nature biotechnology

LETTERS

https://doi.org/10.1038/s41587-020-0456-9

Check for updates

# Massively parallel Cas13 screens reveal principles for guide RNA design

Hans-Hermann Wessels[1,2,3], Alejandro Méndez-Mancilla[1,2,3], Xinyi Guo[1,2], Mateusz Legut[1,2], Zharko Daniloski[1,2] and Neville E. Sanjana[1,2] ✉

**Type VI CRISPR enzymes are RNA-targeting proteins with nuclease activity that enable specific and robust target gene knockdown without altering the genome. To define rules for the design of Cas13d guide RNAs (gRNAs), we conducted massively parallel screens targeting messenger RNAs (mRNAs) of a green fluorescent protein transgene, and CD46, CD55 and CD71 cell-surface proteins in human cells. In total, we measured the activity of 24,460 gRNAs with and without mismatches relative to the target sequences. Knockdown efficacy is driven by gRNA-specific features and target site context. Single mismatches generally reduce knockdown to a modest degree, but spacer nucleotides 15–21 are largely intolerant of target site mismatches. We developed a computational model to identify optimal gRNAs and confirm their 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 gRNAs for all protein-coding transcripts in the human genome.**

Type VI CRISPR (clustered regularly interspaced short palindromic repeats) enzymes have recently been identified as programmable RNA-guided, RNA-targeting Cas proteins with nuclease activity that allows for target gene knockdown without altering the genome. In addition to target RNA knockdown[1–10], Cas13 proteins have been used for viral RNA detection[7,9,11,12], site-directed RNA editing[13], demethylation of m$^6$A-modified transcripts[14], RNA live imaging[15,16] and modulation of splice site choice, as well as cleavage and polyadenylation site usage[5,17,18]. 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 (gRNA) that mediates target recognition by RNA–RNA hybridization. Although Cas13 enzymes exert some nonspecific collateral nuclease activity on activation[4–6,11,19], they have greatly reduced off-target activity in cultured cells compared with RNA interference (RNAi)[2,5,13]. Previous studies have shown that Cas13 gRNAs have minimal protospacer flanking sequence constraints[1,4,13,20] and that RNA target sites should be accessible for Cas13 binding[1,2,4]. Beyond these basic parameters, we currently lack information about optimal Cas13 crRNA designs for effective target RNA knockdown.

To date, three Cas13 effector proteins (*Pgu*Cas13b, *Psp*Cas13b, *Rfx*Cas13d) have been reported as showing high RNA knockdown efficacy with minimal off-target activity[5,13]. We compared the ability of these Cas13 enzymes to knock down green fluorescent protein (GFP) mRNA when directed to either the cytosol or the nucleus. *Rfx*Cas13d (CasRx) consistently showed the strongest target knockdown, especially when fused to a nuclear localization sequence (NLS) (see 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-nucleotide (nt) reference gRNA, and found that 23-nt to 30-nt gRNAs confer the most pronounced target knockdown (see Supplementary Fig. 1d).

To systematically assess the *Rfx*Cas13d target knockdown 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 (PM) and mismatch gRNAs (Fig. 1a). We performed FACS to gate cells in four bins based on their GFP intensity (see Supplementary Fig. 2a). The gRNA counts showed high concordance between bins across three independent transductions, with clear separation of bin 1, which contained cells with the lowest GFP expression (see Supplementary Fig. 2b–d).

We calculated the $\log_2$(fold change) ($\log_2$(FC)) gRNA enrichment between all bins and the unsorted input gRNA distribution (see Supplementary Data 1). PM gRNAs were enriched in bin 1, whereas increasing numbers of mismatches led to a gradual decrease in gRNA enrichment (Fig. 1b and see Supplementary Fig. 3a–c). This was true for the whole gRNA population as well as for individual PM gRNAs and their corresponding gRNAs with one to three mismatches (Fig. 1b,c and see Supplementary Fig. 3d). As a control, the library also contained 537 nontargeting crRNAs which were effectively depleted from bin 1 (Fig. 1b and see Supplementary Fig. 3a–c). As expected, gRNA abundances in bin 1 were negatively correlated to those in bins 2–4, which contained cells with higher GFP intensities (see Supplementary Fig. 3e,f). Taken together, this suggests that the enrichments of gRNAs in bin 1 accurately reflect target mRNA knockdown.

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

To examine whether Cas13 can tolerate mismatches between the gRNA and the target RNA, we calculated the relative $\log_2$(FC) ($\Delta\log_2$(FC)) for each mismatch gRNA by subtracting the $\log_2$(FC) from the reference (PM) gRNA (Fig. 1f). We found a critical (seed) region for Cas13d knockdown efficacy between gRNA nucleotides 15–21, with its center at nucleotide 18 relative to the gRNA 5′-end. Although seed regions have been shown for Cas13a[1,21,22], one group reported no clear seed region for Cas13d[23], whereas another showed

[1]New York Genome Center, New York, NY, USA. [2]Department of Biology, New York University, New York, NY, USA. [3]These authors contributed equally: Hans-Hermann Wessels, Alejandro Méndez-Mancilla. ✉e-mail: neville@sanjanalab.org

