis depicts the  $\log_2$  fold-change (FC) of the gRNA at the indicated time point relative to the Day 0 sample. One-sided KS-test comparing high-scoring and low-scoring guides, \*\*\*  $p = 2 \times 10^{-5}$ , \*\*\*\*  $p = 2 \times 10^{-6}$ . (d) A375 essentiality screen with growth dropout phenotype assaying 20 high-scoring and 20 low-scoring gRNAs per gene ( $n = 35$  essential genes and  $n = 65$  control genes). One-sided KS-test comparing high-scoring ( $n = 698$ ) or low-scoring ( $n = 700$ ) guides to the distribution of non-targeting gRNAs ( $n = 677$ ). \*  $p = 0.043$ , \*\*  $p = 0.0095$ , \*\*\*\*  $p < 1 \times 10^{-44}$ . (e) Gene ranking for essentiality based on the robust rank aggregation (RRA)  $p$ -value across replicates for all 20 high-scoring gRNAs and all  $n = 100$  genes tested. Blue dots denote essential genes from a prior RNAi screen<sup>28</sup>. (f) Spearman rank correlation of Cas13d gene depletion (as in e) with prior CRISPR-Cas9 and RNAi screens in A375 cells. Analysis includes genes represented in all libraries ( $n = 35$  essential genes and  $n = 15$  control genes). (RNAi screen: A375 DEMETER2 v5 score<sup>28</sup>, Cas9 screen: A375 STARS score<sup>29</sup>). Boxes

in *a*, *c* and *d* indicate the median and interquartile ranges, with whiskers indicating 1.5 times the interquartile range, or the most extreme data point outside the 1.5-fold interquartile.



**Figure 3. Improvement of *RfxCas13d* on-target guide RNA prediction model with tiling screens over endogenous transcripts.**

(a-c) Distribution of perfect match guide RNAs (gRNAs) along the coding region of CD46 ( $n = 704$  gRNAs), CD55 ( $n = 925$  gRNAs) and CD71 ( $n = 890$  gRNAs) transcripts and their log<sub>2</sub> fold-change (FC) enrichments. Positive FC values indicate better transcript knock-down. Guide RNAs are separated into targeting efficacy quartiles Q1-Q4 per gene with Q4 containing gRNAs with the best knock-down efficacy. Numbered bars below indicate exons. Lines indicate LOESS fit with 95% confidence interval shading. (d) Spearman correlation  $r_s$  of predictions from the RF<sub>minimal</sub> (RF<sub>GFP</sub>) model and the updated RF<sub>combined</sub> regression model to held-out screen data using bootstrapping across all four tiling screens ( $n = 1000$  bootstraps). (e) Comparison of predicted and measured log<sub>2</sub>FC quartiles across the 10-fold model cross-validation. Quartile definition as in a-c. Bars indicate the mean. Error bars denote s.e.m.. (f) Spearman rank correlation between observed gRNA depletion (target knock-down) and the predicted guide score for the indicated Cas13d essentiality screens and indicated on-target models (10 most essential genes; A375:  $n = 398$  gRNAs, HEK:  $n = 60$  gRNAs) (see Fig 2c-d and Supplementary Fig. 8). Boxes in d indicate the median and interquartile ranges, with whiskers indicating 1.5 times the interquartile range, or the most extreme data point outside the 1.5-fold interquartile.