npj | systems biology and applications
Article
Published in partnership with the Systems Biology Institute
https://doi.org/10.1038/s41540-024-00431-8

Codon usage and expression-based features significantly improve prediction of CRISPR efficiency

Check for updates

Shaked Bergman[1] & Tamir Tuller[1,2]

CRISPR is a precise and effective genome editing technology; but despite several advancements during the last decade, our ability to computationally design gRNAs remains limited. Most predictive models have relatively low predictive power and utilize only the sequence of the target site as input. Here we suggest a new category of features, which incorporate the target site genomic position and the presence of genes close to it. We calculate four features based on gene expression and codon usage bias indices. We show, on CRISPR datasets taken from 3 different cell types, that such features perform comparably with 425 state-of-the-art predictive features, ranking in the top 2–12% of features. We trained new predictive models, showing that adding expression features to them significantly improves their r^2 by up to 0.04 (relative increase of 39%), achieving average correlations of up to 0.38 on their validation sets; and that these features are deemed important by different feature importance metrics. We believe that incorporating the target site's position, in addition to its sequence, in features such as we have generated here will improve our ability to predict, design and understand CRISPR experiments going forward.

CRISPR (clustered regularly interspaced short palindromic repeats) is a powerful technology to induce mutations in precise genomic locations, with potential to substantially aid the development of treatments to various diseases (such as cancer and AIDS), as well as basic science[1–3]. This technology is a marked improvement upon previous genome-editing tools, such as zinc-finger nucleases (ZFNs) and Transcription activator-like effector nucleases (TALENs), both in precision and required resources; instead of engineering DNA-binding proteins, CRISPR utilizes a guide RNA (gRNA), which binds to complementary target sites and catalyzes a double stranded break (DSB) reaction performed by an endonuclease. In the process of repairing the break, a mutation could be induced – either randomly (via non-homologous end joining, or NHEJ) or by inserting a short sequence of interest (via homology-directed repair).

Despite its precision, the design of a sgRNA with high sensitivity and specificity that would efficiently affect its target and only its target remains a challenge; several computational models were developed to aid with various aspects of this task: tools that predict CRISPR efficiency at a given target site[4–23]; tools that predict the mutations induced by NHEJ[24–28]; and tools that find suitable gRNAs and potential off-target sites for an on-target or gene of interest[29–48]. These tools, while considerably advancing the field of CRISPR, have relatively low sensitivity and specificity[49–55].

One possible explanation for the low performance is the nature of predictive features used in these models: almost exclusively sequence-based features, consisting of nucleotide identities at various target site positions, as well as thermodynamic features, which are sequence-based as well, such as the melting temperature of the DNA target site and the free energy of the gRNA. Only a few models include epigenetic features in addition to the sequence-based ones[5,10,12,18]. The usage of sequence-based features is related to the fact that most state-of-the-art models are trained on in vitro data or out-of-context target sites inserted via a lentivirus; in many cases, the dataset does not include the target site's location at all and lists only its sequence. As a result, features based on the target site's genomic location are expected to have low to negligible correlations with CRISPR efficiency, since the data does not fully reflect CRISPR action in its intended, in vivo setting. While the in vitro datasets are larger and more numerous than the ex vivo datasets, computational modeling of CRISPR should strive toward using ex vivo and in vivo datasets.

Gene expression, being the basis for life, is a complex process governed by a multitude of factors and conditions (reviewed in ref. 56). Measurements of gene expression reflect an orchestra of RNAs and proteins working in tandem to produce the correct protein at the correct time and are related to virtually every cellular process. Different aspects of gene expression are encoded in the codon sequence itself (reviewed in ref. 57), and several codon

[1]Department of Biomedical Engineering, Tel-Aviv University, Tel Aviv, Israel. [2]The Sagol School of Neuroscience, Tel-Aviv University, Tel Aviv, Israel.
e-mail: tamirtul@tauex.tau.ac.il

usage bias (CUB) indices were created to estimate the way codon composition affects gene expression (reviewed in ref. 58). For these reasons – the information encoded in expression levels, and the relative ease of estimating expression via CUB indices—we chose to evaluate expression data as potential predictive features of CRISPR efficiency. For example, highly expressed regions may entail a higher rate of cellular functions, which may be related to CRISPR's ease-of-access to target sites, for example via chromatin accessibility or endogenous factors that aid CRISPR action. Here we show that expression measurements and CUB indices can be used as useful features to predict CRISPR efficiency and could be an important tool in understanding the way CRISPR works ex vivo and in vivo.

Results

The general structure of the study

In this study, we have set out to assess the viability of CUB and expression features as predictive features of CRISPR efficiency. To that end, we downloaded empirical CRISPR ex-vivo data (see "Acquiring ex-vivo CRISPR efficiency"), generated features based on their corresponding cell types (see "Calculating CUB and expression features") and checked that our features' relation with CRISPR efficiency is not diminished when controlling for chromatin accessibility (see "CUB/expression correlation is not explained by chromatin accessibility"). We compared our features to well-established features used in state-of-the-art models ("CUB and expression features outperform most classic features") and checked whether they encode new information compared to these models ("CUB and expression features encode information orthogonal to state-of-the-art models"); we then checked for consistent trends in the relationship between our features and CRISPR efficiency ("High-efficiency sites reside in significantly higher-expressed genes relative to low-efficiency sites"). Finally, we trained predictive CRISPR models with the well-established features and checked whether adding our features significantly improved them, and whether our features are marked as important by various feature importance metrics ("CUB and expression features significantly improve CRISPR prediction").