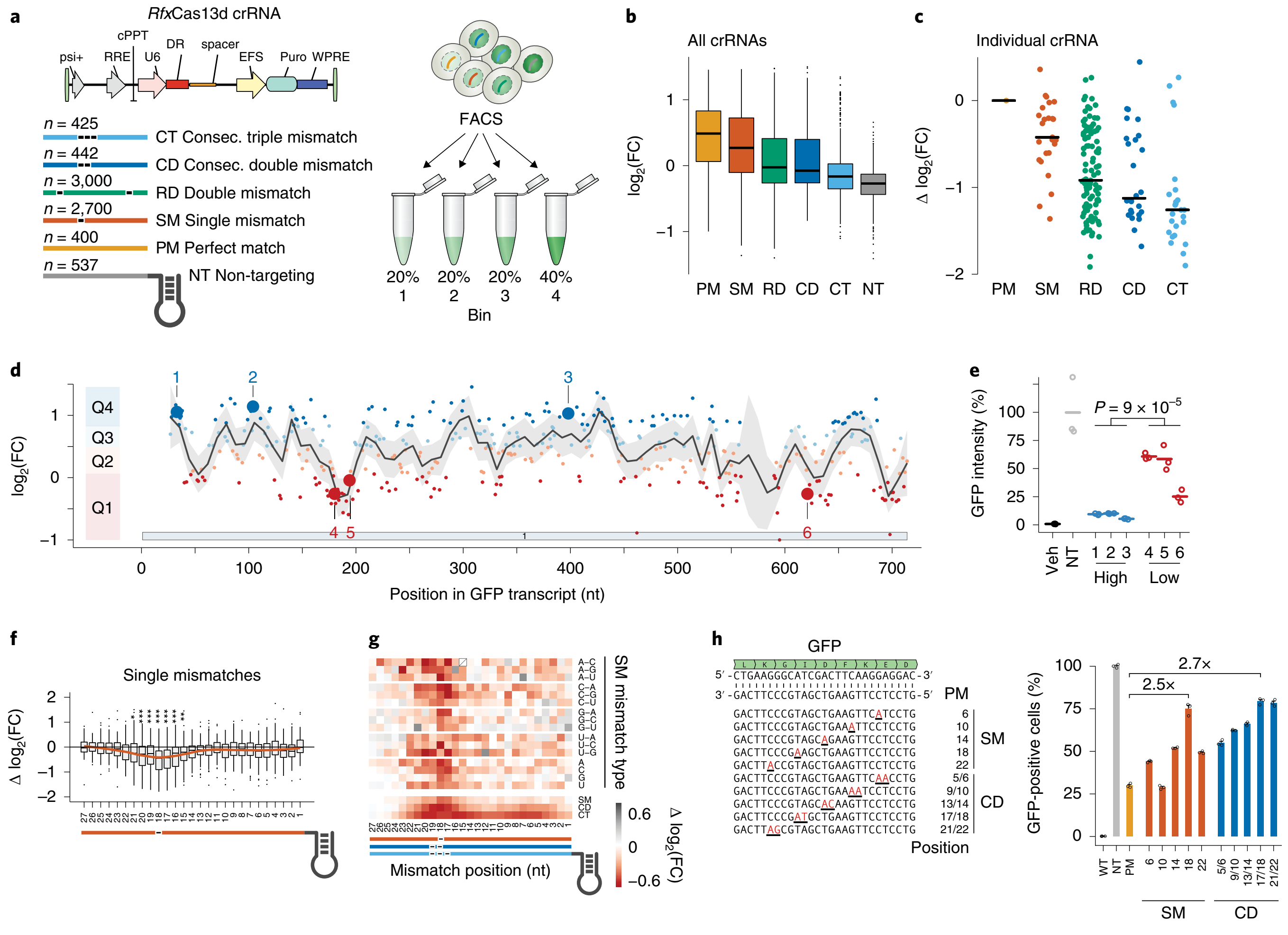

**Fig. 1 | Pooled CRISPR *Rfx*Cas13d GFP knockdown tiling screen.** **a**, The GFP-targeting gRNA library ($n = 7{,}500$) was lentivirally transduced into TetO-*Rfx*Cas13d and GFPd2PEST HEK293 cells in $n = 3$ transduction replicates. After selection, cells are sorted by GFP intensities into four bins. psi+, Psi packaging signal; RRE, Rev response element; cPPT, central polypurine tract; *U6*, human *U6* promoter; DR, *Rfx*Cas13d direct repeat; spacer, gRNA sequence; EFS, elongation factor 1α short promoter; Puro, puromycin selection marker; WPRE, post-transcriptional regulatory element. **b**,**c**, $\log_2$(FC) enrichment scores of gRNAs comparing gRNA counts of the lowest fluorescence (bin 1) with the input (unsorted) cell population. Scores are demarcated by gRNA type, as given by the list in **a**. **b**, All gRNAs. **c**, A single PM gRNA and corresponding derivative gRNAs with mismatches. The gRNA $\log_2$(FC) enrichments are calculated relative to the PM reference gRNA ($\Delta\log_2$(FC)). Black lines denote medians. **d**, Distribution of PM gRNAs along the GFP mRNA and their $\log_2$(FC) enrichment ($n = 399$). The gRNAs are separated into targeting efficacy quartiles Q1–Q4, with Q4 containing guides with the best knockdown efficacy. The line indicates LOESS (locally estimated scatterplot smoothing) fit with 95% confidence interval shading. **e**, Percentage GFP knockdown for six guide RNAs (three with high efficacy and three with low efficacy) highlighted in **d** (lines indicate mean of $n = 3$ biological replicates). Veh, vehicle transfection. **f**, Relative targeting efficacy ($\Delta\log_2$(FC)) of gRNAs with SMs at the indicated position relative to their cognate PM gRNAs ($n = 100$; $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$ from a two-tailed, Student's *t*-test). **g**, Top: change in targeting efficacy by gRNA nucleotide identity or mismatch type for SM gRNAs. Bottom: change in targeting efficacy for SMs, CDs or CTs by position. **h**, Validation of *Rfx*Cas13d seed region. Left: individual PM and mismatch gRNAs relative to GFP target mRNA. Right: percentage of GFP-positive cells after co-transfection of specific GFP-targeting gRNAs normalized to the nontargeting control (mean of $n = 3$ biological replicates). WT, wild type. Boxes in **b** and **f** indicate the median and IQRs, with whiskers indicating 1.5× the IQR or the most extreme data point outside the 1.5-fold IQR.

position-dependent mismatch sensitivity for Cas13d in a cell-free assay[24]. Within the seed region, single mismatches led to diminished gRNA enrichment, whereas mismatches outside the seed region were tolerated better (Fig. 1f). The critical region was present regardless of the mismatch identity (Fig. 1g). Similarly, consecutive double (CD) and triple (CT) mismatches indicated the presence of the critical region (Fig. 1g and see Supplementary Fig. 4a). For randomly distributed double mismatches, the largest change in enrichment was observed in cases where both mismatches were in the seed region (see Supplementary Fig. 4b). Increasing the number of mismatches to three mismatches largely abrogated target knockdowns (see Supplementary Fig. 4a). For this reason, the critical region may have been masked in previous studies on *Es*Cas13d which tested four consecutive mismatches[23].

Given the heterogeneity in enrichment for gRNAs with mismatches in the seed region, we sought to assess the effect of surrounding nucleotide context (see 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 (see 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). Whereas a PM gRNA decreased the percentage of GFP-positive cells to ~29%, a single mismatch at gRNA

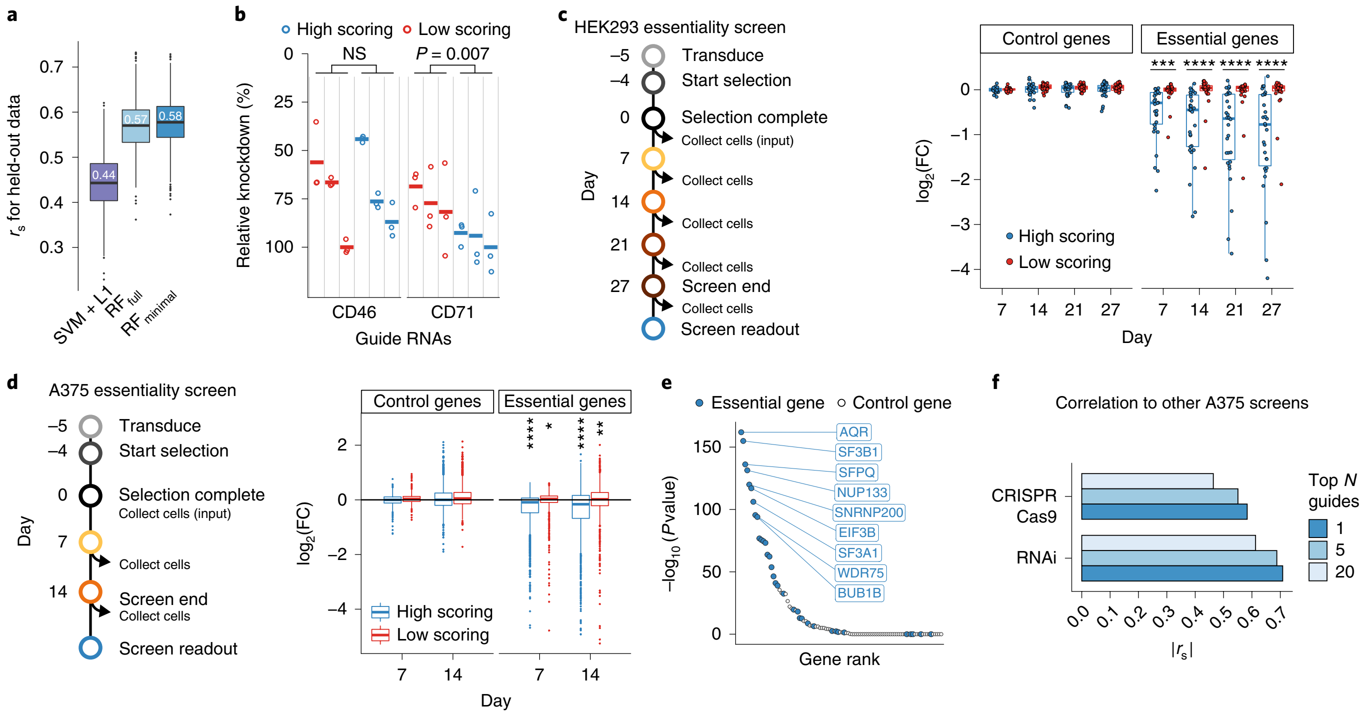

**Fig. 2 | *Rfx*Cas13d on-target gRNA prediction model. a**, Spearman rank correlation $r_s$ of predictions from an RF regression model (either with all features or a minimal set of the most predictive features) and from an SVM with L1 regression to held-out screen data ($n$ = 1,000 bootstraps). **b**, Validation of on-target model testing three high-scoring and three low-scoring gRNAs via targeting of cell-surface proteins and antibody labeling to measure target knockdown by flow cytometry. Relative knockdown indicates the percentage reduction (relative to nontargeting gRNAs) in the mean fluorescence intensity (lines indicate mean of $n$ = 3 biological replicates). NS, not significant. **c**, Validation of on-target model assaying three high-scoring and three low-scoring gRNAs per gene in a gene essentiality screen in HEK293 cells with ten essential genes and ten control genes. Each point represents one gRNA as a mean of three replicate experiments. The $y$ axis depicts the $\log_2$(FC) of the gRNA at the indicated time point relative to the day 0 sample (one-sided Kolmogorov–Smirnov 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 Kolmogorov–Smirnov test comparing high-scoring ($n$ = 698) and low-scoring ($n$ = 700) guides with the distribution of nontargeting gRNAs ($n$ = 677; *$P$ = 0.043, **$P$ = 0.0095, ****$P < 1 \times 10^{-44}$). **e**, Gene ranking for essentiality based on the RRA $P$ value across replicates using all 20 high-scoring gRNAs for the A375 screen in **d**. Blue dots denote essential genes from a prior RNAi screen[28]. **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 version 5 score[28]; Cas9 screen: A375 STARS score[29]). Boxes in **a**, **c** and **d** indicate the median and IQRs, with whiskers indicating 1.5× the IQR or the most extreme data point outside the 1.5-fold IQR.

