ARTICLES
https://doi.org/10.1038/s41551-017-0178-6

nature biomedical engineering

Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs

Jennifer Listgarten[1*], Michael Weinstein[2,3*], Benjamin P. Kleinstiver[4,5,6], Alexander A. Sousa[4,5], J. Keith Joung[4,5,6], Jake Crawford[1], Kevin Gao[1], Luong Hoang[1], Melih Elibol[1], John G. Doench[7*] and Nicolo Fusi[1*]

Off-target effects of the CRISPR–Cas9 system can lead to suboptimal gene-editing outcomes and are a bottleneck in its development. Here, we introduce two interdependent machine-learning models for the prediction of off-target effects of CRISPR–Cas9. The approach, which we named Elevation, scores individual guide–target pairs, and also aggregates them into a single, overall summary guide score. We demonstrate that Elevation consistently outperforms competing approaches on both tasks. We also introduce an evaluation method that balances errors between active and inactive guides, thereby encapsulating a range of practical use cases. Because of the large-scale and computational demands of the prediction of off-target activities, we have developed a fast cloud-based service (https://crispr.ml) for end-to-end guide-RNA design. The service makes use of pre-computed on-target and off-target activity prediction for every genic region in the human genome.

Although the clustered regularly interspaced short palindromic repeats (CRISPR)–CRISPR-associated protein 9 (Cas9) system is routinely used, potentially avoidable off-target effects can complicate or hinder its use. The best way to mitigate off-target effects is to know when and where they occur, and then design guides to avoid them while balancing for on-target efficiency[1,2]. Such a balance may differ for different tasks. For example, the creation of cellular and animal models, or therapeutic uses of CRISPR–Cas9, will in general be far less tolerant of off-target effects than would be genome-wide screens, in which redundancy of targeting (that is, the use of multiple guide RNAs (gRNAs) for one gene) can be used to average out off-target effects. Nevertheless, reduction of off-target effects is desirable in all applications.

While numerous techniques have been developed to quantify off-target effects, such as genome-wide unbiased identification of double-strand breaks enabled by sequencing (GUIDE-seq)[3], high-throughput genome-wide translocation sequencing (HTGTS)[4], integrase-defective lentiviral vector (IDLV) capture[5], Digenome-seq[6,7], CIRCLE-seq[8], selective enrichment and identification of adapter-tagged DNA ends by sequencing (SITE-seq)[9], BLESS/BLISS[10–12] and other laboratory-based assays[1], scaling these assays to all gRNAs genome wide is not currently practically feasible for most research laboratories owing to cost, labour and the availability of general-purpose assays[1]. By contrast, as we show herein, machine-learning-based predictive modelling can leverage a small number of such data to learn statistical regularities of gRNA–target sequence pairs that cause off-target effects, as well as their aggregate effect on a cell. Such modelling therefore enables inexpensive and rapid in silico screening of off-target effects at a genome-wide level for gRNAs never before assayed[1,2,13].

There are two main use cases for off-target predictive modelling. The first is to understand how active a given off-target region is likely to be for a specific gRNA, which we refer to as gRNA–target scoring. This task is useful if one is concerned about a particular region of the genome, such as accidentally knocking out a tumour suppressor gene when trying to make an edit to disable an HIV entry receptor. The second use case is to obtain an overall summary score of all off-target activities for a given gRNA, to obtain a rank-order of good potential gRNAs for a gene, for example.

One can break down the off-target predictive modelling problem into three main tasks. Given a gRNA to evaluate for off-target activity one must:

脱靶预测建模问题可以分解为三个主要的任务。

(1) Search and filter genome wide for potential targets for one gRNA. For example, those regions of the genome that match the gRNA up to *N* number of nucleotide mismatches to the target site. Note, these sites are not deemed to be active off-targets until after step 2, which uses machine learning to distinguish the targets that are expected to be active from those that are not. This merely creates a short-list of potentially active sites.
(2) Score each potential target for activity from step 1. That is, assign a numeric value that indicates how much off-target activity is expected for one gRNA–target pair.
(3) Aggregate the scores from step 2 into a single off-target potential with which to assess the gRNA.

Numerous solutions have been presented for the first task of search and filter, including Cas-OFFinder[14], CRISPOR[15], CHOP-CHOP[16], e-CRISPR[17], CRISPR-DO[18], CROP-IT[19] and COSMID[20], which differ in the algorithms used to search, as well as the completeness of the search. Search completeness is dictated by options such as maximum number of mismatches, allowed protospacer adjacent motifs (PAMs) and the search algorithm used. For infrastructural efficiencies and

搜索完整性由诸如最大不匹配数量，允许的原型间隔符相邻基序（PAM）和使用相应的搜索算法。

[1]Microsoft Research, Cambridge, MA, USA. [2]Molecular, Cell, and Developmental Biology, and Quantitative and Computational Biosciences Institute, University of California Los Angeles, Los Angeles, CA, USA. [3]Zymo Research, Irvine, CA, USA. [4]Molecular Pathology Unit & Center for Cancer Research, Massachusetts General Hospital, Charlestown, MA, USA. [5]Center for Computational and Integrative Biology, Massachusetts General Hospital, Charlestown, MA, USA. [6]Department of Pathology, Harvard Medical School, Boston, MA, USA. [7]Broad Institute of MIT and Harvard, Cambridge, MA, USA. Jennifer Listgarten, Michael Weinstein, John G. Doench and Nicolo Fusi contributed equally to this work. *e-mail: jennifer.listgarten@gmail.com; mweinstein@zymoresearch.com; jdoench@broadinstitute.org; fusi@microsoft.com

Table 1 | Summary of CRISPR gRNA design services that include off-target scoring

Shorthand	On-target scoring	Off-target scoring	Off-target aggregator	On-target interface	Off-target interface
Elevation (this work) & Azimuth[1]	Machine-learning-based models	Machine-learning-based models	Machine-learning-based models	Human exome targets pre-computed website; cloud API (application programming interface) for re-use in code and Excel; source code	Human exome targets pre-computed website; source code for any target
MIT server[2]	Hand-crafted rules	Hand-crafted rules	Hand-crafted rules	Website	Website
CRISPR-DO[18]	Re-uses rules from ref. [21]	Re-uses rules from MIT server[2]	As in MIT server[2]	Website; source code	Website; source code
CRISPOR[15]	Re-uses rules from multiple papers[1,21–26]	Re-uses rules from MIT server[2]	As in MIT server[2]	Website; source code	Website; source code
Broad GPP[1]	Machine-learning-based models	Newly developed rules based on data	Not genome wide (only relative within-gene scores)	Website; source code	Website; source code
E-CRISPR[17]	Hand-crafted rules, and rules from refs [18,21,23]	Hand-crafted rules	Hand-crafted rules	Website	Website
CHOP-CHOP[16]	Re-uses rules from ref. [21] by default, and refs [1,22,23]	Counts the number of off-targets but does not score them	Not available	Website	Website

ease of integration with our cloud service, we created our own system to perform search and filtering (Elevation-search); for the purposes herein, we used the same parameters as in ref. [1]. The second and third steps, of scoring and aggregation, have been explored considerably less than the search and filter step, and are the focus of this work. A list of available gRNA design services that perform one or both of these steps is shown in Table 1. The only existing tools that return aggregation scores are the Massachusetts Institute of Technology (MIT) web server[2], CRISPOR[15] and CRISPR-DO[18]; the latter two re-implement the MIT web server rules. CHOP-CHOP[16] counts the number of potential off-targets without scoring them; CROP-IT[19] uses a hand-crafted series of rules and has been shown to be substantially outperformed by the MIT web server[15]. In addition, the CFD (cutting frequency determination) method[1] has been shown to outperform the MIT web server on gRNA–target pair scoring[15], but ~~the CFD web server does not perform genome-wide off-target aggregation.~~

Although alternative CRISPR systems that may possess improved specificity (for example, Cpf1 endonuclease[27,28]) are being developed, these systems are still in their relative infancy, and Cas9 from *Streptococcus pyogenes* remains the workhorse endonuclease of choice.

For each of gRNA–target scoring and gRNA summary scoring, we developed ~~a machine-learning approach~~ that substantially improved on the state-of-the-art for the respective task, as demonstrated through our experiments. Together, we call our end-to-end modelling of off-targets, Elevation, which complements our on-target model, Azimuth. A schematic of our approach is shown in Fig. 1.

For the first task, of gRNA–target scoring, we developed a two-layer regression model (Elevation-score), in which the first layer learns to predict the off-target activity for single-mismatch (that is, between the target and the intended target, thus including alternative PAMs) gRNA–target pairs. The second-layer model learns how to combine predictions from the single-mismatch model for gRNA–target pairs with multiple mismatches into a single gRNA–target score—our 'combiner' model. For the combinatorial explosion of possible mismatch combinations, the amount of training data for the combiner model is extremely small. Consequently, we used a relatively simple model here. Note that insertions or deletions (indels) contribute to the off-target problem[29] but to a much lesser extent[3]; hence, we have focused our modelling efforts on mismatches.

第二步：汇总。

For the second task, of aggregating the individual target scores for a guide into a single numeric score, we first applied our gRNA–target scoring model to a list of potential targets (those short-listed by Elevation-search), and then use our model, Elevation-aggregate, to aggregate the individual scores. ~~The aggregation model takes into account whether each potential target lies in a gene, and allows these and other features to interact with each other by way of a non-linear modelling approach (boosted regression trees).~~ Details and intuitions for the development of the two-layer Elevation-score model and the Elevation-aggregate model are provided in the Methods.