Acquiring ex-vivo CRISPR efficiency

To properly evaluate the predictive power of expression-based features, we required efficiency data measured ex vivo rather than in vitro; we downloaded the 3 largest ex vivo CRISPR efficiency datasets, which are based on 3 different cell types: T cell data from Leenay et al., including 1574 sites[26]; HEK293 data from TTISS, including 666 sites[59]; and U2OS data from GUIDE-Seq, including 260 sites[60]. The Leenay dataset includes on-target data, whereas the other two sets are largely off-target data.

Calculating CUB and expression features

To generate expression features, we downloaded the transcript sequences and coordinates available in Ensembl v.109[61] and found for each target site its nearby genes (which we defined as genes whose genomic distance from the target sites is up to 1000 codons, or 3000nt). The features are calculated for each gene, and the feature value for each site is defined as the average feature value over its nearby genes.

We had initially generated 7 different features, 3 of which were lowly correlated with CRISPR efficiency (See methods and Supplementary Table 1). We discarded these features and continued our analysis with the remaining 4: (A) Expression, i.e. normalized mRNA levels of each gene, downloaded from the Expression Atlas[62] for T, HEK293 and U2OS cells. (B) ChimeraARS[63]; this index captures high-dimensional patterns in gene sequences based on a reference set. We chose the top 2% expressed genes (based on the Expression Atlas data) as a reference set. (C) Normalized translational efficiency (nTE)[64], which estimates the tRNA supply-and-demand in the cell and how suited the gene is to that supply. (D) Relative codon bias strength (RCBS), which estimates the codon usage bias of the transcript[65]. We categorize feature A as an "Expression" feature, and features B-D as "CUB" features. The first three features are calculated using transcript sequences and expression levels, based on the dataset's cell type; whereas RCBS is based solely on transcript sequences, and is identical for the 3 cell types.

CUB/expression correlation is not explained by chromatin accessibility

Epigenetic accessibility was shown to be related to CRISPR efficiency and is used in a few predictive models[5,10,12,18]. To validate whether the correlations between CUB/expression and CRISPR efficiency are due to that relation, we calculated the partial correlation between our features and CRISPR efficiency when controlling for chromatin accessibility based on DNAse measurements (see the methods section for details). While the correlations were indeed reduced after controlling for accessibility, the reduction was not substantial for the highly correlative features (partial correlations around 0.1, similar to the full correlations); on the other hand, the lowly correlative features' correlation was reduced more noticeably, with a halving (or more) of the correlation in some cases (Supplementary Table 2). This indicates that accessibility can explain some, but not all, of the CUB/expression correlations with CRISPR efficiency. We note here and in the discussion section that CUB/expression features are substantially easier to acquire and calculate for a variety of cell types and tissues, compared to chromatin accessibility.

CUB and expression features outperform most classic features

To evaluate the CUB and expression, we first calculated 425 well-established features used in most CRISPR models; of these, 420 were sequence features, indicating the identity of each nucleotide at each position, as well as the fraction of each nucleotide and dinucleotide in the target site; and 5 were thermodynamic features, estimating the free energy of different regions in the sgRNA and the melting temperature of the DNA target site (see "Calculating classic features" for more details). We then calculated the correlation between the features and CRISPR efficiency (Fig. 1a). The CUB and expression features ranked highly in all 3 datasets, with the best CUB feature ranked in the top 12%/2%/7% in the T/HEK293/U2OS datasets, respectively. The expression feature was highly correlative in the T and HEK293 dataset, ranking in the top 5%/3%, respectively.

In the T and HEK293 datasets, the CUB features outperformed the thermodynamic features: average correlation of 0.06 (CUB) vs. 0.03 (thermodynamic) in T cells, and 0.13 (expression) vs. 0.03 (thermodynamic) in HEK293. In the U2OS set, while the best thermodynamic feature was more correlative than the best CUB feature (0.11 vs. 0.1, respectively), the averages of the feature sets were similar, with a slight advantage for the CUB features (0.06 vs. 0.07).

The expression feature outperformed the thermodynamic features in the T and HEK293 sets as well, with correlations of 0.13/0.1 (expression) vs. 0.07/0.06 (best thermodynamic feature).

Since 400 out of the 425 classic features are binary while our expression features are continuous, we conducted an additional comparison by converting all continuous features to binary features. For each feature, we calculated the average of its maximal and minimal values on the dataset; we then defined feature values higher than that average as 1 and features lower (or equal) than 0. We calculated the median efficiencies of sites with feature value 1/0 and took the ratio between the lower and higher values such that the ratio is ≤1. We expect this ratio to be lower for more informative features since this would denote a clearer difference between high-efficiency and low-efficiency sites (Supplementary Fig. 1). The best CUB feature was ranked in the top 13%/37%/46% out of all features and the expression feature was ranked in the top 14%/32%/7% out of all features, for the T cell/HEK293/U2OS dataset, respectively. When comparing these features only to the 29 features that were originally continuous, the best CUB feature was ranked 3rd/12th/12th, and the expression feature was ranked 5th/9th/1st.