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 knockdown efficacy for PM 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 PM crRNAs showed a positive correlation between the MFE and gRNA efficacy (see Supplementary Fig. 6a). In particular, 'G'-dependent structures, such as predicted G-quadruplexes, showed diminished target knockdown. Given that the crRNA folding is critical for effective target knockdown, we sought to further stabilize and improve the DR through repair of a predicted bulge in the DR, by varying the length of the stem loop or by disrupting bases in the proximal DR stem (see Supplementary Fig. 6b). Analysis of the crystal structure of *Es*Cas13d and *Ur*Cas13d, 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[23,24], similar to that previously found to enhance Cas9 activity[25,26]. We observed that any change in stem length abrogated target knockdown completely (see Supplementary Fig. 6c). Also, repair of the bulged nucleotide within the stem decreased target knockdown. However, disruption of the first base-pair within the proximal stem further increased Cas13d targeting efficacy, leading to a novel *Rfx*Cas13d DR with improved knockdown. We tested the modified DR on six additional gRNAs targeting GFP and found that the modified DR improved target knockdown, especially for gRNAs with low knockdown efficacy (see Supplementary Fig. 6d).

We defined 15 crRNA and target RNA features based on their correlation with observed gRNA enrichment (see Supplementary Table 1 and Supplementary Note 1). With these features, we sought to derive a generalizable 'on-target' model to predict Cas13d target knockdown. 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 (see Supplementary Fig. 7a), weighting the crRNA-folding energy, the local target C context and the upstream target U context as the most important features (see Supplementary Fig. 7b). Other learning approaches frequently chose similar features, suggesting that these features are the main drivers of Cas13d GFP knockdown (see Supplementary Fig. 7c). To identify the key predictor of gRNA efficacy, we iteratively reduced the number of features, monitoring the model performance and deriving a minimal model that explained about 37% of the variance ($r^2$) with Spearman's correlation ($r_s$) of ~0.58 for the held-out data (Fig. 2a and see Supplementary Fig. 7d–f). In comparison, a support vector machine (SVM) regression

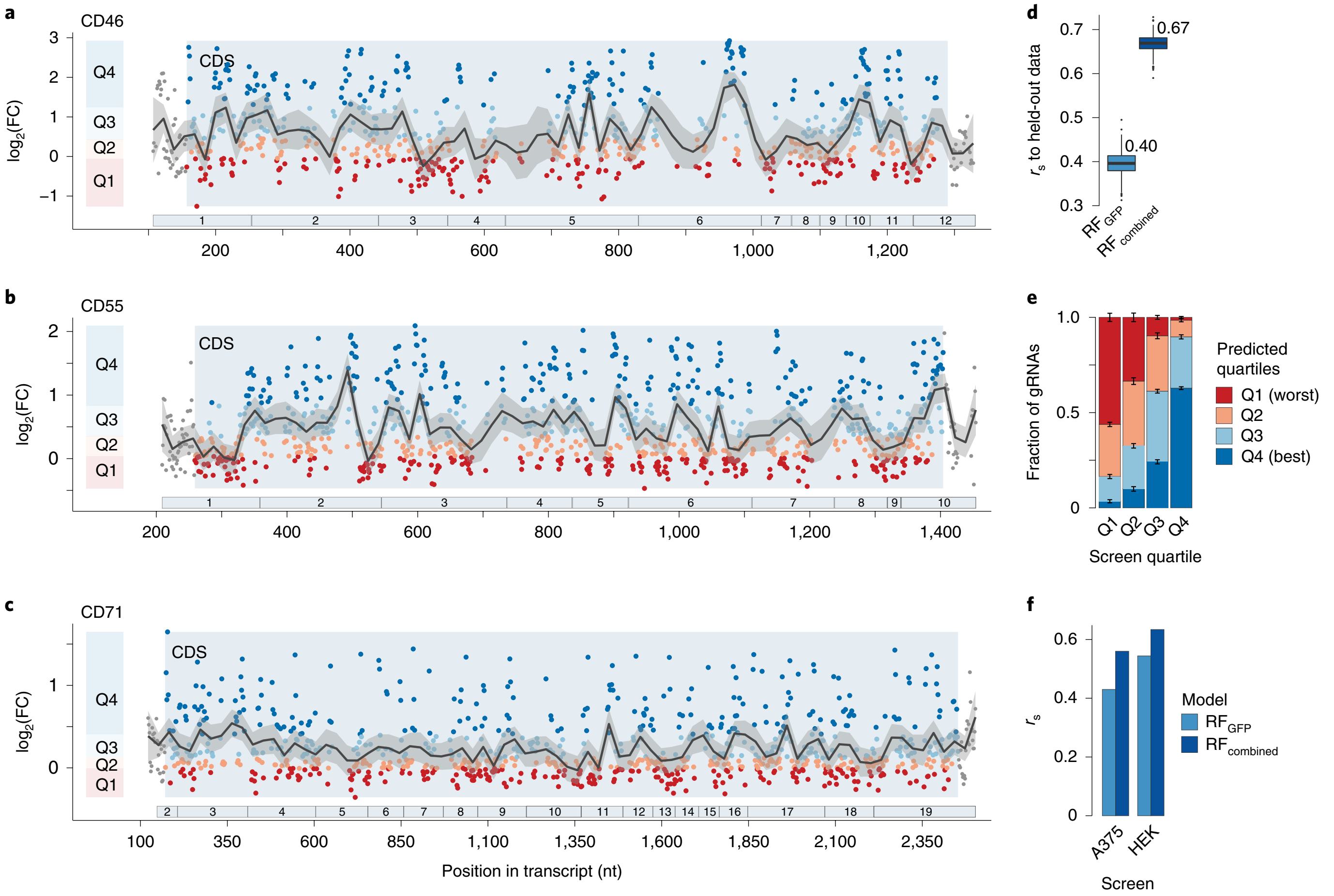

**Fig. 3 | Improvement of *Rfx*Cas13d on-target gRNA prediction model with tiling screens over endogenous transcripts. a–c**, Distribution of PM gRNAs along the coding region of CD46 (**a**, $n=704$ gRNAs), CD55 (**b**, $n=925$ gRNAs) and CD71 (**c**, $n=890$ gRNAs) transcripts and their $\log_2$(FC) enrichments. Positive FC values indicate better transcript knockdown. The gRNAs are separated into targeting efficacy quartiles Q1–Q4 per gene with Q4 containing gRNAs with the best knockdown efficacy; the numbered bars below indicate exons. Lines indicate LOESS fit with 95% confidence interval shading. **d**, Spearman correlation $r_s$ of predictions from the $RF_{minimal}$ ($RF_{GFP}$) model and the updated $RF_{combined}$ regression model for the held-out screen data using bootstrapping across all four tiling screens ($n=1{,}000$ bootstraps). **e**, Comparison of predicted and measured $\log_2$(FC) quartiles across the tenfold model cross-validation. Quartile definition as in **a**–**c**. Bars indicate the mean and error bars denote the s.e.m. **f**, Spearman rank correlation between observed gRNA depletion (target knockdown) and the predicted guide score for the indicated Cas13d essentiality screens and indicated on-target models (ten 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 IQRs, with whiskers indicating 1.5× the IQR or the most extreme data point outside the 1.5-fold IQR.

model with a similar structure to a Cas9 gRNA prediction algorithm[27] performed worse when applied to these data ($r^2=0.21$, $r_s=0.44$) (Fig. 2a).

To show that our model is generalizable, we designed gRNAs to target the endogenous transcripts of *CD46* and *CD71*, which encode cell-surface proteins, and measured the gRNA knockdown efficacy by flow cytometry. For each gene, we chose three gRNAs predicted to have high knockdown efficacy (top two quartiles) and three gRNAs predicted to have low knockdown efficacy (bottom two quartiles). On an individual gRNA level, we found that most 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 knockdown between all three high-scoring gRNAs and all three low-scoring gRNAs, we found a notable improvement for CD71, whereas 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 ten essential and ten control genes with both three high-scoring and three 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[28] (see Supplementary Fig. 8a). Most high-scoring gRNAs targeting essential genes were progressively depleted over time, whereas low-scoring gRNAs showed largely no depletion (Fig. 2c, and see Supplementary Fig. 8b).

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

We calculated a gene depletion score 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[28] 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[29] (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 could largely separate gRNAs with low knockdown 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 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 and see Supplementary Fig. 9a–c). In addition to PM gRNAs, we added several additional gRNA classes, including gRNAs to target noncoding elements (see Supplementary Fig. 9a). For each screen, PM gRNAs showed the strongest gRNA enrichment relative to the unsorted input samples, whereas two different negative controls, reverse complement gRNAs and nontargeting gRNAs, were depleted (see 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 nt to 36 nt. Starting from 23-nt length, gRNAs exerted full knockdown efficacy, whereas longer gRNA 3′ ends did not have any deleterious effects (see Supplementary Fig. 9e).