Results

In this section, we first evaluate gRNA–target pair prediction models, including our Elevation-score. We demonstrate that Elevation-score yields state-of-the-art performance. In the next section, we evaluate Elevation-aggregate alongside the two competing summarization approaches—the MIT web server and CFD aggregation, where only the former has an accompanying web service that provides summary scores (CFD only provides within-gene rankings). Again, we find that our approach performs best, sometimes by an order of magnitude.

Individual gRNA–target pair off-target predictive modelling. We started by evaluating our Elevation-score approach using two independent data sets generated from genome-wide unbiased assays—one based on GUIDE-seq[3], and the other an aggregated data set curated by ref. [15] (we removed the GUIDE-seq data set from it as this is used as an independent data set). Elevation-score outperformed all other models—CFD[1], the current state-of-the-art, Hsu-Zhang[2] and CRISPR–Cas9 target online predictor (CCTop)[13]—in predicting off-target activity (Fig. 2). For a breakdown of performance by the number of mismatches, please see Supplementary Fig. 1.

Note that, for off-target prediction, it is generally more consequential to mistake an active off-target site for an inactive one, rather than the other way around, because only the first type of error can disrupt the cell or confound experimental interpretation, whereas the second may only require designing another gRNA. Consequently, we chose an evaluation measure that accounts for this asymmetry—the weighted Spearman correlation, in which each gRNA–target pair is weighted by an amount that is a (monotonic) function of its measured activity. Because the precise error asymmetry

脱靶的端对端建模

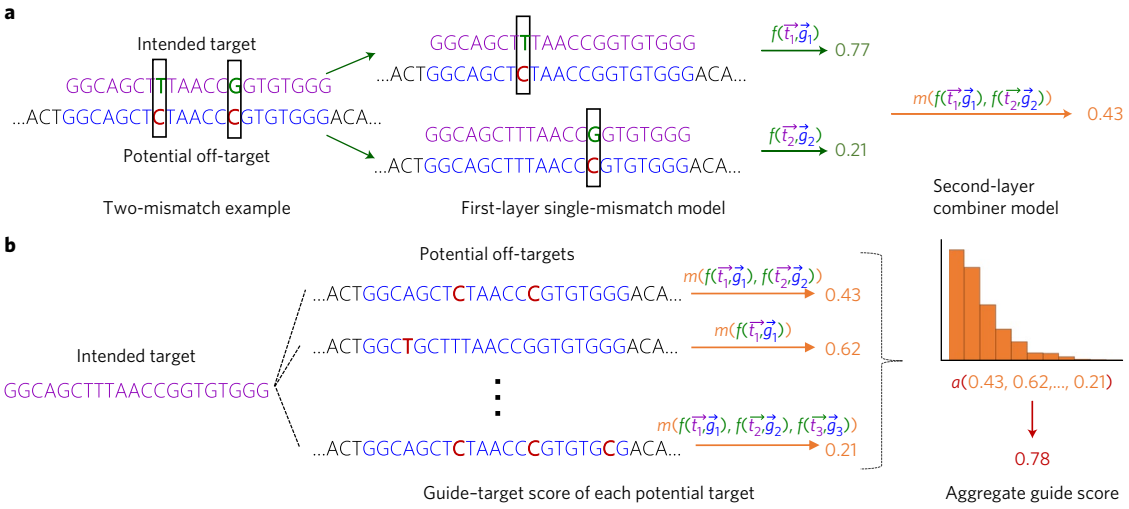

Fig. 1 | Schematic of Elevation off-target predictive modelling. a, An example of how to score a gRNA–target pair with two mismatches. First, the gRNA–target pair is broken down into two single-mismatch pseudo-pairs, ($\{t_1, g_1\}$, $\{t_2, g_2\}$), (where t_i and g_i, respectively, denote the target and gRNA in the ith pair), each of which is scored with the first-layer (single-mismatch) model, $f(\vec{t}_1, \vec{g}_1)$. Then, these scores are combined with the second-layer model, $m(f(\vec{t}_1, \vec{g}_1), f(\vec{t}_2, \vec{g}_2))$, yielding a single gRNA–target score that accounts for all mismatches. **b**, An example of how to aggregate the set of gRNA–target scores for a single gRNA into one summary off-target score for a gRNA. The aggregator model, $a()$, computes statistics of the input distribution of gRNA–target scores as features and runs them through a model, producing the aggregate score for a gRNA (for example, 0.78).

is not known a priori and may vary for different applications, we varied the weight continuously between two extremes: from being directly proportional to the measured activity (such that false negatives effectively do not count), to a uniform weighting (that is, yielding standard Spearman correlation).

For first-layer (single-mismatch) model features we used: (1) the position of the mismatch, (2) the nucleotide identities of the mismatch, (3) the joint position and identities of the mismatch in a single feature, and (4) whether the mutation was a transition or transversion. The relative importance of these features is shown in Fig. 3. It is interesting to note that using both the joint 'position and mismatch nucleotide identity' features—those effectively used by CFD—are aided by additionally decoupling these into additional features of position and nucleotide identity, even though regression trees can, in principle (with enough data), recover the joint features from the decoupled ones. Using only the CFD features in our model, or using classification instead of regression, or omitting the second layer of our model, each caused the model to perform worse

the nucleotide identity--核苷酸一致性

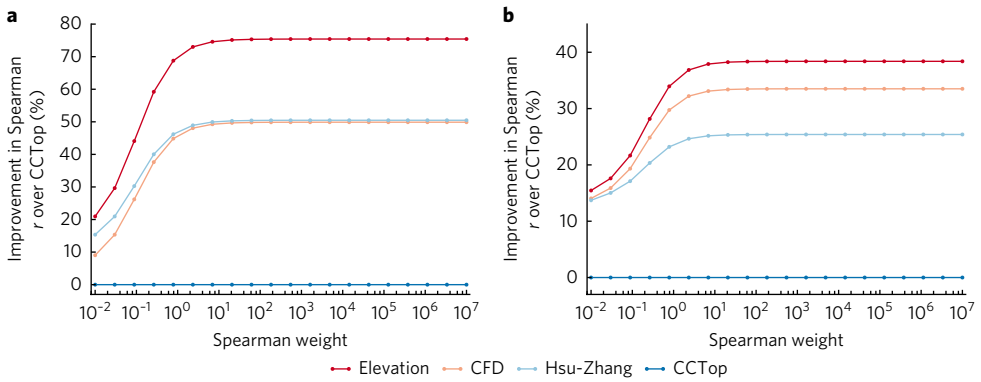

Fig. 2 | gRNA–target pair scoring. Comparison of Elevation-score performance to other methods, evaluated using a weighted Spearman correlation between predictions and assay measurements. The *x* axis shows different weights in the weighted Spearman; at the far left, the weight is effectively proportional to the rank-normalized GUIDE-seq counts/cutting frequency, whereas at the far right, the weight is effectively uniform, yielding a traditional Spearman correlation. For ease of visualization, the *y* axis denotes the per cent improvement of each model over CCTop, which, by design, thus lies constant at zero. **a**, CD33 ($N = 4{,}853$) and GUIDE-seq ($N = 294{,}534$) data were used to train, whereas data (after removing the GUIDE-seq data) from ref. [15] ($N = 10{,}129$) were used to test. **b**, The role of the GUIDE-seq and Haeussler data are reversed from **a**. The final Elevation-score model deployed in our cloud service uses the model trained on GUIDE-seq data. Note that, only 0.12% and 0.51% of count values in GUIDE-seq and Haeussler, respectively, are non-zero, making the traditional Spearman correlation difficult to interpret. However, for completeness, the right-most points correspond to a correlation of 0.117, 0.100, 0.101 and 0.007 for Elevation, CFD, Hsu-Zhang and CCTop, respectively, in **a**, and 0.059, 0.057, 0.053 and 0.043 in **b**. The *P* values computed for each Elevation correlation were less than floating point error (approximately 1×10^{-16}); these demonstrate that, despite the apparent low correlations, a tremendous amount of signal is present. Note that the apparent low correlations probably arise from the massive imbalance of inactive to active gRNAs.

(Supplementary Fig. 2). Feature importances for the second-layer (multiple-mismatch combiner) model show that the total number of mismatches and the sum of the first-layer single-mismatch predictions are driving the model (Supplementary Fig. 3).

Validation of Elevation-score. Finally, we performed two validation experiments of our final Elevation-score model by assessing its performance on two independent GUIDE-seq data sets—the first using previously published wild-type Cas9 experiments[30], here referred to as validation 1 ($N = 103{,}040$ guide–target pairs, of which 53 are active, arising from 5 single gRNAs), and the other newly generated experiments we performed, here referred to as validation 2 ($N = 381{,}249$ guide–target pairs, of which 57 are active, arising from 22 single gRNAs; Supplementary Table 1; Methods). On the whole, Elevation outperforms the other models, with one tie (Fig. 4). However, the performance ordering of CFD and Hsu-Zhang changes between the two experiments, so even though the performance of CFD ties with Elevation-search in one data set, it loses to both Hsu-Zhang and Elevation-search in the other data set, and hence is not performing consistently well. A breakdown by the number of mismatches is provided in Supplementary Fig. 4.