From these results, we concluded that CUB and expression features perform well compared to the classic features and should be considered for use in predictive models.

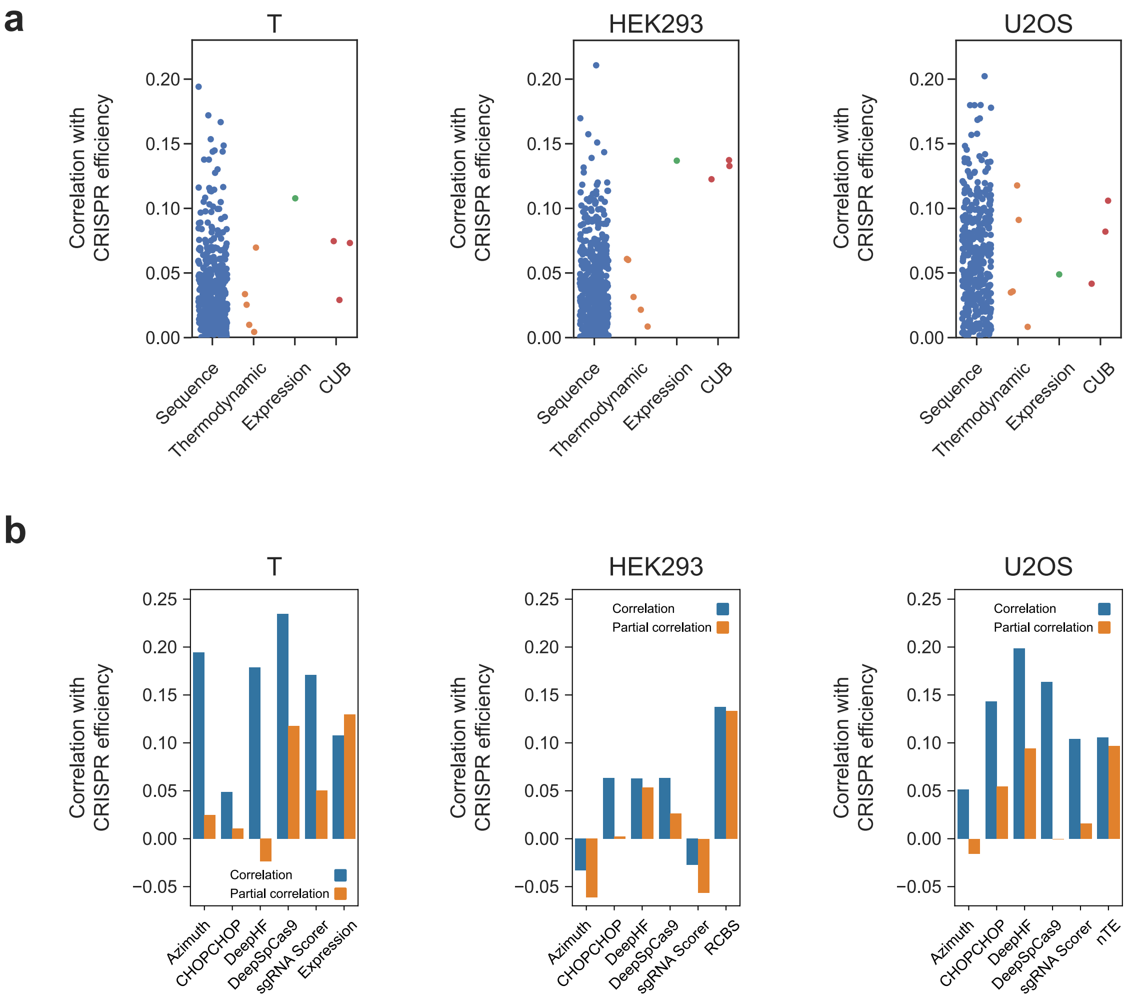


Fig. 1 | CUB and Expression features are correlative with CRISPR efficiency and encode new information. a Spearman correlations (absolute values) between predictive features and CRISPR efficiency. **b** Bar charts of Spearman correlations (blue) and partial correlations (orange) between 5 state-of-the-art models and CUB/expression feature with CRISPR efficiency. The partial correlation of each vector controls for all other vectors in the cell subplot.

CUB and expression features encode information orthogonal to state-of-the-art models

Next, we assessed whether our features encode information that is not included in current state-of-the-art prediction models; to that end, we used 5 such models to predict CRISPR efficiency: Azimuth[66], DeepSpCas9[67], CHOPCHOP[47] (using the Moreno-Mateos[4], i.e., CRISPRscan, scoring scheme), DeepHF[16] and sgRNA Scorer[9]. For each dataset we found the best CUB/expression feature (based on its correlation with CRISPR efficiency) and calculated its partial correlation with CRISPR efficiency when controlling for the 5 model scores (Fig. 1b). For all 3 datasets, the partial correlation was almost identical to the full correlation, indicating that the CUB/expression features' information is not included in the state-of-the-art models and cannot be explained by them.