PM gRNAs targeting coding sequence (CDS) were more enriched compared with gRNAs targeting untranslated regions (UTRs) or introns (see Supplementary Fig. 9f). UTR-targeting gRNAs may show lower enrichments because each target gene may be represented by multiple transcript isoforms with alternative UTR usage. Hence, gRNAs targeting coding regions have a higher likelihood of finding the cognate target site whereas, for example, 3′-UTR-targeting gRNAs find their target site only in a fraction of the expressed transcript 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 may enable the transcript to evade Cas13 targeting. For this reason, gRNA knockdown 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 nt upstream of the 3′-splice site (see Supplementary Fig. 9g). These sites are typically bound by the spliceosome[30], 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 would affect knockdown of the mature transcript in the same way. The exon–junction complex typically binds ~20–24 nt 5′ of the exon–exon junction during splicing[31,32]. Indeed, we observed a depletion of high-scoring gRNAs within a window of −20 to 0 nt 5′ to the exon junction (see Supplementary Fig. 9h).

To improve our on-target model, we focused on PM gRNAs that target CDSs and increased the number of high-confidence model input observations from ~400 to nearly 3,000. Similar to the initial GFP screen, gRNA efficacies were distributed along the coding region in a nonrandom manner (Fig. 3a–c). We repeated the assessment of features that may affect knockdown efficacy (see Supplementary Note 2 for details). Notably, the increased number of observations uncovered positional nucleotide preferences (see Supplementary Fig. 10a,b). The gRNA 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 knockdown. We derived an updated on-target model using 2,918 CDS-targeting gRNAs across all four tiling screens, and selected 35 of 644 evaluated features in a similar fashion to previously (see Methods, and also Supplementary Table 2, Supplementary Note 2 and Supplementary Data 2).

The $\mathrm{RF}_{\mathrm{combined}}$ model displayed improved prediction accuracy compared with the initial $\mathrm{RF}_{\mathrm{minimal}}$ model (from here on referred to as $\mathrm{RF}_{\mathrm{GFP}}$), explaining ~47% of the variance ($r^2$) with Spearman's correlation ($r_s$) of ~0.67 for the held-out data (Fig. 3d and see Supplementary Fig. 10c). Using tenfold 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 knockdown efficacy) separate gRNAs that performed well from those that performed poorly better than expected by chance (see Supplementary Fig. 10d). Furthermore, we performed leave-one-out cross-validation training on three datasets while predicting guide scores for the held-out fourth screen. The $\mathrm{RF}_{\mathrm{combined}}$ model generalized well for endogenous genes (mean ± s.d., $r_s = 0.63 \pm 0.01$) but was less predictive for the GFP transgene ($r_s = 0.33$) (see Supplementary Fig. 10e).

Finally, we compared the ability of both models, $\mathrm{RF}_{\mathrm{GFP}}$ and $\mathrm{RF}_{\mathrm{combined}}$, to correctly predict the knockdown efficacies for the two essentiality screens. Both screens were designed based on gRNA predictions made by the $\mathrm{RF}_{\mathrm{GFP}}$ model. In both cases, the $\mathrm{RF}_{\mathrm{combined}}$ model was in better agreement with the observed knockdown efficacies across all genes (Fig. 3f). Likewise, we found that the $\mathrm{RF}_{\mathrm{combined}}$ model showed improved agreement with the observed gRNA depletion on a per-gene basis for the ten most depleted genes in the A375 fitness screen ($\mathrm{RF}_{\mathrm{GFP}}$: $r_s = 0.46 \pm 0.16$; $\mathrm{RF}_{\mathrm{combined}}$: $r_s = 0.58 \pm 0.14$). Taken together, we show that our updated on-target model, $\mathrm{RF}_{\mathrm{combined}}$, can predict Cas13d gRNA target knockdown 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 ten highest-scoring crRNAs for the 5′-UTR, CDS and 3′-UTR of each transcript (see Supplementary Fig. 11a and 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 knockdown) quartile (Q4) (see Supplementary Fig. 11b). A large fraction of gRNAs is 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 knockdown efficacy. However, almost all transcripts contain top-scoring gRNAs (see 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 knockdown efficacy and, using these 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. Although all three perturbation 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, whereas RNAi is restricted to the cytosol, Cas13 allows for compartmentalized targeting (nucleus, cytosol and other subcellular 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 content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-020-0456-9.

Received: 1 July 2019; Accepted: 11 February 2020;
Published online: 16 March 2020

## References

1. Abudayyeh, O. O. et al. C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science* **353**, aaf5573 (2016).
2. Abudayyeh, O. O. et al. RNA targeting with CRISPR–Cas13. *Nature* **550**, 280–284 (2017).
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).
4. 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).
5. Konermann, S. et al. Transcriptome engineering with RNA-targeting type VI-D CRISPR effectors. *Cell* **173**, 665–676 (2018).
6. 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).
7. Freije, C. A. et al. Programmable inhibition and detection of RNA viruses using Cas13. *Mol. Cell* **76**, 1–12 (2019).
8. Poosala, P., Lindley, S. R., Anderson, K. M. & Anderson, D. M. Targeting toxic nuclear RNA foci with CRISPR-Cas13 to treat myotonic dystrophy. Preprint at *bioRxiv* https://doi.org/10.1101/716514 (2019).
9. Mahas, A., Aman, R. & Mahfouz, M. CRISPR-Cas13d mediates robust RNA virus interference in plants. *Genome Biol.* **20**, 1–16 (2019).
10. Kushawah, G. et al. CRISPR-Cas13d induces efficient mRNA knock-down in animal embryos. Preprint at *bioRxiv* https://doi.org/10.1101/2020.01.13.904763 (2020).
11. Gootenberg, J. S. et al. Nucleic acid detection with CRISPR-Cas13a/C2c2. *Science* **356**, 438–442 (2017).
12. Gootenberg, J. S. et al. Multiplexed and portable nucleic acid detection platform with Cas13, Cas12a, and Csm6. *Science* **360**, 439–444 (2018).
13. Cox, D. B. T. et al. RNA editing with CRISPR-Cas13. *Science* **358**, 1019–1027 (2017).
14. Li, J.et al. Targeted mRNA demethylation using an engineered dCas13b-ALKBH5 fusion protein. Preprint at *bioRxiv* https://doi.org/10.1101/614859 (2019).
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).
17. Jillette, N. & Cheng, A. W. CRISPR artificial splicing factors. Preprint at *bioRxiv* https://doi.org/10.1101/431064 (2018).
18. Anderson, K. M., Poosala, P., Lindley, S. R. & Anderson, D. M. Targeted cleavage and polyadenylation of RNA by CRISPR-Cas13. Preprint at *bioRxiv* https://doi.org/10.1101/531111 (2019).
19. Meeske, A. J., Nakandakari-Higa, S. & Marraffini, L. A. Cas13-induced cellular dormancy prevents the rise of CRISPR-resistant bacteriophage. *Nature* **570**, 241–245 (2019).
20. Meeske, A. J. & Marraffini, L. A. RNA guide complementarity prevents self-targeting in type VI CRISPR systems. *Mol. Cell* **71**, 791–801 (2018).
21. Liu, L. et al. The molecular architecture for RNA-guided RNA cleavage by Cas13a. *Cell* **170**, 714–720 (2017).
22. Tambe, A., East-seletsky, A., Knott, G. J., Connell, M. R. O. & Doudna, J. A. RNA binding and HEPN-nuclease activation are decoupled in CRISPR-Cas13a. *Cell Rep.* **24**, 1025–1036 (2018).
23. Zhang, C. et al. Structural basis for the RNA-guided ribonuclease activity of CRISPR-Cas13d. *Cell* **175**, 212–223 (2018).
24. Zhang, B. et al. Two HEPN domains dictate CRISPR RNA maturation and target cleavage in Cas13d. *Nat. Commun.* **10**, 2544 (2019).
25. Konermann, S. et al. Genome-scale transcriptional activation by an engineered CRISPR–Cas9 complex. *Nature* **517**, 583–588 (2015).
26. Replogle, J. M.et al. Direct capture of CRISPR guides enables scalable, multiplexed, and multi-omic Perturb-seq. Preprint at *bioRxiv* https://doi.org/10.1101/503367 (2018).
27. Doench, J. G. et al. Rational design of highly active sgRNAs for CRISPR–Cas9-mediated gene inactivation. *Nat. Biotechnol.* **32**, 1262–1267 (2014).
28. McFarland, J. M. 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).
29. Doench, J. G. et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR–Cas9. *Nat. Biotechnol.* **34**, 184–191 (2016).
30. Briese, M. et al. A systems view of spliceosomal assembly and branchpoints with iCLIP. *Nat. Struct. Mol. Biol.* **26**, 930–940 (2019).
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).
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).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Methods