As a secondary measure of performance, one could consider how each model ranks only the active off-targets (that is, those detected by GUIDE-seq). Supplementary Tables 2, 3 and 4 show such results on the validation data. Elevation again outperforms the competing methods.

聚合的最后任务是要求获得对于一个gRNA所给定的所有单个gRNA-target评分的单个脱靶总结评分

Aggregating individual off-target scores into a single summary score. The final task, of aggregation, requires obtaining a single off-target summary score for a gRNA given all of its individual gRNA–target scores. A solution to this task is particularly useful for gRNA design, in which users want to scan numerous gRNAs for overall activity. To evaluate our approach on this task, we made use of two data sets with gRNAs targeting non-essential genes in viability screens—the Avana[1] and Gecko[31] libraries. Because each gRNA is designed to target one non-essential gene in these screens, the cell should be viable if no off-target effects are present. In particular, at least three papers have shown evidence that a cell is more likely to die when sustaining numerous DNA breaks[1,31,32]. In addition, a fourth paper leverages this phenomenon to assess off-target cutting[33].

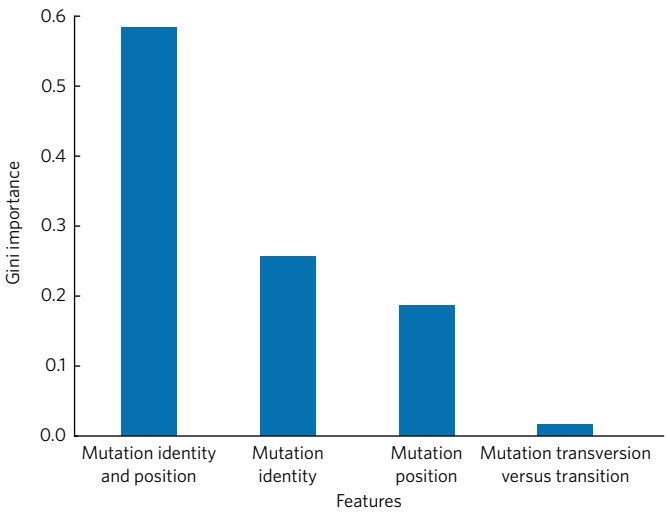

Fig. 3 | First-layer gRNA–target scoring feature importances. Average importances (Gini importances; see Methods) for the features in the first-layer single-mismatch model. This model was trained with CD33 single-mismatch data. Feature importances from the second-layer model are shown in Supplementary Table 2.

Thus, there is now substantial evidence that cell viability is determined at least in part by the number of DNA breaks per cell. A second effect on viability could be off-target activity at an essential gene. However, essential genes cover merely 0.2% of the human genome and are therefore not likely to have much effect in our experiments. To further elucidate this point, we evaluate the performance of our model using only gene essentiality as a feature, which performs vastly worse than when we either ignore it altogether, or additionally use the scores from our gRNA–target model as features (Supplementary Fig. 6). This empirically shows that gene essentiality is not adversely affecting our conclusions using the viability data. Hence, these viability-based experiments serve as bronze standard for the combined task of scoring and aggregation.

Using the viability data, we found that Elevation-aggregation was the best summary score model, yielding up to an eightfold improvement (and never performed worse) in weighted Spearman correlation over the best approach for this end-point task, namely, CFD aggregation. Elevation-aggregation yielded an even larger improvement over crispr.mit.edu[2], the most widely used but now no longer supported gRNA design tool (Fig. 5). The importance of each aggregation feature is shown in Fig. 6.

核染色质开放性--真核生物染色质DNA在核小体或转录因子等蛋白与其结合后，对其他蛋白能否再结合的开放程度，这一特征反应了染色质转录活跃程度

Predicting with chromatin accessibility. Several studies have suggested that chromatin accessibility may have a role in the activity of CRISPR–Cas9 (refs [19,35,36]). To investigate the effect of this feature on off-target activity, we were restricted to using the Gecko viability data, which was performed in 33 cell types, of which three (K562, PANC-1 and T-47D) had matching chromatin accessibility data (DNase I). The other data we had access to were performed in cell types that did not have chromatin accessibility data. Thus, we augmented our aggregation model to include DNase I features, independently for each of the three Gecko cell types. We included these DNase I features in several ways in the aggregation model (Supplementary Fig. 5; Methods). We found an increase in performance in PANC-1 for just one of the four models that included DNase I information. In the other two cell types, two different models that included DNase I information increased prediction over DNase I-agnostic models, but only over half of the weighted Spearman evaluation regime. Importantly, the same type of DNase I model was never consistently best. Next, we used the averaged DNase I data across all 95 available cell types instead of using cell-type-specific DNase I data. We found that this cell-type-averaged DNase I information did not increase the model performance in any of the cell types. Because so few of our data sets have matching chromatin data available at this time, and because of the inconclusive results, we decided to forgo including this information in our final deployed model for the time being. Users who are interested in augmenting our model with DNase I can use our source code to retrain such models, although we believe that it would be better to include it into the Elevation-score rather than Elevation-aggregate, even though we were not able to do so here.

Outlook

We have introduced a machine-learning-based approach to predictive modelling of off-target effects of the CRISPR–Cas9 system. Through systematic investigation, we demonstrated that our newly developed suite of models, Elevation, performs better for each of the two main off-target-related tasks in gRNA design: gRNA–target scoring and aggregation. In addition, we systematically evaluated available competing approaches on the task of summary scoring (aggregation), showing that Elevation consistently outperformed competing approaches by a substantial margin. We also considered how to balance errors between active and inactive gRNAs, developing a metric to do so, based on the weighted Spearman correlation. This type of evaluation encapsulates a range of practical use cases, and enabled us to show that Elevation is consistently

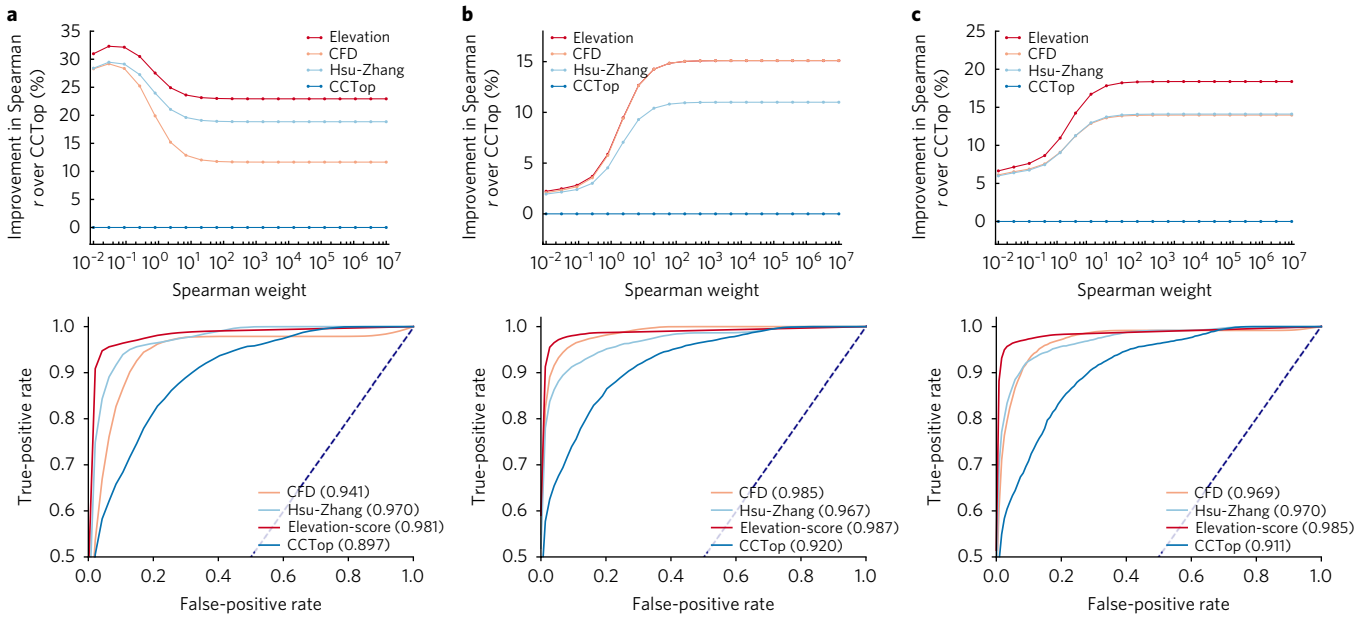

Fig. 4 | Validation of the Elevation gRNA–target scoring model. Performance of our final Elevation-score model on two independent validation data sets. **a**,**b**, 'Validation 1' ($N = 103{,}040$ guide–target pairs, of which 53 are active, arising from 5 single gRNAs (**a**) and 'validation 2' ($N = 381{,}249$ guide–target pairs, of which 57 are active, arising from 22 single gRNAs (**b**). **c**, Combined validation data sets ($N = 484{,}289$ guide–target pairs, of which 110 are active, arising from 27 single gRNAs). Although we believe our weighted Spearman correlation metric (top row) to be a particularly suitable evaluation metric, it is not necessarily intuitive to understand. Thus, we also included (bottom row) receiver operating characteristic (ROC) curve plots for classifier performance, such as in ref. [15], which used this for the same purpose. Note that random performance on the ROC is the dashed diagonal line and corresponds to area under the curve (AUC) = 0.50. Their corresponding AUC is written in the key (higher is better), as these are more intuitive. The ROC/AUC evaluation measure is suboptimal in that it only uses whether GUIDE-seq found activity, rather than how much (which our Spearman-based metric does make use of). However, one can see that the ROC evaluation roughly tracks our Spearman-based metric. (For ease of visualization, ROC curves and AUCs are averages of 100 random samples of inactive guides equal in number to the number of active guides in each data set.) Any true-positive rates that were missing for a given false-positive rate (owing to the sampling) were linearly interpolated from the two nearest neighbours within the appropriate curve.