High-efficiency sites reside in significantly higher-expressed genes relative to low-efficiency sites

We compared the CUB and expression feature values of sites with the top 20% to sites with bottom 20% efficiency, to discern whether these values differ significantly between the two groups, and whether the direction of the relationship between expression and efficiency is consistent (Fig. 2); for each dataset, 2–4 out of the 4 features were significantly different between the groups (*p*-values were calculated using Wilcoxon's rank-sum test and appear in Supplementary Table 3), and in all cases the high-efficiency sites resided in regions with higher expression than the low-efficiency sites. Thus, CUB and expression significantly differ between high-efficiency and low-efficiency sites and is consistently positively related to CRISPR efficiency.

CUB and expression features significantly improve CRISPR prediction

In order to check whether using expression-based features can aid in predicting CRISPR efficiency, we trained models on the 3 datasets using repeated 5-fold cross validation (with 200 repeats). For each train-test split, we trained a model using only the classic features, and – separately – a model using these classic features and our 4 CUB/expression features. Thus, we received a distribution of 1000 correlations for each feature set (classic vs. classic + CUB/expression) in each cell type, and compared their performances (Fig. 3). We conducted this analysis for LASSO and xgboost models (using the Python scikit-learn and xgboost packages, respectively).

In all 6 cases, the models were significantly improved when adding the CUB/expression features. The average r^2 of the T cell/HEK293/U2OS dataset increased by 0.01/0.04/0.01 (4%/39%/11%) in the LASSO models, and 0.02/0.03/0.01 (15%/32%/15%) in the xgboost models, respectively.

We then evaluated our features' importance in the trained models relative to the classic features (Fig. 4, Supplementary Table 4). For the LASSO models, we counted the number of times each feature was selected out of the 1000 cross validation repeats, as well as its permutation feature importance (using scikit-learn). For xgboost, we used the permutation feature importance, xgboost's built-in "gain" importance measure, and the popular SHAP importance[68]. The features were ranked highly, with the best CUB/expression feature being ranked in the top 21 features (out of 429) across all 5 measures and 3 cell types. In 14 out of the 15 per-cell measures, the best CUB/expression features ranked in the top 10 features. These results demonstrate that CUB/expression features can be used to improve CRISPR predictive models.

Discussion

In this paper, we have analyzed the value of using predictive CUB and expression features, that are based on the position – rather than only the sequence – of CRISPR target sites. After generating 4 CUB/expression features – based on mRNA levels, ChimeraARS score, nTE score and RCBS – we evaluated them on the 3 largest ex vivo CRISPR datasets that were published. The 3 dataset experiments were conducted on 3 different cell types: T cells, HEK293 and U2OS.

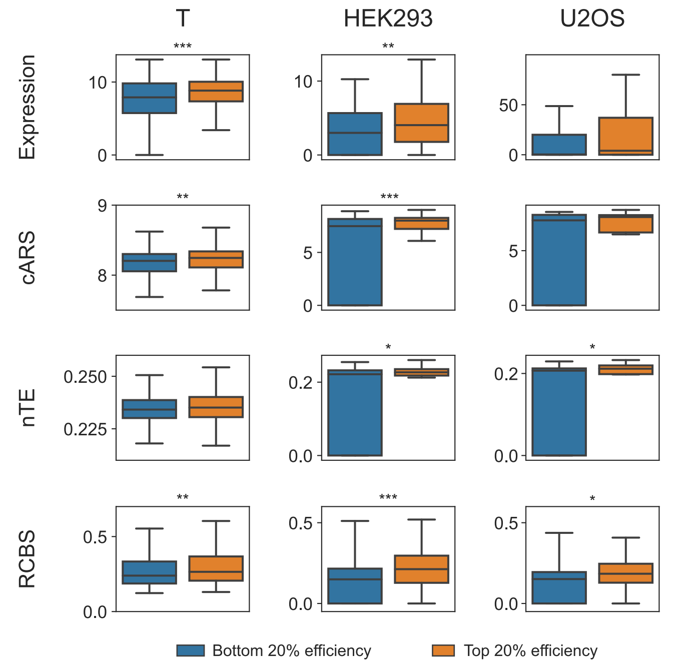


Fig. 2 | High-efficiency sites reside in higher-expression regions. Boxplots of the CUB/expression features for the sites with top/bottom 20% efficiency (orange/blue plots, respectively) in the T cell, HEK293 and U2OS datasets. Asterisks denote significance of difference between the top and bottom sites using Wilcoxon's rank-sum test. $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

Comparing the CUB and expression features to classic, sequence- and thermodynamics-based features that are used in most CRISPR prediction models, we found the CUB/expression features to correlate relatively highly with CRISPR efficiency - with the best CUB/expression feature in each dataset ranking in the top 2–12% of features. In 2 of the 3 datasets, the CUB and expression features clearly outperformed established thermodynamic features, and in the third database the two types of features performed relatively similarly. Our features' correlations with CRISPR efficiency were not substantially reduced when controlling for 5 state-of-the-art models, indicating our features indeed bring new information to the table, information not encoded in current models.