**Cloning of Cas13 nuclease, gRNAs and destabilized EGFP plasmids.** Using Gibson cloning, we modified the EF1a-short (EFS) promoter-driven lentiCRISPRv2 (Addgene, catalog no. 52961) and lentiCas9-Blast (Addgene, catalog no. 52962) plasmids with several different transgenes[33]. For the destabilized enhanced GFP (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 Cas13 orthologs, we cloned effector proteins (*Pgu*Cas13b: Addgene, catalog no. 103861; *Psp*Cas13b: Addgene, catalog no. 103862; *Rfx*Cas13d: Addgene, catalog no. 109049) and their DR sequences (*Pgu*Cas13b: Addgene, catalog no. 103853; *Psp*Cas13b: Addgene, catalog no. 103854; *Rfx*Cas13d: Addgene, catalog no. 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 *Pgu*Cas13b, *Psp*Cas13b or *Rfx*Cas13d and (NLS/NES) was either a nuclear localization signal or a nuclear export signal. To generate doxycycline-inducible Cas13d cell lines, we cloned NLS-*Rfx*Cas13d-NLS (Addgene, catalog no. 109049) into TetO-(Cas13)-WPRE-EFS-rtTA3-2A-Blast. For the screens, we changed the DR in the lentiGuide-Puro vector (Addgene, catalog no. 52963) to contain the *Rfx*Cas13d DR using Gibson cloning to create lentiRfxGuide-Puro (pLentiRNAGuide)[33].

Guide cloning was done as described previously[33]. 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 made available by Addgene (https://www.addgene.org/Neville_Sanjana/).

**Cell culture and monoclonal cell line generation.** HEK293FT cells were acquired from Thermo Fisher Scientific (catalog no. R70007) and A375 cells were acquired from American Type Culture Collection (catalog no. CRL-1619). HEK293FT and A375 cells were maintained at 37 °C with 5% CO$_2$ in D10 medium: Dulbecco's modified Eagle's medium with high glucose and stabilized L-glutamine (Caisson, catalog no. DML23) supplemented with 10% fetal bovine serum (Serum Plus II, Sigma-Aldrich, catalog no. 14009C) and no antibiotics.

To generate doxycycline-inducible *Rfx*Cas13d-NLS HEK293FT and A375 cells, we transduced cells with an *Rfx*Cas13d-expressing lentivirus at a low multiplicity of infection (MOI $<$ 0.1) and selected with 5 μg ml$^{-1}$ of blasticidin S (Thermo Fisher, catalog no. A1113903). Single-cell colonies were picked after sparse plating. Clones were screened for Cas13d expression using western blotting and mouse anti-FLAG M2 antibody (Sigma, catalog no. F1804).

For the GFP-tiling screen, *Rfx*Cas13d-expressing cells were transduced with pLentiEGFPdestabilized lentivirus at a low MOI (<0.1) and selected with 100 μg ml$^{-1}$ of hygromycin B (Thermo Fisher, catalog no. 10687010) for 2 d. Single-cell colonies were grown by sparse plating. Resistant and GFP-positive clonal cells were expanded and screened for homogeneous GFP expression by flow cytometry.

**Transfection and flow cytometry.** For all transfection experiments, we seeded $2\times10^5$ HEK293FT cells per well of a 24-well plate before transfection (12–18 h) and used 500 or 750 ng plasmid together with a 5:1 ratio of Lipofectamine 2000 (Thermo Fisher, catalog no. 11668019) or 1 mg ml$^{-1}$ of polyethylenimine (PEI, Polysciences, catalog no. 23966) to DNA (for example, 2.5 μl of Lipofectamine 2000 or PEI mixed with 0.5 μg of plasmid DNA). Flow cytometry or FACS was performed at 48 h post-transfection. All transfection experiments were performed in biological triplicates.

For the Cas13 ortholog comparison (see Supplementary Fig. 1a–c), we cloned the effector proteins (*Pgu*Cas13b: Addgene, catalog no. 103861; *Psp*Cas13b: Addgene, catalog no. 103862; *Rfx*Cas13d: Addgene, catalog no. 109049) and their direct repeat sequences (*Pgu*Cas13b: Addgene, catalog no. 103853; *Psp*Cas13b: Addgene, catalog no. 103854; *Rfx*Cas13d: Addgene, catalog no. 109053) as described above. We co-transfected the pLentiRNACRISPR constructs together with a GFP expression plasmid (pLentiEGFPdestabilized) in a 2:1 molar ratio. The gRNA length comparison (see Supplementary Fig. 1d) was performed using previously published *Rfx*Cas13d constructs (Addgene, catalog nos. 109049 and 109053), except that we removed the GFP cassette from the *Rfx*Cas13d plasmid. The modified *Rfx*Cas13d construct and guide plasmids were co-transfected together with pLentiEGFPdestabilized in a 2:2:1 molar ratio. For the DR modification experiment (see Supplementary Fig. 6c) we transfected *Rfx*Cas13d-expressing cells, starting doxycycline induction (1 μg ml$^{-1}$) 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 (see Fig. 2b) and CD46 screen validation (see Supplementary Fig. 9c), we transfected *Rfx*Cas13d-expressing cells with a gRNA-expressing plasmid. Then, 48 h post-transfection, the cells were stained for the respective cell-surface protein for 30 min at 4 °C and measured by FACS (BioLegend: CD46 no. 352405 clone TRA-2–10, CD71 (TFRC) no. 334105 clone CYIG4).

For the GFP screen validation (see Fig. 1e) and seed validation experiments (see Fig. 1h), we co-transfected *Rfx*Cas13d-expressing cells with a gRNA-expressing plasmid and pLentiEGFPdestabilized at a 1:1 molar ratio. At 48 h post-transfection, the cells were analyzed by flow cytometry.

To assess the upstream U context (see Supplementary Note 1), we transfected upstream-U-context-modified pLentiEGFPdestabilized-MCS plasmid together with a crRNA plasmid into *Rfx*Cas13d-expressing cells in a 2:1 molar ratio. Each GFP-upstream, U-context plasmid was co-transfected with both a targeting and a nontargeting gRNA used to calculate the knockdown, because 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 10,000 in silico randomized 52mers with $(A_{24},C_{14},G_{14})$ with minimal predicted RNA secondary structure, as determined by RNAfold[34] 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, catalog no. L34963). For each sample we analyzed at least 5,000 cells. If cell numbers varied, we randomly downsampled all conditions to the same number of cells before calculating the mean fluorescence intensity. For GFP co-transfection experiments, we considered only the percentage of transfected cells with the highest GFP expression, determined by comparing the nontargeting control with wild-type control cells. For the upstream, U-context, co-transfection experiments, we considered whole-cell populations.

For knockdown experiments of endogenous genes (see Fig. 2b and Supplementary Fig. 9c), we determined the percentage of transfected cells with a lower target gene signal than the nontargeting control, in the condition with the highest observed knockdown. For all conditions, we analyzed the same bottom percentage of cells. For the selected cells, we compared the mean fluorescence intensity of targeting guides relative to nontargeting guides to determine the percentage knockdown. 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 the percentage GFP-positive cells in the nontargeting guide control.

**Screen library design and pooled oligo cloning.** To design the *Rfx*Cas13d gRNA library for GFP, we used the entire EGFP coding sequence (without the start codon). In silico, we generated all PM 27mer gRNAs 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 gRNAs and introduced one random nucleotide conversion at each position (single-nucleotide mismatch (SM) set $n$ = 2,700). From these 100, we randomly sampled 17 gRNAs and introduced 26 or 25 consecutive double (CD set $n$ = 442) and consecutive triple (CT set $n$ = 425) mismatches, respectively. We sampled an additional 13 gRNAs from the SM set (in total, 30 gRNAs) and introduced 100 random double mismatches at any position for each gRNA, if not present already in the set of 17 CD mismatches (RD set $n$ = 3,000). In total, we designed 6,967 GFP-targeting guides and added 533 nontargeting guides (NT set) of the same length from randomly generated sequences that did not align to the human genome (hg19) with fewer 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; GENCODE v.19) and longest 3′-UTR isoform (CD46: ENST00000367042.1; CD55: ENST00000367064.3; CD71: ENST00000360110.4). As described above, we generated all PM 23mers, and selected ~2,000 evenly spaced gRNAs per target. In addition to PM, SM, RD and NT sets, as described earlier, we included for each target a set of guide-length variant (LV set $n$ = 450) gRNAs 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 PM sequences (RC set $n$ = 300). Further details are given in Supplementary Data 5.