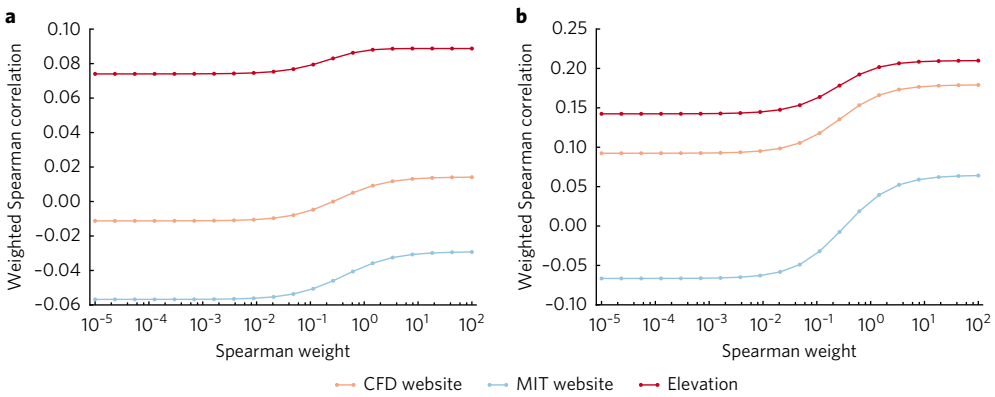

Fig. 5 | Joint scoring and aggregation on viability screens. Weighted Spearman correlation of Elevation to the crispr.mit.edu server. **a**, Avana data ($N = 4{,}950$) was used to train, and Gecko data was used to test ($N = 4{,}697$). **b**, The reverse of **a**. Note that the MIT website often yields correlation in the wrong direction. The final Elevation model deployed in our cloud service uses the model trained on Avana.

superior across the entire range. We recommend that the community use such metrics in the future when comparing new and existing models for off-target modelling.

As data become available for a richer set of scenarios, including different endonucleases, different organisms, in vitro and in vivo settings, and epigenetics on more cell types, models and tools should be updated accordingly.

Elevation-score and Elevation-aggregate, which together we call Elevation, complement our on-target predictive model, Azimuth[1]. Together, Azimuth and Elevation, along with our cloud service and web front-end (https://crispr.ml), provide an integrated end-to-end guide design tool that enables users to more effectively deploy CRISPR–Cas9 for research screening experiments and that may provide a useful pre-screening tool for identifying potential gRNAs for therapeutic applications—one based on the state-of-the-art machine-learning-based methods. To make as an efficient a service as possible, the back-end of our web portal contains pre-computed on-target and off-target activity prediction for every genic region

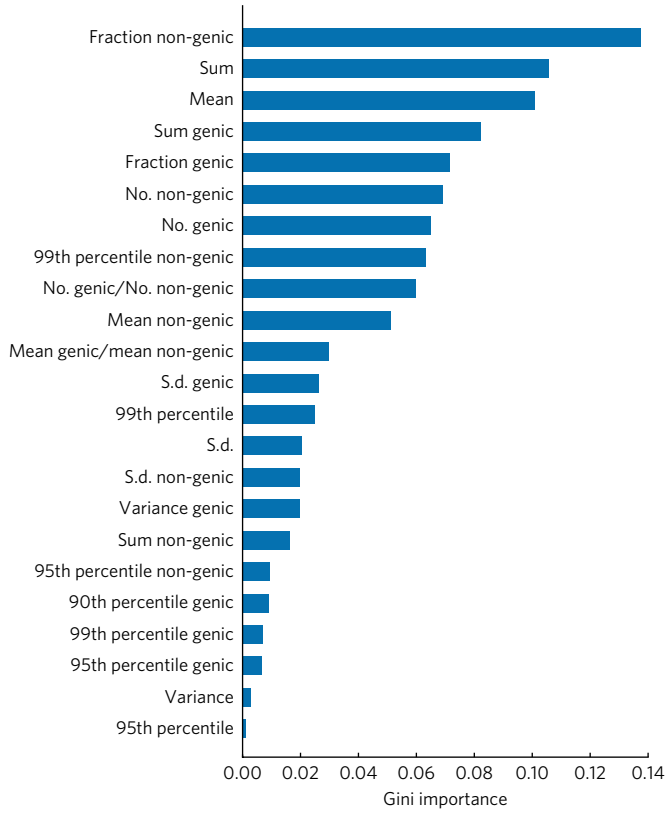

Fig. 6 | Aggregator feature importances. Weights from the aggregator model in Elevation, which uses gradient boosted regression trees. The features were: the mean, median, variance, standard deviation (s.d.), 99th, 95th and 90th percentiles, and the sum of the Elevation gRNA–target scores for each gRNA. We compute these for each of: all off-targets (no post-fix), only genic off-targets ('genic') and only non-genic targets ('non-genic'), where is-genic is obtained from ENSEMBL[34]. In addition, we compute these further features: fraction of targets that are genic, fraction of targets that are non-genic, ratio of the number of genic to non-genic targets, and ratio of the mean genic to non-genic score. The Gini importance is described in Methods.

in the human genome. In future work, we will also more carefully investigate the issue of search and filter.

Methods

Data. To train our first-layer, single-mismatch model, we used CD33 data from ref. [1] where all single-mismatch mutations (between the intended target and the potential off-target, thus including alternate PAMs) were introduced into the target DNA for 65 perfect-match gRNAs that were effective at knockout. CD33-negative cells were isolated by flow cytometry so that their log-fold-change (LFC) prior to CRISPR–Cas9 introduction could be measured by sequencing. After filtering as in ref. [1], we retained 3,826 single-mismatch observations and 1,027 alternate PAM observations for a total of 4,853 gRNA–target training examples, of which 2,273 were considered active by ref. [1]. These data measure protein knockout efficiency rather than DNA cleavage. We refer to these data as the CD33 data.

To evaluate our second-layer, multiple-mismatch model, we used two unbiased/genome-wide multiple-mismatch data sets. The first were already-published GUIDE-seq data[3] comprising nine gRNAs that were assessed for off-target cleavage activity. These gRNAs yielded 354 active off-target sites (that is, non-zero counts) with up to six mismatches. Non-active sites were obtained from ref. [1] who used Cas-OFFinder[14] to identify all 294,534 sites with six or fewer mismatches. The second data comprised off-target data aggregated by ref. [15], after removing GUIDE-seq data to make it independent from the previously mentioned data set. These data consisted of 52 active targets among 10,129 non-active potential targets. We set 0 counts in GUIDE-seq data to 0.001, which is the estimated sensitivity of the assay as determined and also done in ref. [15]. Finally, for both the GUIDE-seq and ref. [15] measured activity, we linearly re-scaled the values to lie in [0,1] before applying a Box–Cox transform[37].

For our validation of the gRNA–target model, Elevation-score (trained on the CD33 and GUIDE-seq data), we applied it to two previously unseen data sets that were assayed with GUIDE-seq: (1) five gRNAs yielding a total of 103,040 potential off-targets, of which 53 are active, from ref. [30], and (2) 22 unique gRNAs in a newly generated data set, yielding a total of 381,249 potential off-targets, of which 57 are active (see Supplementary Table 1). The list of potential off-target sites was obtained using Elevation-search, as described below. The gRNAs in the newly generated data set were chosen in a manner that was unbiased with respect to favouring any of the predictive models. In particular, we used CFD, Hsu-Zhang, CCTop and Elevation-score to make off-target predictions for gRNAs in the Gecko library[31] (which were not used to train any of those models), excluding any gRNAs that yielded non-viable cells (that is, assay read-out of less than −1.0). We then converted the predictions within each method to ranks, to make the predictions comparable in scale across methods, and then averaged the ranks across methods to obtain one estimated activity for each gRNA, which was model-agnostic. This yielded an ordering of gRNAs from expected most to least active (an ordering that was not biased to any one method). From that list, we then chose 10 consecutive gRNAs, each starting at the top 10%, 20% and 30% of overall activity. For the 20% set, one gRNA had two perfectly matched sites in the genome, so instead, we used the next gRNA on the list. Only gRNAs assayed with wild-type Cas9 were used from ref. [30]; these gRNAs had been selected without any predictive modelling of off-target effects, and hence were unbiased with respect to the methods being compared herein. In particular, G-N19-NGG sites in a few of the commonly assayed genomic amplicons/genes had been selected[3,38].

To evaluate the aggregation of off-target effects, we used two data sets arising from gRNAs targeting non-essential genes in a viability screen. The first, from the Avana library[1], used 4,950 gRNAs targeting 880 non-essential genes. The second, from the Gecko library[31], used 4,697 gRNAs targeting 837 non-essential genes. Other than for the DNase I experiments, we used cell type A375 from Gecko.