The expression feature performed well on the T cell and HEK293 datasets, but had a relatively low correlation on the U2OS dataset; whereas the CUB features as a whole were highly correlative on all 3 datasets, with different features being ranked the best in each one. Codons encode multiple aspects of gene function via complex and high-level codes[57], and it is possible the CUB indices capture some of these codes better than the direct gene expression measurements, which – despite having evolved considerably in the last decades – can still be noisy, biased, and do not capture all gene expression steps (e.g., the post transcriptional steps). Thus, we believe incorporating both expression-based and CUB-based features would lead to the best results.

We also found that gene expression was significantly different between high-efficiency and low-efficiency sites, with high-efficiency sites residing in regions with higher expression. This is consistent with the correlations between the features and efficiency, which were all positive.

We evaluated the predictive power our features confer by training LASSO and xgboost models with and without the CUB and expression features, and comparing their performances; adding the 4 CUB/expression features significantly improved the models, by up to 0.04 (relative increase of 39%), and the predictive models achieved average correlations of up to 0.38 on their validation sets. We believe the fact that only 4 features achieve that much shows their potential for CRISPR prediction, especially considering new CRISPR models continue to include sequence-based features almost exclusively. The features were also ranked as highly important in their

Fig. 3 | CUB and expression features significantly improve CRISPR prediction. Histograms of the correlations between measured and predicted efficiency in the T cell, HEK293 and U2OS datasets, when testing LASSO and xgboost models. The blue/orange histogram indicates the model with/without the CUB and expression features, and the average correlation is marked with a solid/dashed line, respectively. Arrow: direction and magnitude of difference between histograms' averages. p-values were calculated using Wilcoxon's signed rank test.

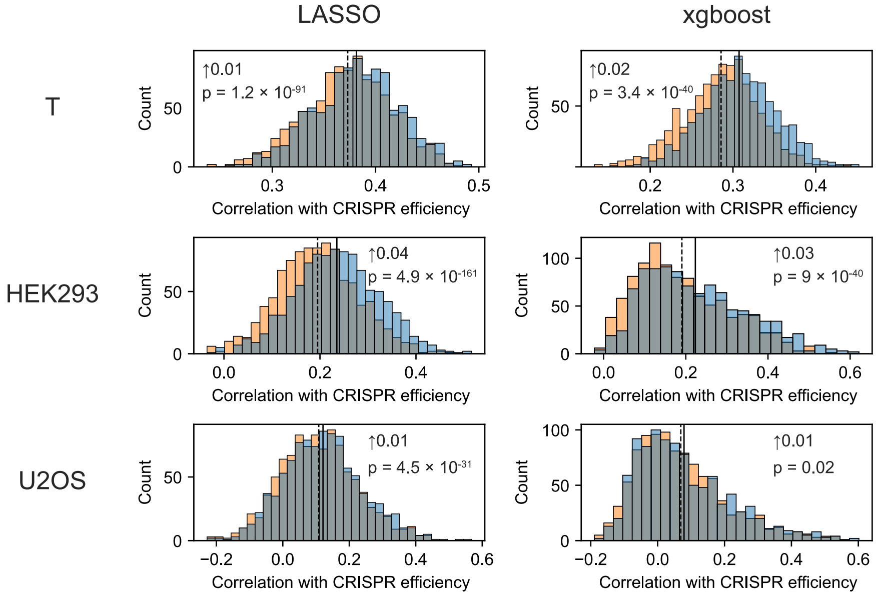
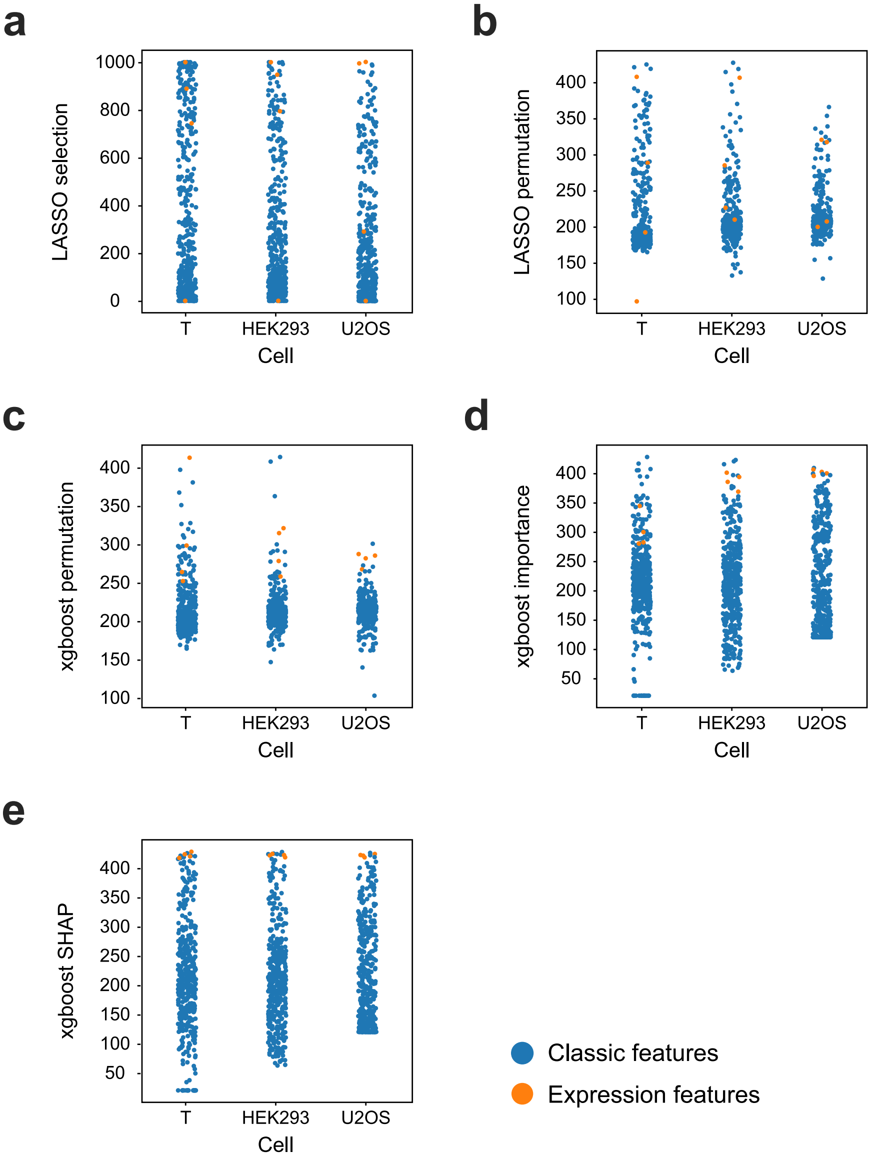