For both targeted essentiality screens, we used the DEMETER2 v.5 (ref. [28]) dataset from the Cancer Dependency Map portal (DepMap) to determine essential and control genes. Specifically, we selected essential genes with low log$_2$(FC) enrichments across all cell lines and in the respective assay cell line (see 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 an expression level of fewer than ~150 transcripts per million. We predicted gRNA efficiencies using the minimal RF$_{GFP}$ model and removed all guides with matches or partial matches elsewhere in the transcriptome. We allowed up to three mismatches when looking for potential off-targets. From the set of remaining PM gRNA 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 nonoverlapping regions of the target transcripts. For the A375 cell line targets, we selected the top 20 high-scoring gRNAs. For the set of 20 low-scoring guides, we chose among the bottom 60 to reduce the overlap of gRNAs that fall into the same region. In this way, we assayed twenty genes in HEK293FT cells targeting ten essential and ten control genes with three low-scoring and three high-scoring guides, as well as three nontargeting guides ($n=123$). For the A375 screen, we targeted 100 genes (35 essential and 65 control genes) with 40 guides each (20 high scoring and 20 low scoring) and included 680 nontargeting sequences ($n=4{,}680$).

All large-scale, 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 synthesized as standard oligonucleotides (IDT), array cloned, confirmed by Sanger sequencing and subsequently pooled using equal amounts. Complete library representation with minimal bias (90th percentile:10th percentile crRNA read ratio: 1.68–2.17) was verified by Illumina sequencing (MiSeq).

**Pooled lentiviral production and screening.** Lentivirus was produced via transfection of library plasmid with appropriate packaging plasmids (psPAX2: Addgene, catalog no. 12260; pMD2.G: Addgene, catalog no. 12259) using PEI reagent in HEK293FT. At 3 d post-transfection, viral supernatant was collected and passed through a 0.45-µm filter and stored at −80 °C until use.

Doxycycline-inducible *Rfx*Cas13d-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 1,000× guide representation in the selected cell pool per infection replicate using a standard spinfection protocol. We generated either two or three independent replicate experiments. After 24 h, *Rfx*Cas13d expression was induced by addition of 1 µg ml$^{-1}$ of doxycycline (Sigma, catalog no. D9891) and cells were selected with 1 µg ml$^{-1}$ of puromycin (Thermo Fisher, catalog no. A1113803), resulting in ~30% cell survival. Puromycin selection was complete by 48 h post-puromycin addition. Assuming independent infection events, we determined that ~83% of surviving cells received a single guide RNA construct. Cells were passaged every 2 d, maintaining at least the initial cell representation and supplementing with fresh doxycycline.

The tiling screens were terminated after 5–10 d. For all targets we noted maximal knockdown after 2–4 d (data not shown). For cell-surface proteins, cells were stained in batches of $1\times10^7$ cells for 30 min at 4 °C (BioLegend: CD46 clone TRA-2–10 no. 352405, 3 µl per $1\times10^6$ cells; CD55 clone JS11 no. 311311, 1.5 µg per $1\times10^6$ cells; CD71 clone CYIG4 no. 334105, 4 µl per $1\times10^6$ cells). We collected unsorted samples for input gRNA representation of approximately 1,000× coverage for each sample and sorted at least another 1,000× representations into the assigned bins based on their signal intensities (GFP: lowest 20%, 20%, 20% and remaining highest 40%, see Supplementary Fig. 2a; CD proteins lowest 20% and highest 20%, see Supplementary Fig. 9b and Supplementary Data 5). Cells were washed in phosphate-buffered saline and frozen at −80 °C until sequencing library preparation. In each case, the bin containing the lowest 20% represented the strongest target knockdown.

The essentiality screens were started (day 0) on complete puromycin selection, which was at 5 d after transduction. Cells were passaged every 2–3 d, maintaining at least the initial cell representation and supplementing with fresh doxycycline. At day 0 (input) and every 7 d, we collected a >1,000× 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 (Thermo Fisher, catalog no. K0722) using $2\times10^6$ cells or fewer per column. The crRNA readout was performed using two rounds of PCR[35]. 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 µl volume with 2 µg of gDNA in each reaction, limited by the amount of extracted gDNA (total gDNA ranged from 8 µg to 50 µg per sample with an estimated representation of $10^6$ diploid cells per ~6.6 µg of gDNA. PCR1: 4 µl$^{-1}$ of 5× Q5 buffer, 0.02 U µl$^{-1}$ of Q5 enzyme (M0491L), 0.5 µM forward and reverse primers, and 100 ng gDNA µl$^{-1}$. PCR conditions: 98 °C for 30 s, 24× (98 °C for 10 s, 55 °C for 30 s, 72 °C for 45 s), 72 °C for 5 min).

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 adapters, barcodes and stagger sequences to prevent monotemplate sequencing issues. Complete sequences of the five forward and three reverse Illumina PCR2 readout primers used are shown in Supplementary Data 4 (PCR2: 50 µl of 2× Q5 master mix (NEB no. M0492S), 10 µl of PCR1 product and 0.5 µM forward and reverse PCR2 primers in 100 µl. PCR conditions: 98 °C for 30 s, 17× (98 °C for 10 s, 63 °Cor 30 s, 72 °C for 45 s), 72 °C for 5 min).

Amplicons from the second PCR were pooled by screen experiment (for example, all GFP screen samples) in equimolar ratios (by gel-based band densitometry quantification) and then purified using a QiaQuick PCR Purification kit (Qiagen, catalog no. 28104). Purified products were loaded on to a 2% E-gel and gel extracted using a QiaQuick Gel Extraction kit (Qiagen, catalog no. 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 1×150 v.2.5.

Reads were de-multiplexed 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 gRNA length by searching for known anchor sequences relative to the guide sequence. For the tiling screens, preprocessed reads were either aligned to the designed crRNA reference using bowtie[36] (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 PM and unique alignments. Preprocessed gRNA sequences from the essentiality screens were aligned allowing for up to one mismatch (-v 1 -m 1). Alignment statistics are available in Supplementary Data 6. The raw gRNA counts (see Supplementary Data 7) were normalized, separated by screen dataset using a median-of-ratios method as in DESeq2 (ref. [37]) and underwent batch correction using combat implemented in the SVA R package[38]. Nonreproducible 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 gRNA across all sample-centric comparisons. We removed all crRNA counts within the top *X*% of residuals across all samples (GFP: 2%; CD proteins: 0.5%; essentiality screen: no outlier removal). For the GFP screen, we remove only outliers on a per-sample basis as needed (but not the entire gRNA). For CD46, CD55 and CD71 screens, as the number of outliers was small, we decided to remove the entire gRNA from the analysis. Supplementary Table 3 indicates all filtering steps applied.

Processed crRNA counts are available in Supplementary Data 8. The gRNA enrichments were calculated by taking the $\log_2$ of the count ratios between a bin or time point and the corresponding input sample. Consistency between replicates was estimated using robust rank aggregation (RRA)[39]. $\Delta\log_2$(FC) for mismatching guides was calculated by subtracting the $\log_2$(FC) of the PM reference guide. For the tiling screens, all plots and analyses were performed using the mean gRNA enrichments of bin 1 (=bottom 20%) across replicates, unless indicated otherwise. Similarly, we used the mean gRNA enrichments relative to day 0 across replicates for the essentiality screen. The gRNA enrichment scores ($\log_2$(FC)) are available in Supplementary Data 1. In all combined analyses across all four tiling screens, we scaled the observed $\log_2$(FC) separately to improve comparability. For the generation of the combined on-target model, we normalized $n = 2{,}918$ selected CDS-targeting gRNAs across the four tiling screens to the same scale before training and testing the model. To do so, for each dataset *D*, we computed the upper and lower quartiles of the guide $\log_2$(FC) ($\mathrm{UQ}_D$ and $\mathrm{LQ}_D$, respectively), as well as the corresponding quartiles for the $\log_2$(FC) among all the datasets pooled together ($\mathrm{UQ}_P$ and $\mathrm{LQ}_P$). We then updated each fold change, *x*, as follows: $x\hat{} = ((x - \mathrm{LQ}_D) / (\mathrm{UQ}_D - \mathrm{LQ}_D) \times (\mathrm{UQ}_P - \mathrm{LQ}_P) + \mathrm{LQ}_P)$. 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.** The crRNA secondary structure and MFEs were derived using RNAfold [--gquad] on the full-length crRNA (DR + guide) sequence[34]. For building the combined on-target model and for testing the $\mathrm{RF}_{\mathrm{GFP}}$ model on the combined dataset, we assumed 23mer gRNAs 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 previously[40]. We performed a grid-search calculating the RNA accessibility for each target nucleotide in a window of −20 bases downstream of the target site to +20 bases upstream of the target site, assessing the unpaired probability of each nucleotide over 1–50 bases for all PM guides. Then, we calculated Pearson's correlation coefficient between the $\log_{10}$(transformed unpaired probabilities) and the observed gRNA $\log_2$(FC) for each point and window relative to the gRNA. RNA–RNA hybridization between the gRNA and its target site was calculated using RNAhybrid [-s -c][41]. We calculated the RNA-hybridization MFE for each gRNA 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 $\log_2$(FC) or used partial correlation to account for the crRNA-folding MFE. In each case, we computed Pearson's correlation.