DNase I peak file data for 95 human cell types, measuring chromatin accessibility, were downloaded from the http://genome.ucsc.edu/cgi-bin/hgTables?hgsid=581299277_DBUyFx88KdBssISoFyqBLBdKNq2M&clade=mammal&org=Human&db=hg38&hgta_group=regulation&hgta_track=wgEncodeRegDnase&hgta_table=0&hgta_regionType=genome&position=chr9%3A133252000-133280861&hgta_outputType=primaryTable&hgta_outFileName on 22 March 2017.

GUIDE-seq. U2OS cells (ATCC) were cultured at 37 °C with 5% CO_2 in advanced DMEM supplemented with 10% heat-inactivated fetal bovine serum, 2 mM GlutaMax, and penicillin/streptomycin (all cell culture reagents from Thermo Fisher Scientific). Cell line identity was validated by short tandem repeat (STR) profiling (ATCC) and routine mycoplasma testing was negative for contamination. GUIDE-seq experiments were performed with 22 unique single gRNAs (and the EMX1 site 1 single gRNA as a control) as previously described[3]. Briefly, roughly 2×10^5 human U2OS cells were transfected (SE kit and DN-100 program on a 4D nucleofector; Lonza) with 750 ng nuclease plasmid, 250 ng gRNA RNA plasmid, and 100 pmol end-protected double-stranded oligo (dsODN) GUIDE-seq tag. Approximately 72 h following nucleofection, genomic DNA was extracted via Agencourt DNAdvance Genomic DNA Isolation (Beckman Coulter). Gene disruption and GUIDE-seq tag-integration efficiencies were evaluated using T7E1 and RFLP assays, respectively, as previously described[30]. GUIDE-seq sample libraries (prepared as previously described[3]) were sequenced on an Illumina MiSeq sequencer, and data were analysed using an updated version 1.1 of the open-source GUIDE-seq software[39]. All data related to GUIDE-seq experiments can be found in Supplementary Table 1. New GUIDE-seq data generated for this study have been deposited in the NCBI Sequence Read Archive (SRA) under accession number SRP117146.

Predictive modelling for scoring individual gRNA–target pairs. Here, we describe the CFD model[1], what assumptions it makes, and then describe our model, Elevation-score, and how it relates conceptually to CFD. The predictive off-target model, CFD, first computes the observed frequency of gRNA–target pair activity for each single-mismatch type in the CD33 data. CFD then combines these single-mismatch frequencies by multiplying them together for gRNA–target pairs with multiple mismatches. For example, if a gRNA–target pair had a A:G mismatch in position 3, a T:C mismatch in position 5 and a PAM of 'CG' in the target region, then CFD would take the off-target score of this gRNA to be $CFD\ score = \mathrm{P}(active\,|\mathrm{A}:G,3)\times\mathrm{P}(active\,|\mathrm{T}:C,5)\times\mathrm{P}(active\,|CG)$, where each of these terms are computed from observed frequencies in the CD33 training data (which contained only single mismatch, or alternate PAMs, but never both).

~~*CFD as naive Bayes.*~~ One can interpret the CFD algorithm in terms of a known classification model called naive Bayes[40] as follows. First, denote $Y=1$ to mean a gRNA–target pair was active, and $Y=0$ to denote that the pair was not. Next, denote features such as T:C,5 as X_i, where i simply indexes some enumeration of these

观测频率

朴素贝叶斯

1、Y=1表示有活性，Y=0表示无活性。

features (that is, a one-hot encoding). If that feature (mismatch) occurred, then $X_i = 1$, and if it did not occur then $X_i = 0$. Thus, in the CD33 data (with only single mismatches), a particular gRNA–target pair has only one $X_i = 1$ and all others have $X_i = 0$. In this notation, one can re-write CFD as follows for one gRNA–target pair:

贝叶斯模型将会计算在给定特征值下，一个gRNA-target对是有活性的概率。

$$CFD \equiv \prod_{i \in \{i | X_i = 1\}} P(Y = 1 | X_i = 1).$$

By contrast, a naive Bayes model would compute the probability that a gRNA–target pair is active given the feature values as:

$$Naive\ Bayes \equiv P\left(Y = 1 \middle| \{X_j\}\right) = \frac{P(Y=1)}{P\left(\{X_j\}\right)} \prod_i P(X_i | Y = 1)$$

where X_j is the set of all features X_i. Naive Bayes makes only one assumption: conditioned on a gRNA being active, the features X_i are independent so that $P\left(\{X_j\} \middle| Y = 1\right) = \prod_i P(X_i | Y = 1)$. Using Bayes' rule, one can re-write the naive Bayes classifier as:

$$Naive\ Bayes \equiv P\left(Y = 1 \middle| \{X_j\}\right) = \frac{P(Y=1)}{P\left(\{X_j\}\right)} \prod_i P(Y = 1 | X_i) \frac{P(X_i)}{P(Y=1)}$$

$$= \frac{1}{P\left(\{X_j\}\right)} \prod_i P(Y = 1 | X_i) P(X_i).$$

If we make two further assumptions, we find that the naive Bayes classifier exactly matches CFD. The first assumption is that the features are marginally independent, namely, that $\prod_i P(X_i) = P\left(\{X_j\}\right)$, in which case naive Bayes simplifies to:

$$Naive\ Bayes_{\text{feat.ind.}} = \prod_i P(Y = 1 | X_i).$$

The CFD assumption of marginal feature independence seems reasonable and yielded good results. Consequently, we make the same assumption in our Elevation-score model. If we also assume that $P(Y = 1 | X_i = 0) = 1$, then CFD and naive Bayes become identical. This second CFD assumption ($P(Y = 1 | X_i = 0) = 1$) seems a more difficult one to accept, but with some careful thought (and the fact that CFD performs so well), also seems reasonable as we explain next; hence, we also make this assumption. The key insight is to ask which properties of the training data one expects to generalize to unseen data sets where the model might be applied. In particular, it seems reasonable to assume that $P(Y = 1 | X_i = 1)$ is a quantity that will generalize to other data sets; intuitively, the quantity reflects how probable a gRNA is to be active given that we observed a particular kind of mismatch—as such, it is independent of the distribution of the types of mismatch in the training versus test data sets. By contrast, $P(Y = 1 | X_i = 0)$ defines how probable a gRNA is to be active given that we did not observe a feature. When computing this quantity, one marginalizes (averages) over all examples in the CD33 data set where $X_i = 0$, which includes all gRNA–target pairs for which $X_{j \neq i} = 1$; as such, this probability specifically depends on the distribution of mismatch types in the off-target data set and their corresponding activities. Thus, we do not necessarily expect these quantities, $P(Y = 1 | X_i = 0)$ to generalize from our training data (CD33 specific) to general test sets. ~~Now the question remains, how can we therefore make a reasonable approximation?~~ One could try to posit a canonical theoretical or actual data set that will best generalize; however, it is extremely difficult to come up with such a set. Furthermore, in light of how we are going to use our naive Bayes probabilities (described next), getting it exactly right is not critical. Hence, we make the CFD assumption that $P(Y = 1 | X_i = 0) = 1$.

We have now shown that, with two assumptions, the CFD model can be interpreted as a naive Bayes classifier. The reason for making this connection is not only to put the CFD in a proper probabilistic framework, including its assumptions, but more importantly, to then generalize this model so as to improve its performance, which we now do in describing our Elevation-score model.

Elevation-score as two-layer stacked regression. We generalized away from CFD in three main ways: (1) instead of classification we use regression, (2) we augment the feature space, and (3) we replace the a priori manner of combining probabilities by multiplication to combining using machine learning. We call the model implementing only the first two, Elevation-naive (Supplementary Fig. 1), whereas we refer to the model resulting from all three as Elevation-score, or final model class. We now explain these in more detail.

The first observation in generalizing away from CFD is that it is a classification algorithm, which means it discards the real-valued assay measurements, converting them to be binary active/inactive. Thus, by design, CFD is unable to capture the more nuanced information available in the data. In moving from classification to regression, the model has access to more fine-grained information. Although not widely used, there exist generalizations of naive Bayes from classification to regression[41]; however, owing to the specifics of our problem, they are not convenient to apply. Thus, we developed our own approach in which we first convert the CD33 LFC values to lie in the range [0,1] so that they can, loosely speaking, be interpreted as probabilities. To do so, we used a kernel density estimator to transform each LFC to the cumulative density of that LFC in the kernel density estimate. We used a ~~Gaussian kernel~~ and chose the bandwidth by ~~tenfold cross-validation~~ (yielding a bandwidth of 0.23).

Recall that Elevation-score is a two-layer stacked regression model, in which the first layer makes predictions for gRNA–target pairs with only a single mismatch, whereas the second layer combines these for gRNA–target pairs with multiple mismatches (in contrast to CFD, which simply multiplies them together).