Fig. 4 | CUB and expression features rank highly compared to classic features. Ranks of features used in the LASSO and xgboost models, averaged over the 1000 cross validation iterations; a higher rank indicates a more important feature. The classic and CUB/expression features are marked with blue and orange dots, respectively. **a** Number of times each feature was selected in the LASSO models. **b** Feature rank based on LASSO permutation importance. **c** Feature rank based on xgboost permutation importance. **d** Feature rank based on xgboost's "gain" importance. **e** Feature rank based on SHAP values of the xgboost models.



models by the number of times they were selected by LASSO, permutation importance, xgboost's "gain" importance and SHAP values.

As described in the previous paragraphs, we endeavored to assess our features using multiple angles and tests: correlation vs. other features, full/partial correlation vs. other models, separation between high-efficiency and low-efficiency sites, model improvement, feature selection and importance. Correlations in the field are relatively low – this is the current state of CRISPR prediction efficiency, which we aim to improve here. Correlations of up to 0.2 were observed in multiple previous studies[17,52–55]; this is demonstrated by the model performance shown in Fig. 1b as well, where state-of-the-art models achieve correlations of up to 0.2 – and in some cases considerably lower – with CRISPR efficiency. These 5 models are well established in the field and cited in numerous papers; they mark the de facto state-of-the-art in CRISPR prediction.

Our features are based on ex vivo conditions, and as such have to be evaluated on such data. Ex vivo measurements are a better representation of the true efficiency of CRISPR in the cell, compared to in vitro measurements – but they include more noise, heterogeneity, and relevant affecting variables compared to in vitro measurements, making CRISPR prediction a challenge.

Since gene expression is an outcome of many different processes, it is hard to pinpoint the exact reason for which expression is useful in predicting CRISPR efficiency. We hypothesize that highly expressed regions are more active regions, in which the CRISPR complex can find its way more easily to its target – for example, by a higher physical accessibility of the target sites due to the three-dimensional conformation of the DNA, or a higher presence of endogenous factors in these areas which could aid CRISPR action. We have shown that the correlation between expression/CUB and CRISPR efficiency is reduced, but not completely diminished, when controlling for chromatin accessibility; indicating some (but not all) of the expression-CRISPR relation may be attributed to accessibility. Such regions may have a lower density of DNA, leading to fewer sites with partial complementarity to the gRNA; these sites can compete with the CRISPR on-target, thus having fewer sites nearby is expected to increase CRISPR efficiency[69].

One major advantage of CUB/expression features, compared to other position-based features such as those based on epigenetics, is that expression levels (measured by RNA-seq) are available for a large number of cells and tissues, since they are relatively simpler and cheaper to acquire. In the case of non-model organisms, for which experimental data is scarce, many CUB features (such as the ChimeraARS and RCBS used here) require only gene sequences to calculate, and so can be generated on them as well. Thus, these features could easily be calculated for the vast majority of CRISPR experiments.