**Assessing gRNA nucleotide composition.** The gRNA composition was derived by calculating the nucleotide probability within the respective gRNA sequence length. To assess the presence of sequence constraints similar to a previously described anti-tag[20] or 5′- and 3′-protospacer flanking sequences, we ranked all PM gRNAs by their $\log_2$(FC) enrichment within each screen separately. We selected the top and bottom 20% enriched/depleted gRNAs, and calculated the positional nucleotide probability for the 4 nt upstream and downstream relative to the gRNA match. To assess nucleotide preferences at any gRNA match position, in addition to upstream and downstream nucleotides, we selected the top 20% of the $\log_2$(FC)-ranked PM guides as described above and calculated nucleotide preferences as described previously[27]. In brief, we calculated the probability of each nucleotide at each position for the top gRNAs and all gRNAs. The effect size is the difference of nucleotide probability by subtracting the values from all guides from the top guides (Δ nt probability). *P* values were calculated from the binomial distribution with a baseline probability estimated from the full-length mRNA target sequence for all PM gRNAs. *P* values were adjusted using Bonferroni's 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 1–50 nt centered around the position of interest (for example, $p=-18$ with $w=11$ summarizes the nucleotide probability in a window from −23 to −13, with +1 being the first base of the gRNA). We evaluated $p$ for all positions within 75 nt upstream and downstream of the gRNA. The nucleotide probability of each point was then correlated with the observed $\log_2$(FC) gRNA enrichments for all PM gRNAs, either directly or using partial correlation accounting for crRNA-folding MFE. In each case we used Pearson's correlation.

The RNA context around single-nucleotide mismatches was assessed accordingly with a slight modification. Here the nucleotide context was assessed relative to the mismatch position summarizing the nucleotide probability in a window of 1–15 nt to either side (for example, $p=8$ with $w=5$ summarizes the nucleotide content in a window of 11 nt from 23 to 13). For more details on $p$ and $w$, please see Supplementary Fig. 5b. We used all 2,700 single-nucleotide mismatch guides in the GFP-tiling screen (100 gRNAs × 27 mismatched positions per guide). The nucleotide context of each position and each window size was then correlated with the observed $\Delta\log_2$(FC) relative to the PM reference gRNA, either directly or using partial correlation to account for crRNA-folding MFE. In each case, we used Pearson's correlation.

**On-target model selection.** An explanation for all selected features for the $RF_{GFP}$ and $RF_{combined}$ models can be found in Supplementary Tables 1 and 2, respectively. The $RF_{combined}$ 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 5th and 95th percentiles, with a mean set to the 5th 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 RF models can be found in Supplementary Table 4.

To evaluate and compare model performances, we randomly sampled 1,000 bootstrap datasets from the data of PM gRNA $\log_2$(FC) response values and selected features. We used 399 data points for the initial $RF_{GFP}$ model and 2,918 data points for all CDS-annotating PM guides across the four tiling screens. For the $RF_{combined}$ model, we normalized the observed $\log_2$(FC) values data before training and testing as described earlier (see 'Screen readout and read analysis' above). Normalized response values showed better generalizability compared with unnormalized or scaled $\log_2$(FC). For each bootstrap sample, 70% of the data was used for training and the remaining 30% was held out for testing, ensuring a 70:30 split for each screen dataset when testing the $RF_{combined}$ 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 Spearman's correlation coefficient $r_s$ and Pearson's $r^2$ for the held-out data. We compared a variety of different methods[40] (Supplementary Table 5).

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 nucleotides and dinucleotides of the guide target sites, and their upstream and downstream flanking 4 nt. 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 evaluated a set of 15 defined features (see Supplementary Table 1), alongside 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 $RF_{GFP}$ model and monitored the model performance as described in the paragraph above. At each iteration, the RF model performed slightly better than any other learning approach. Reducing the features to fewer than the selected six features ($RF_{minimal}=RF_{GFP}$) reduced the model performance. For the combined on-target model, we did not iteratively reduce the set of 35 selected features. We compared the $RF_{GFP}$ model with an SVM+L1 model similar to one of the first CRISPR–Cas9 on-target models. Specifically, we used one-hot encoding for all 35 (31) nt positions considered (27 gRNA positions in the case of the GFP screen, 23 for the combined model, and 8 additional positions with 4 nt upstream and 4 nt downstream). Considering all positions, the feature space contained 140 single-nucleotide features, 544 dinucleotide features and the GC content (685 non-all-zero features). Here we used tuning (see Supplementary Table 5 for parameters) to increase model performance for SVM+L1 specifically. For both RF models, one-hot-encoded features did not lead to high Spearman's correlation coefficient $r_s$ for the held-out data.

For further evaluation of the RF models we used tenfold cross-validation by randomly partitioning the data into ten equally sized partitions, ensuring even contributions from each screen to each partition. We trained the model ten times on 90% of the data and predicted the held-out 10%. For each data point, we assigned the known gRNA efficacy quartile based on the $\log_2$(FC) enrichment and compared it with the predicted efficacy quartiles in the held-out data. For the $RF_{GFP}$ model, we found that the model could readily separate poorly performing from effective gRNAs. 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 (see Supplementary Fig. 7e). We also assessed the predicted guide score by calculating the median predicted guide score for the top- and bottom-ranked gRNAs in the 10% held-out data, based on the known $\log_2$(FC) rank for all ten cross-validation folds (top/bottom $n=2$, 4, 8, 16, 32, 64, 128 or 256 gRNAs). To compute the null distribution, we calculated the median predicted guides scores of randomly selected gRNAs 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's rank correlation for the predicted guide efficiency of the held-out fourth screen to the observed $\log_2$(FC) enrichments.

To make the guide score easier to interpret, we standardized the guide score to a [0, 1] interval preserving the distribution between the 5th and 95th percentiles. Normalized values exceeding the [0, 1] interval were set to 0 or 1, respectively. The final $RF_{GFP}$ model was trained on all data points for PM guides using the 6 selected features with 1,500 regression trees. The model explains 36.9% of the observed variance, with a mean of squared residuals of 0.139. Supplementary Table 6 shows the feature contribution for the $RF_{GFP}$ model.

Similarly, final $RF_{combined}$ 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 negligible performance improvements compared with default settings. The model (mtry = 12, ntree = 2,000) explains 47.16% of the observed variance, a mean of squared residuals of 0.168, and the feature contribution as indicated in Supplementary Table 7 ranked by importance.

***Rfx*****Cas13d guide scoring.** We created a user-friendly R script that readily predicts *Rfx*Cas13d on-target guide scores. The only user-provided argument is a single-entry FASTA file input, minimally, of 30 nt which represents the target sequence, such as a transcript isoform sequence. The software first generates all possible 23mer gRNAs, collects all required features and predicts gRNA efficacies. The only filter applied removes gRNAs with homopolymers of five or more Ts and six or more Vs (V = A, C, G). Such gRNAs may trigger early transcript termination for PolIII transcription or cause difficulties during oligo synthesis. The software returns a FASTA file with gRNA sequences ranked by the predicted standardized guide score. In addition, a csv file is created 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 at 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 gRNA sequences (see Supplementary Data 3). We have made all guide score predictions available online (https://cas13design.nygenome.org).

***Rfx*****Cas13d guide-scoring validation.** To validate that our initial $RF_{GFP}$ 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 three guides present in the low-scoring quartiles (Q1 and Q2) and three guides in the high-scoring quartiles (Q3 and Q4). We selected the guides to be nonoverlapping and to reside in three 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 gRNAs as described (see Screen library design and pooled oligo cloning). For the HEK293FT cell screen, we compared the guide depletion of 4 groups of 30 guides (essential gene targeted by high-scoring or low-scoring guide, and control genes targeted by high-scoring or low-scoring guide). We expected the greatest depletion for the 30 high-scoring gRNAs targeting essential genes. Similarly, we compared the relative guide depletion of the same 4 groups of gRNAs 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 RRA[39] to assign a $P$ value based on the consistency of $\log_2$(FC)-based rank of the best (most depleted) $N$ gRNAs per gene (for $N = 1$, 5 or 20) across the two A375 screen replicates. The $-\log_{10}$($P$ values) were then compared with other growth screens (RNAi and Cas9) using Spearman's rank correlation. Specifically, we compared the RRA-derived $\log_{10}$($P$ value) with the $\log_2$(FC) from an RNAi-based DEMETER2 version 5 repository[28] and the merged STARS scores from a Cas9-based approach[29]. For the correlation, we used only genes with values present in all scores/modalities (all essential genes: $n=35$; control genes: $n=15$).