To learn each first-layer (single-mismatch) regression model, $p\left(y \middle| \{X_j\}\right)$, we used ~~boosted regression trees~~ (using default settings in scikit-learn) on the CD33 data[42]. As each gRNA–target pair has only a single $X_i = 1$ in these data, we could have also just used a ~~linear regression model~~. However, we wanted to include a richer featurization of the gRNA–target pair than just features of the form A:G, 5, resulting in the fact that, even for single-mismatch data, more than one $X_i = 1$ could occur (also some features were real-valued rather than 0/1). Furthermore, we wanted these ~~features to be able to interact in a non-linear manner~~. In particular, in addition to the CFD features, we also used '~~decoupled~~' versions of them—one of the form 'A:G' encoding the mismatch nucleotide types, which was one-hot encoded (described at the end of this section) and the other ~~an integer feature for the position~~ (for example, 5). Note that CFD uses these together as a single feature. ~~We also included whether the mutation was a transversion or a transition. We call the model that uses these improvements and combines each mismatch just as CFD does, by multiplying the values together,~~ Elevation-naive. As can be seen in Supplementary Fig. 1, moving from classification to regression improved the performance of the off-target model, as did augmenting the features. Next, we describe how we improved Elevation-naive to obtain Elevation-score.

boosted 回归树

特征

Although Elevation-naive improved on CFD, there were several aspects of the modelling approach that suggested areas for further improvement. The first was that the naive Bayes assumption of class-conditional independence may not be fully justified. The second is that our regression model's predicted values are not calibrated probabilities of gRNA–target activity; hence, when we combine them under that assumption (as does naive Bayes and CFD), we may suffer in performance. Thus, it stands to reason that if we could somehow loosen these assumptions, we might achieve better performance still. One way to do this is to augment the model, here with a second layer, and then to use the limited amount of multiple-mismatch/PAM gRNA–target pair data to learn the newly added parameters. We refer to this second layer of Elevation-score as the combiner because it learns how to combine the predictions from the single-mismatch model in a more nuanced way than simply multiplying them together, thereby allowing some of the stated assumptions to be mitigated. Thus, where a CFD/naive Bayes approach would simply multiply single-mismatch probabilities together, we instead use a data-driven machine-learning approach to fine-tune how they should be combined. In particular, we first use our first-layer boosted regression trees model J times to make predictions for each of the J single mismatches for a gRNA–target pair (that is, J features for which $X_j = 1$), yielding J predictions $\hat{y}_j \in [0, 1]$ (one for each feature with $X_j = 1$, and setting $\hat{y}_k = 1$ for the remaining K features that have $X_k = 0$). Thus, each gRNA–target pair has $T = J + K = 21$ boosted regression tree predictions $\{\hat{y}_t\}$ (20 for each possible mismatch position and one for an alternate PAM). The log of these 21 features ($\{\log(\hat{y}_t)\}$), along with their sum, their product and J—the number of mismatches/alternate PAM—are then the input to an L1-regularized linear regression combiner model—the second-layer model. We used both the GUIDE-seq data and data from ref. [15], each in turn, to train a combiner model. Each time we tested on the other (non-training) data set, using tenfold cross-validation to set the L1 penalty. Note that, Elevation-score's two-layer model is inherently different from both a two-layer neural network[43] and from stacked generalization[44] because the data used to train each Elevation-score layer are different (single versus multiple mismatch). Furthermore, note that, owing to the tiny proportion of non-zero values in these data (for example, 0.5%), we subsampled the zero activity examples to match the number of non-zero values within each data set, only for training, finding this to be helpful.

不是校准概率

数据驱动机器学习方法

Finally, because what we ultimately want are predictions of the probability that a gRNA–target pair is active, we also applied one final transformation to the output from the L1-regression combiner model. Namely, we put the output of that model through a calibration model. This calibration model estimates $P(\text{active} | \text{GUIDE} - \text{seq normalized counts})$ using a logistic regression model trained on predictions for the CD33 data using Elevation-naive as inputs (this model makes prediction in GUIDE-seq normalized count space), and using the corresponding CD33 binarized observed activities[1] (LFC > 1) as the target variable. Note that this transformation is monotonic and, as such, only affects the performance of aggregation, not gRNA–target scoring, given that we use a (weighted) Spearman rank correlation. For the aggregation task, the Spearman correlation is computed only after aggregation of scores for a gRNA; thus, any change in scale of the pre-aggregated scores, even if monotonic (such as our calibration model), can dramatically influence the quality of the final aggregation. In other words, ~~even a simple linear transformation could change the aggregation scores.~~

二值观测活性---目标变量

One-hot encoding of categorical variables. A 'one-hot' encoding refers to taking a single categorical variable and converting it to more variables, each of which can take on the value 0 or 1, with at most one of them being 'hot', or on. For example, with a categorical nucleotide feature, which can take on values A/C/T/G, each letter would get converted into a vector of length four, with only one entry 'hot' (equal to 1 instead of 0), corresponding to one of the four letters.

Elevation分数值只提供了选择具有最小期望脱靶活性的gRNA的初始条件。

Elevation-aggregate. Elevation-score provides only the starting ingredients for choosing a gRNA with the least expected off-target activity. To actually rank gRNAs, one may want to summarize the scores from all gRNA–target pairs for a given gRNA into a single number so that gRNAs can be ranked by this number for off-target activity. Thus, we developed Elevation-aggregate, a model based on ~~gradient boosted regression trees~~, to perform this task. Hyper-parameter settings were chosen by ~~cross-validation~~ using a ~~random search~~ over these parameters and ranges: losses ∈ {least squares; least absolute deviation; Huber}, learning rates ∈ [1.0×10^{-6}, 1.0] equally spaced in log-space with 100 points, the number of estimators ∈ [20, 50, 80, 100, 200, 300, 400, 500], maximum tree depth from 1 to 7, the minimum number of samples to split = [2, 3, 4], splitting criterion ∈ [Friedman mean-squared error, mean-squared error, mean absolute error]. We evaluated 10 randomly chosen samples from these sets. For Fig. 5 and Supplementary Fig. 1, we performed fivefold cross-validation on the training data to select the best model before using it to measure performance on the test set. For the DNase I experiments, we were only able to use Gecko data owing to cell-type compatibility and therefore had to evaluate Gecko itself using cross-validation. Thus, in this setting, we used 20-fold cross-validation to evaluate a model (for example, *dnase1*), where within each fold, we performed an inner 5-fold cross-validation to select the best hyper-parameter setting. The input features used for the model were computed from the distribution of gRNA–target Elevation-score predictions and comprised: the mean, median, variance, standard deviation, 99th, 95th and 90th percentiles, and the sum of off-target scores. We compute these for each of: all off-targets, only genic off-targets and only non-genic off-targets. In addition, we computed these further features: the sum of genic (non-genic) off-targets divided by the total number of off-targets, the fraction of targets that were genic, the fraction of targets that were non-genic, the ratio of the number of genic to non-genic targets, and the ratio of the average genic to non-genic score. The final deployed model was trained only on the Avana data, as combining it with Gecko did not increase cross-validation performance.

将多个值聚合成一个以进行排序。

将DNase1峰值数据合并到聚合模型中。

Incorporation of chromatin accessibility features. To incorporate DNase I peak data into the aggregation model (the only model for which we had training data with corresponding DNase I data), we tried four different approaches: (1) using the DNase I as a 'mask' on the values output from the gRNA–target scorer (that is, taking the element-wise product between the original features and their corresponding DNase I peak values); (2) adding to our original features, statistics of the DNase I features (same summary statistics as with our original features, but computed only on the DNase I features) as independent features; (3) using our original aggregation features in addition to those newly added features in (1); and (4) using our original aggregation features in addition to those newly added features in (1) and (2). The track files used in this experiment were downloaded from UCSC[46]. The data's Gene Expression Omnibus (GEO) accession numbers are GSM736629 and GSM736566 for K562, GSM736517 and GSM736519 for PANC-1, and GSM1024761 and GSM1024762 for T-47D.

Comparison to other approaches. We compared models using the Spearman correlation between predicted and measured off-target activity. Furthermore, as discussed in the main text, we also evaluated the weighted Spearman correlation for various weight settings, to account for an asymmetrical loss with respect to false-positive active errors as compared to false-negative active errors. Specifically, we set the weights as follows. Let $\{g_i\}$ be the values of the normalized GUIDE-seq or Haeussler data (all lying in [0,1]). Then, we set the weight for each data point to be $w_i = \frac{g_i + v}{\max_j g_j + v}$ (such that $w_i \in [0, 1]$) where v is varied through 10^{-5} to 100 and is denoted on the x axis of the relevant figures. When $v = 10^{-5}$, the weights are effectively equal to the GUIDE-seq/Haeussler measured values (this is the left-hand side of our plots), meaning that data with zero counts effectively almost vanish from the computation. When $v = 100$ (right-hand side of our plots), the weights become effectively identical for all gRNAs, yielding a standard Spearman correlation. For each v, we assessed the effective sample size (the sum of the weights) and only considered weights that produced effective sample sizes of ≥50, to remove high-variance estimates, which could be misleading.

For implementation of CFD, we used Supplementary Table 19 from ref. [1]. We re-implemented CCTop based on the description in their paper. For Hsu-Zhang gRNA–target pair scoring, we re-implemented the approach based on the equation in their paper.

To compare Elevation to the aggregation scores of the MIT web server, each gRNA sequence was submitted to the MIT CRISPR Design Tool using their RESTful API provided for single sequences (http://crispr.mit.edu/). Every sequence was queried using sequence type 'other region (23–50 nucleotides)' and target genome 'human (hg19)' to obtain an off-target score. The server failed to produce scores for the following three sequences, which we therefore removed from consideration in our comparison: sequence TGACCTGTGACCATGATCACCACAGGGTTG from Avana and sequences CAAGCCTGTGTGCTGCAAGCCTGTCTGCTCTGTGCC and TCTCTGGCCATCATTTCCTGGGAGAGATGGATGGTG from Gecko. All queries were submitted and their results processed between 15 and 29 August 2016, inclusive. No software version number was found in the output or web page.