The main limitation of our work remains the relatively scarce measurements of ex vivo CRISPR efficiency; there is a growing trend of conversion from in vitro measurements to ex vivo measurements, but the largest datasets are still measured in vitro. While certainly useful to the field, the intended use of CRISPR is, naturally, in vivo. Since ex vivo experiments can prove to be very different from in vitro ones, additional ex vivo CRISPR datasets would allow more accurate assessments of CRISPR usage in real-world situations. In addition, while we assessed three different cell types, all three were eukaryotic cells – and specifically, human cells. It remains to be seen whether our features could improve CRISPR prediction in other organisms; for example, since chromatin is absent in prokaryotes, our hypothesis regarding the accessibility information contained in our features would be relevant to a lesser extent there. Nevertheless, since expression levels are the culmination of numerous cell processes, we believe it feasible for expression-based features to be informative in prokaryotes as well, perhaps indicating other favorable conditions for CRISPR action.

This study demonstrates the usefulness of using features that are based on the wealth of information accumulated regarding each genomic location. Incorporating such features, specific to the cell type in which the CRISPR experiment is conducted, would enhance our ability to predict, understand and utilize CRISPR technology.

Methods

Acquiring CRISPR efficiency measurements

For T cells we used the "MutationEfficiency", "IndelCounts" and "Insertion" files published by Leenay et al.[26]; we filtered out sites with fewer than 1000 mapped reads, and calculated the efficiency as the fraction of edited reads out of all reads mapped to the site. For HEK293 cells we used the TTISS dataset[59]; we defined each site's efficiency as the average number of SpCas9 Seq reads mapped to it, keeping only sites with at least one such read. For U2OS cells we used the number of GUIDE-Seq reads mapped to each site as efficiency scores[60].

Acquiring gene expression levels

We downloaded mRNA levels from the Expression Atlas[62]. For the T cell and HEK293 datasets, we used normalized expression values from accession IDs E-GEOD-36766 and E-GEOD-14429, respectively. For the U2OS dataset, we used the TPM values from the osteosarcoma (MG63) measurements in E-MTAB-2706.

Calculating CUB features

We calculated two versions of the codon adaptation index (CAI)[70] scores for each gene, using two different reference sets: the whole human transcriptome, and transcripts with the top 2% expression. For each codon in each reference set, we calculated its CAI weight as its frequency in the set relative to the most frequent synonymous codon. The CAI weight of a gene is then the geometric mean of its codons' CAI weights.

We calculated the ChimeraARS score for each gene using the ChimeraUGEM program[63], setting the reference set as the transcripts with the top 2% expression for each cell type. ChimeraARS calculates the similarity between a given sequence and a reference set by finding the longest substring, common to the gene and the set, in each position of the sequence.

We calculated the tRNA adaptation index (tAI)[71] using human tAI weights from ref. 72; the tAI score of a gene is the geometric mean of its codons' tAI weights.

We used these same weights, and the downloaded per-cell mRNA levels, to calculate the nTE weights; a codon's nTE weight is defined as its tAI weight (the "tRNA supply") divided by its number of appearances in the transcriptome (i.e., the sum of its appearances in each transcript, multiplied by the transcript's abundance; this is the "tRNA demand").

We defined each genomic target site's CUB/expression feature as the average corresponding feature value of all genes which reside up to 3000nt away from the target site.

Calculating epigenetic features

We downloaded DNAse bigwig files from ENCODE[73], and defined the accessibility feature as the average value over the target site's coordinates. The accession ID and files used are listed in Supplementary Table 5.

Calculating classic features

We calculated 420 sequence-based features and 5 thermodynamic features.

For the sequence-based features, we used binary features denoting the identity of each nucleotide and dinucleotide along the 20nt target site (e.g., the feature pos5_G is 1 if there is a G in position 5, 0 otherwise). This results in 4×20 (nucleotide at each position) $+ 16 \times 20$ (dinucleotide at each position), i.e, 400, features. We also calculated the overall frequency of each nucleotide and each dinucleotide over the target site, resulting in 20 features.

The method of calculating thermodynamic features was taken from DeepHF[16]: 4 features denoting the melting temperatures of various regions along the DNA target site (positions 1–20, 1–4, 5–12 and 13–20), and the gRNA's free energy. Melting temperatures were calculated with using Biopython, and free energy was calculated using RNAfold[74].

The evaluation metric r^2

We report the percentage increase in the performances on the r^2 values. r^2 is a well-established metric that was used in previous studies in the field as it estimates the percentage in the variance of the CRISPR efficiency that can be explained by our models. We believe that the percentage of improvement is a very important metric as it provides a good comparison to the current "state-of-the-art".

Data availability

The datasets utilized in this study have been downloaded from publicly available datasets. CRISPR efficiency datasets were downloaded from refs. 26,59,60. mRNA levels were downloaded from Expression Atlas[62] (accession IDs E-GEOD-36766, E-GEOD-14429, E-MTAB-2706). tAI weights were downloaded from ref. 72. DNAse data was downloaded from ENCODE[73] (accession IDs ENCFF268DVI, ENCFF412ONC, ENCFF437CNA, ENCFF526NOL, ENCFF635ZUA, ENCFF529BOG, ENCFF418OBI).