Furthermore, we used the $\log_2$(FC) guide depletion values to compare the predictive value of the $RF_{GFP}$ and $RF_{combined}$ models. Specifically, for both essentiality screens we used ten essential genes (all in HEK293FT and the ten most depleted in A375 cells) and correlated the predicted guide scores from both models to the observed $\log_2$(FC) guide depletion scores (normalized to 0–100% per gene) of all detected gRNAs (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 gRNAs per gene in the A375 screen.

**Data representation.** In all box plots, boxes indicate the median and interquartile ranges (IQRs), with whiskers indicating either 1.5× the IQR or the most extreme data point outside the 1.5-fold IQR. All transfection experiments show the mean of three replicate experiments, with individual replicates plotted as points.

**Reporting Summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

Screen data have been deposited at the Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/) with the accession no. GSE142675. All code and software to reproduce our entire analyses are available on our gitlab repository (https://gitlab.com/sanjanalab/cas13). Moreover, we provide precomputed gRNA 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

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

## References

33. Sanjana, N. E., Shalem, O. & Zhang, F. Improved vectors and genome-wide libraries for CRISPR screening. *Nat. Methods* **11**, 783–784 (2014).
34. Lorenz, R. et al. ViennaRNA package 2.0. *Algorithms Mol. Biol.* **6**, 26 (2011).
35. Shalem, O. et al. Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84–88 (2014).
36. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, 212 (2009).
37. 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).
38. 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).
39. Kolde, R., Laur, S., Adler, P. & Vilo, J. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics* **28**, 573–580 (2012).
40. Agarwal, V., Subtelny, A. O., Thiru, P., Ulitsky, I. & Bartel, D. P. Predicting microRNA targeting efficacy in *Drosophila*. *Genome Biol.* **19**, 1–23 (2018).
41. Krueger, J. & Rehmsmeier, M. RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic Acids Res.* **34**, 451–454 (2006).

## 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 New York University and New York Genome Center startup funds, National Institutes of Health (NIH)/National Human Genome Research Institute (grant nos. R00HG008171, DP2HG010099), NIH/National Cancer Institute (grant no. R01CA218668), Defense Advanced Research Projects Agency (grant no. D18AP00053), the Sidney Kimmel Foundation, the Melanoma Research Alliance, and the Brain and Behavior Foundation. A.M.-M. is supported by a CONACyT-Mexico Fellowship (no. 412653).

## Author 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 and analyzed the experiments. H.H.W. analyzed the screen data, and built the gRNA prediction software and 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 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 adviser to Vertex.

## Additional information

**Supplementary information** is available for this paper at https://doi.org/10.1038/s41587-020-0456-9.

**Correspondence and requests for materials** should be addressed to N.E.S.

**Reprints and permissions information** is available at www.nature.com/reprints.

nature research

Corresponding author(s): Sanjana, Neville

Last updated by author(s): Jan 30, 2020

# Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

## Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a | Confirmed

☐ ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement

☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*

☐ ☒ A description of all covariates tested

☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

☐ ☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted
*Give P values as exact values whenever suitable.*

☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

☐ ☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

☐ ☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

*Our web collection on statistics for biologists contains articles on many of the points above.*

## Software and code

Policy information about availability of computer code

Data collection: No custom code was used for data collection.

Data analysis: FlowJo for FACS, DNAStar and Benchling for sequence analysis. The data processing and analysis was done using UNIX, python and RStudio. See Method section within Supplementary File for a detailed description. Custom code for guide scoring and the web-tool has been made available. All code and software to reproduce our entire analyses are available on our gitlab repository (https://gitlab.com/sanjanalab/cas13).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

## Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data and resource availability. Screen data are being deposited to GEO (https://www.ncbi.nlm.nih.gov/geo/) with the accession number GSE142675. Plasmids and libraries have been deposited to Addgene. All code and software to reproduce our entire analyses are available on our gitlab repository (https://gitlab.com/sanjanalab/cas13). Other data and materials that support the findings of this research are available from the corresponding author upon request.

# Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

# Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

All screens (tiling screens and fitness screens) have been conducted in two or three replicate experiments. Detailed statistics for included guide numbers and guide classes can be found in Supplementary Data 5. All transfection experiments have been conducted in three independent biological replicate transfections. The final predictive guide model (RFcombined) was constructed using features for all 2,918 perfect matching guides that match to coding sequences of the four targetd genes (GFP, CD46, CD55, CD71).

Data exclusions

The screen was conducted in two or three replicates. Non-reproducible technical outliers were removed by applying pair-wise linear regression for each sample, collecting the residuals and taking the median value for each crRNA 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, we remove the entire guide RNA from the analysis. For the fitness screens, we did not apply outlier removal. Each step in this process is detailed in the Supplementary Methods. All code to reproduce our procedure is available here: https://gitlab.com/sanjanalab/cas13. Applied exclusion criteria have been tailored to our data and were not pre-established.

Replication

The GFP and CD46-tiling screens, the HEK293 fitness screen and all transfections have been conducted in three replicate experiments. The CD55 and CD71-tiling screens have been conducted in two replicate experiments. We confirm that all finding have been replicated in each screen independently.

Randomization

The evaluation of machine learning approaches to derive a predictive model entailed random sampling and the 10-fold cross-validation (R base::sample)

Blinding

Blinding is not relevant to our study because it is not a subjective trial and the results presented here are purely based on objective description of our novel experimental technology.

# Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

## Materials & experimental systems

| n/a | Involved in the study |
|---|---|
| □ | ☒ Antibodies |
| □ | ☒ Eukaryotic cell lines |
| ☒ | □ Palaeontology |
| ☒ | □ Animals and other organisms |
| ☒ | □ Human research participants |
| ☒ | □ Clinical data |

## Methods

| n/a | Involved in the study |
|---|---|
| ☒ | □ ChIP-seq |
| □ | ☒ Flow cytometry |
| ☒ | □ MRI-based neuroimaging |

## Antibodies

Antibodies used

BioLegend:
CD46 #352405 clone TRA-2-10 Lot: B286717 (3µl per 1x10^6 cells),
CD55 #311311 clone JS11 Lot: B249208 (1.5µg per 1x10^6 cells),
CD71 (TFRC) #334105 clone CYIG4 Lot:B276134 (4µl per 1x10^6 cells)

Validation

Antibodies were validated by the vendors. All antigens used were validated before. An extended list is available from the vendor. Examples are:
CD46 (Antigen: Cardone J, et al. 2010. Nat. Immunol. 11:862.);
CD55 (Antigen: Peyron P, et al. 2000. J. Immunol. 165:5186; Product: Rhys HI, et al. 2018. EBioMedicine. 29:60);
CD71 (Antigen: Hentze M, et al. 1996. P. Natl. Acad. Sci. USA 93:8175; Product: Segal JM, et al. 2019. Nat Commun. 10:3350)

# Eukaryotic cell lines

Policy information about cell lines

| | |
|---|---|
| Cell line source(s) | HEK293FT cells (also denoted to as HEK293 cells) were acquired from Thermo Fisher (R70007). A375 cells were acquired from ATCC (CRL-1619). |
| Authentication | All cell lines used have been authenticated by the original vendors. |
| Mycoplasma contamination | All cell lines were tested as mycoplasma-free using Lonza MycoAlert (#LT07-518). |
| Commonly misidentified lines<br>(See ICLAC register) | No commonly misidentified cell lines were used. |

# Flow Cytometry

## Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

## Methodology

| | |
|---|---|
| Sample preparation | For GFP, cells were washed with PBS before analysis/sorting by flow cytometry. For CD46, CD55 and CD71 the cells were stained for the respective cell surface protein for 30min at 4°C , washed and measured by flow cytometry. |
| Instrument | Flow cytometry data acquisition and sorting were performed on a Sony SH800 sorter. |
| Software | FlowJo (Threestar) was used for flow cytometry data analysis. |
| Cell population abundance | For the GFP-tiling screen, we collected 7.5x10^6 cells for an unsorted input crRNA representation (1000x coverage) and sorted at least another 7.5x10^6 cells into 4 bins based on their GFP-intensity (lowest 20%, 20%, 20% and remaining highest 40%).<br>For CD46, CD55 and CD71 tiling screens, we we collected 6x10^6 cells for an unsorted input crRNA representation (1000x coverage) and sorted at least another 6x10^6 cells into 2 bins based on their CD protein intensity (lowest 20% and highest 20%).<br>For the HEK293 fitness screen, we collected 2x10^6 cells per sample. For the A375 fitness screen, we collected 7x10^6 cells per sample. For both fitness screens this represents >1000x coverage for each sample. |
| Gating strategy | For FACS analysis of transfection experiments, 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. For flow cytometry experiments aside from the pooled screens, the gating strategy is shown in Supplementary Figure 1b. For the pooled screens, the gating strategies are shown in Supplementary Figure 2a and Supplementary Figure 9b. |

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.