To compare Elevation to the CFD server (http://portals.broadinstitute.org/gpp/public/analysis-tools/sgrna-design), each gene corresponding to an Avana or Gecko gRNA was submitted between 21 and 23 September 2016, inclusive, and the relevant gRNA rows retrieved. A score for a gRNA was obtained by adding the values in the two fields 'Tier I Match Bin I Matches' and 'Tier I Match Bin II Matches' as done by ref. [1]. Although the server returns a field off-target rank, this field cannot be readily compared across gRNAs, as it is within-gene only.

Elevation-search. To perform efficient genomic searches for potential off-targets, we developed the program Elevation-search (also known as dsNickFury) which uses seed and extension[45] (using two tandem seeds) to find near-match CRISPR–Cas9 targets. In brief, Elevation-search can leverage distributed computing to efficiently catalogue every potential CRISPR–Cas9 target in a genome for any CRISPR–Cas9 system with targets that can be abstracted into some maximum length of RNA gRNA followed by a set of potential PAM sequences of fixed length. These potential targets are then ~~organized into a tree data structure based on two tandem seed sequences~~ (lengths of 8 and 6 nucleotides were used as the first and second seeds, respectively, but this is a user-specified parameter that affects performance and not results) taken from the gRNA sequences that were immediately proximal to the PAM site. ~~The first branch layer of the tree structure comprised all observed first tandem seeds~~ (most proximal to the PAM), whereas ~~the second layer contains branches for each second tandem seed~~. Each first and second seed combination links to a file containing all of the potential CRISPR–Cas9 targets in the genome that have that specific combination of tandem seeds proximal to the PAM site.

Potential off-target matches are defined by a maximum number of mismatches relative to the intended target (set by the user), with a certain number of bases distal to the PAM being ignored if desired. For experiments herein, mismatch tolerance was set to three, with the three most distal bases from the PAM being ignored. All PAMs that were deemed to have non-zero activity in ref. [1] are considered (NAG, NCG, NGA, NGC, NGG, NGT and NTG). This strategy was based on previous observations that much of the CRISPR–Cas9 off-target activity risk is determined by the number of mismatches between on-target and off-target sequences, with bases that are more distal to the PAM sequence being more mismatch tolerant and contributing less to specificity[3]. Potential sites were searched initially based on their tandem seeds, using a depth-first search of the cached tree structure. Any leaves with fewer mismatches than the maximum allowed had the same check then applied to their extended sequences, ignoring bases that were distal to the PAM, as determined by a user-specified parameter. Those sequences that passed the filters were considered as potential off-targets and were scored by Elevation. Note that the resulting search is not an approximate. For example, if the search parameters define a search where up to three mismatches are tolerated, then all such sites will be returned. We also used the Ensembl database to determine whether each off-target was in an annotated gene, such that users can obtain an aggregated off-target score across one, the other, or both. The resulting sites can be sorted based on their mismatch counts and/or Elevation score, and in general, can be reported directly to the user by way of a file or by deposition into a NoSQL database, as we have done to populate the back-end of https://crispr.ml.

深度优先搜索

Because most sites can be disqualified based on their seeds without loading the extended sequence and have already been annotated by both sequence and locus, searches can be conducted using substantially fewer resources than an alignment-based search. This allows for many searches to be conducted in parallel on a distributed computing environment. For results reported herein, we pre-computed and stored all human genome-wide results for both on-target and off-target predicted activities in a cloud-based database that we make available to the community.

Our system is designed to function on several different CRISPR–Cas9 systems with PAM sites at the 3′ end of the target. Parameters may be set for different lengths of gRNA sequence, PAM sequences with higher activity and species of origin for the reference genome. Potential targets can be ranked for on-target efficiency and off-target risk. The system is currently using Azimuth for on-target activity prediction and Elevation for off-target activity prediction for the *S. pyogenes* CRISPR–Cas9 system.

A summary of the search parameters used for all experiments in this paper and the online cloud service are as follows: we included all off-targets in the genome with no more than 3 mismatches in the 4–20 of the gRNA, with any number of mismatches in the first 3 gRNA nucleotides (1–3), and considering any PAM deemed to have non-zero probability according to the CFD model (namely, NAG, NCG, NGA, NGC, NGG, NGT and NTG)[1]. We stopped any searches that yielded >40,000 potential off-targets according to these criteria. For those searches yielding >40,000 potential off-targets, we set our Elevation gRNA potential (that is, the final

gRNA aggregate value) to be equal to 1,000. Furthermore, note that, although we searched over a wide range of PAMs, when evaluating models, we used only NRG, to be consistent with the work in ref. [1] that used this filter to try to accommodate CCTop, which could only evaluate NRG PAMs.

Gini importance as feature importance in regression trees. The Gini importance refers to the decrease in the mean-squared error (the criterion used to train each regression tree) when that feature is introduced as a node in the tree. This measure has a close, empirical correspondence with the importance of the feature that would be obtained with a permutation test, and can also be viewed as a relative decrease in entropy provided by splitting of that feature. This measure of importance does not convey whether having that feature makes a gRNA better or worse in the model because such a notion is impossible for regression trees in which the effect of one feature is dependent on the presence/absence of other features (that is, there are non-linear interactions between the features).

Model interpretability. Our primary goal in this paper was to provide a state-of-the-art tool that the community could use, alongside evidence of it outperforming alternative tools. Although any biological insights that can be derived from such analyses are extremely interesting, we caution that the more powerful (and hence complex) a model is, the less interpretable it is. That is not to say that people do not try to assign interpretations to complex models, but these interpretations are by definition not ideal summaries, and should not be over-interpreted. To give some intuition into why, consider first a linear regression model. In a linear regression model, each 'feature' (for example, in our context might be 'A in position 2') contributes to a final regression prediction in a linear, additive manner. Thus, it is fairly trivial to assign some importance (and directionality) to each feature independently. By contrast, with a more complex model, which allows for interactions, the importance of each feature depends entirely on what the value for other features are, and cannot easily be interpreted on its own. In addition, because of this complexity, it is often the case that a near infinity of models, each with slight perturbations in effective ranking of features, could all achieve the same predictive performance. Thus, although we have provided these feature rankings, we have not focused too much on them because we do not want to encourage their over-interpretation.

Web portal. We pre-computed all on-target and off-target scores for the human exome (GRCh38) and made them available at https://crispr.ml. The on-target scores were computed using Azimuth[1]. Aggregated off-target values were computed using Elevation-aggregation. To further drill down in the specific off-targets for a given guide, we also list all the individual gRNA off-target scores (that is, not aggregated), computed using Elevation-score. Please check the tool tips carefully to see whether a higher score means more active or less active, as there may be an inversion relative to scores stated in the paper.

Life Sciences Reporting Summary. Further information on experimental design is available in the Life Sciences Reporting Summary.

Code availability. All source code and a front-end website for the cloud service can be found from links at https://www.microsoft.com/en-us/research/project/crispr.

Data availability. The authors declare that all data supporting the findings of this study are available within the paper and its Supplementary Information. The new GUIDE-seq data generated for this study have been deposited with the NCBI SRA under accession number SRP117146. All other data used are publicly available, as noted in Methods.

Received: 3 December 2016; Accepted: 23 November 2017;
Published online: 10 January 2018

References

1. 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).
2. Hsu, P. D. et al. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* **31**, 827–832 (2013).
3. Tsai, S. Q. et al. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR–Cas nucleases. *Nat. Biotechnol.* **33**, 187–197 (2015).
4. Frock, R. L. et al. Genome-wide detection of DNA double-stranded breaks induced by engineered nucleases. *Nat. Biotechnol.* **33**, 179–186 (2015).
5. Wang, X. et al. Unbiased detection of off-target cleavage by CRISPR–Cas9 and TALENs using integrase-defective lentiviral vectors. *Nat. Biotechnol.* **33**, 175–178 (2015).
6. Kim, D. et al. Digenome-seq: genome-wide profiling of CRISPR–Cas9 off-target effects in human cells. *Nat. Methods* **12**, 237–243 (2015).
7. Kim, D., Kim, S., Kim, S., Park, J. & Kim, J.-S. Genome-wide target specificities of CRISPR–Cas9 nucleases revealed by multiplex Digenome-seq. *Genome Res.* **26**, 406–415 (2016).
8. Tsai, S. Q. et al. CIRCLE-seq: a highly sensitive in vitro screen for genome-wide CRISPR–Cas9 nuclease off-targets. *Nat. Methods* **14**, 607–614 (2017).
9. Cameron, P. et al. Mapping the genomic landscape of CRISPR–Cas9 cleavage. *Nat. Methods* **14**, 600–606 (2017).
10. Ran, F. A. et al. In vivo genome editing using *Staphylococcus aureus* Cas9. *Nature* **520**, 186–191 (2015).
11. Yan, W. X. et al. BLISS is a versatile and quantitative method for genome-wide profiling of DNA double-strand breaks. *Nat. Commun.* **8**, 15058 (2017).
12. Crosetto, N. et al. Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing. *Nat. Methods* **10**, 361–365 (2013).
13. Stemmer, M., Thumberger, T., del Sol Keyer, M., Wittbrodt, J. & Mateo, J. L. CCTop: an intuitive, flexible and reliable CRISPR/Cas9 target prediction tool. *PLoS ONE* **10**, e0124633 (2015).
14. Bae, S., Park, J. & Kim, J. S. Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. *Bioinformatics* **30**, 1473–1475 (2014).
15. Haeussler, M. et al. Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biol.* **17**, 148 (2016).
16. Labun, K., Montague, T. G., Gagnon, J. A., Thyme, S. B. & Valen, E. CHOPCHOP v2: a web tool for the next generation of CRISPR genome engineering. *Nucleic Acids Res.* **44**, W272–W276 (2016).
17. Heigwer, F., Kerr, G. & Boutros, M. E-CRISP: fast CRISPR target site identification. *Nat. Methods* **11**, 122–123 (2014).
18. Ma, J. et al. CRISPR-DO for genome-wide CRISPR design and optimization. *Bioinformatics* **32**, 3336–3338 (2016).
19. Singh, R., Kuscu, C., Quinlan, A., Qi, Y. & Adli, M. Cas9–chromatin binding information enables more accurate CRISPR off-target prediction. *Nucleic Acids Res.* **43**, e118 (2015).
20. Cradick, T. J., Qiu, P., Lee, C. M., Fine, E. J. & Bao, G. COSMID: a web-based tool for identifying and validating CRISPR/Cas off-target sites. *Mol. Ther. Nucleic Acids* **3**, e214 (2014).
21. Xu, H.et al. Sequence determinants of improved CRISPR sgRNA design. *Genome Res.* **25**, 1147–1157 2015).
22. Chari, R., Mali, P., Moosburner, M. & Church, G. M. Unraveling CRISPR–Cas9 genome engineering parameters via a library-on-library approach. *Nat. Methods* **12**, 823–826 (2015).
23. Doench, J. G.et al. Rational design of highly active sgRNAs for CRISPR–Cas9-mediated gene inactivation. *Nat. Biotechnol.* **32**, 1262–1267 (2014).
24. Wang, T., Wei, J. J., Sabatini, D. M. & Lander, E. S. Genetic screens in human cells using the CRISPR–Cas9 system. *Science* **343**, 80–84 (2014).
25. Moreno-Mateos, M. A. et al. CRISPRscan: designing highly efficient sgRNAs for CRISPR–Cas9 targeting in vivo. *Nat. Methods* **12**, 982–988 (2015).
26. Housden, B. E. et al. Identification of potential drug targets for tuberous sclerosis complex by synthetic screens combining CRISPR-based knockouts with RNAi. *Sci. Signal.* **8**, rs9 (2015).
27. Kim, D. et al. Genome-wide analysis reveals specificities of Cpf1 endonucleases in human cells. *Nat. Biotechnol.* **34**, 863–868 (2016).
28. Kleinstiver, B. P. et al. Genome-wide specificities of CRISPR–Cas Cpf1 nucleases in human cells. *Nat. Biotechnol.* **34**, 869–874 (2016).
29. Lin, Y. et al. CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences. *Nucleic Acids Res.* **42**, 7473–7485 (2014).
30. Kleinstiver, B. P. et al. High-fidelity CRISPR–Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* **529**, 490–495 (2016).
31. Aguirre, A. J. et al. Genomic copy number dictates a gene-independent cell response to CRISPR/Cas9 targeting. *Cancer Discov.* **6**, 914–929 (2016).
32. Munoz, D. M. et al. CRISPR screens provide a comprehensive assessment of cancer vulnerabilities but generate false-positive hits for highly amplified genomic regions. *Cancer Discov.* **6**, 900–913 (2016).
33. Morgens, D. W. et al. Genome-scale measurement of off-target activity using Cas9 toxicity in high-throughput screens. *Nat. Commun.* **8**, 15178 (2017).
34. Yates, A. et al. Ensembl 2016. *Nucleic Acids Res.* **44**, D710–D716 (2016).
35. Lee, C. M., Davis, T. H. & Bao, G. Examination of CRISPR/Cas9 design tools and the effect of target site accessibility on Cas9 activity. *Exp. Physiol.* https://doi.org/10.1113/EP086043 (2017).
36. Horlbeck, M. A. et al. Nucleosomes impede Cas9 access to DNA in vivo and in vitro. *eLife* **5**, e12677 (2016).
37. Box, G. E. P. & Cox, D. R. An analysis of transformations. *J. R. Stat. Soc. Ser. B Methodol.* **26**, 211–252 (1964).
38. Reyon, D.et al. FLASH assembly of TALENs for high-throughput genome editing. *Nat. Biotechnol.* **30**, 460–465 (2012).
39. Tsai, S. Q., Topkar, V. V., Joung, J. K. & Aryee, M. J. Open-source guideseq software for analysis of GUIDE-seq data. *Nat. Biotechnol.* **34**, 483 (2016).

40. Russell, S. & Norvig, P. *Artificial Intelligence: A Modern Approach: International Edition* 3rd edn (Pearson, New Jersey, 2010).
41. Frank, E., Trigg, L., Holmes, G. & Witten, I. H. Naive Bayes for regression. *Mach. Learn.* **41**, 5–25 (2000).
42. Freund, Y. & Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comp. Syst. Sci.* **55**, 119–139 (1997).
43. Bishop, C. M. *Pattern Recognition and Machine Learning* (Springer, New York, 2007).
44. Wolpert, D. H. Stacked generalization. *Neural Netw.* **5**, 241–259 (1992).
45. Baeza-Yates, R. A. & Perleberg, C. H. Fast and practical approximate string matching. *Inf. Process. Lett.* **59**, 21–27 (1996).
46. Hoffman, M. M. et al. Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res.* **41**, 827–841 (2013).

Acknowledgements

We thank A. Annavajhala for Azure cloud support, C. Kadie for use and support of his HPC cluster code, J. Jernigan, O. Losinets and the HPC team for cluster support, M. Hegde for help with the data, J. Lopez and M. Aryee for assistance with GUIDE-seq data analysis, M. Haeussler for help accessing the data from his paper, and J.-P. Concordet for feedback on the manuscript. M.W. is supported by a UCLA Collaboratory Fellowship. This work used computational and storage services associated with the Hoffman2 Shared Cluster provided by the UCLA Institute for Digital Research and Education's Research Technology Group, and also an Azure-for-Research grant to UCLA. We acknowledge the ENCODE Consortium, the UW ENCODE group for generating these data, and UCSC for processing these data and making them available for download.

Author contributions

J.L. and N.F. designed, implemented and evaluated the machine learning and statistical methods (Elevation-score and Elevation-aggregate). M.W. designed and implemented the Elevation-search infrastructure, also known as dsNickFury. J.G.D. provided biological expertise. B.P.K., J.K.J., J.L., N.F. and J.G.D. selected validation gRNAs. B.P.K., A.A.S. and J.K.J. assayed the validation gRNAs for off-target activity. J.L., N.F., M.W. and J.G.D. designed the web interface. L.H. and K.G. created the front-end webpage for the cloud service. M.E. and J.C. helped run the experiments and populated the cloud server. J.L., M.W., J.G.D., N.F., B.P.K. and J.K.J. wrote the paper.

Competing interests

J.L., L.H., M.E., J.C. and N.F. performed research related to this manuscript while employed by Microsoft. J.K.J. has financial interests in Beam Therapeutics, Editas Medicine, Monitor Biotechnologies, Pairwise Plants, Poseida Therapeutics and Transposagen Biopharmaceuticals. J.K.J.'s interests were reviewed and are managed by Massachusetts General Hospital and Partners HealthCare in accordance with their conflict of interest policies.

Additional information

Supplementary Information is available for this paper at https://doi.org/10.1038/s41551-017-0178-6.

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

Correspondence and requests for materials should be addressed to J.L. or M.W. or J.G.D. or N.F.

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

nature research

Corresponding author(s): Jennifer Listgarten, Michael Weinstein, John Doench, Nicolo Fusi

☐ Initial submission ☐ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

We generated as much new data as it was feasible to perform in the laboratory, taking into account funding and time constraints. We also made use of publicly available data.

2. Data exclusions

Describe any data exclusions.

None.

3. Replication

Describe whether the experimental findings were reliably reproduced.

The GUIDE-Seq assay reported on off-target activities for 22 gRNAs for the newly generated data. Computational predictive modelling yielded superior performance on these and the public data.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

No randomization was required.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

No blinding was required.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	A statement indicating how many times each experiment was replicated
☐	☒	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☒	☐	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒	The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☒	☐	A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☒	☐	Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	The custom software is available on GitHub at https://github.com/MicrosoftResearch

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	The new GUIDE-seq data generated for this study has been deposited with the NCBI Sequence Read Archive (SRA) under accession number SRP117146.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	N/A

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	U2OS cells were a gift from Toni Cathomen, Freiburg.
b. Describe the method of cell line authentication used.	Cell-line identities were validated by STR profiling (ATCC) and deep sequencing.
c. Report whether the cell lines were tested for mycoplasma contamination.	Cells were tested monthly for contamination, and near the end of passaging them before they were discarded. Cell lines tested negative for mycoplasma contamination.
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	No commonly misidentified cell lines were used.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	N/A

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	N/A