Received: 23 February 2024; Accepted: 27 August 2024;
Published online: 03 September 2024

References

1. Doudna, J. A. & Charpentier, E. The new frontier of genome engineering with CRISPR-Cas9. *Science* **346**, 1258096 (2014).
2. Pickar-Oliver, A. & Gersbach, C. A. The next generation of CRISPR–Cas technologies and applications. *Nat. Rev. Mol. Cell Biol.* **20**, 490–507 (2019).
3. Li, H. et al. Applications of genome editing technology in the targeted therapy of human diseases: mechanisms, advances and prospects. *Signal Transduct. Target. Ther.* **5**, 1 (2020).
4. Moreno-Mateos, M. A. et al. CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. *Nat. Methods* **12**, 982–988 (2015).
5. 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–e118 (2015).
6. Xu, H. et al. Sequence determinants of improved CRISPR sgRNA design. *Genome Res.* **25**, 1147–1157 (2015).
7. Kaur, K., Gupta, A. K., Rajput, A. & Kumar, M. ge-CRISPR - an integrated pipeline for the prediction and analysis of sgRNAs genome editing efficiency for CRISPR/Cas system. *Sci. Rep.* **6**, 30870 (2016).
8. Labuhn, M. et al. Refined sgRNA efficacy prediction improves large- and small-scale CRISPR–Cas9 applications. *Nucleic Acids Res.* **46**, 1375–1385 (2018).
9. Chari, R., Yeo, N. C., Chavez, A. & Church, G. M. sgRNA Scorer 2.0: a species-independent model to predict CRISPR/Cas9 activity. *ACS Synth. Biol.* **6**, 902–904 (2017).

10. Abadi, S., Yan, W. X., Amar, D. & Mayrose, I. A machine learning approach for predicting CRISPR-Cas9 cleavage efficiencies and patterns underlying its mechanism of action. *PLOS Comput. Biol.* **13**, e1005807 (2017).
11. Listgarten, J. et al. Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs. *Nat. Biomed. Eng.* **2**, 38–47 (2018).
12. Chuai, G. et al. DeepCRISPR: optimized CRISPR guide RNA design by deep learning. *Genome Biol.* **19**, 80 (2018).
13. Lin, J. & Wong, K.-C. Off-target predictions in CRISPR-Cas9 gene editing using deep learning. *Bioinformatics* **34**, i656–i663 (2018).
14. Peng, H., Zheng, Y., Blumenstein, M., Tao, D. & Li, J. CRISPR/Cas9 cleavage efficiency regression through boosting algorithms and Markov sequence profiling. *Bioinformatics* **34**, 3069–3077 (2018).
15. Alkan, F., Wenzel, A., Anthon, C., Havgaard, J. H. & Gorodkin, J. CRISPR-Cas9 off-targeting assessment with nucleic acid duplex energy parameters. *Genome Biol.* **19**, 177 (2018).
16. Wang, D. et al. Optimized CRISPR guide RNA design for two high-fidelity Cas9 variants by deep learning. *Nat. Commun.* **10**, 4284 (2019).
17. Xue, L., Tang, B., Chen, W. & Luo, J. Prediction of CRISPR sgRNA activity using a deep convolutional neural network. *J. Chem. Inf. Model.* **59**, 615–624 (2019).
18. Zhang, G., Dai, Z. & Dai, X. A novel hybrid CNN-SVR for CRISPR/Cas9 guide RNA activity prediction. *Front. Genet.* **10**, 1303 (2019).
19. Dimauro, G. et al. CRISPRLearner: a deep learning-based system to predict CRISPR/Cas9 sgRNA on-target cleavage efficiency, GiovanniAU - Colagrande. *Electronics* **8**, 1478 (2019).
20. Hiranniramol, K., Chen, Y., Liu, W. & Wang, X. Generalizable sgRNA design for improved CRISPR/Cas9 editing efficiency. *Bioinformatics* **36**, 2684–2689 (2020).
21. Niu, R., Peng, J., Zhang, Z. & Shang, X. R-CRISPR: a deep learning network to predict off-target activities with mismatch, insertion and deletion in CRISPR-Cas9 system. *Genes* **12**, 1878 (2021).
22. Zhang, G., Dai, Z. & Dai, X. C-RNNCrispr: prediction of CRISPR/Cas9 sgRNA activity using convolutional and recurrent neural networks. *Comput. Struct. Biotechnol. J.* **18**, 344–354 (2020).
23. Konstantakos, V., Nentidis, A., Krithara, A. & Paliouras, G. CRISPRedict: a CRISPR-Cas9 web tool for interpretable efficiency predictions. *Nucleic Acids Res.* **50**, W191–W198 (2022).
24. Shen, M. W. et al. Predictable and precise template-free CRISPR editing of pathogenic variants. *Nature* **563**, 646–651 (2018).
25. Chen, W. et al. Massively parallel profiling and predictive modeling of the outcomes of CRISPR/Cas9-mediated double-strand break repair. *Nucleic Acids Res.* **47**, 7989–8003 (2019).
26. Leenay, R. T. et al. Large dataset enables prediction of repair after CRISPR–Cas9 editing in primary T cells. *Nat. Biotechnol.* **37**, 1034–1037 (2019).
27. Allen, F. et al. Predicting the mutations generated by repair of Cas9-induced double-strand breaks. *Nat. Biotechnol.* **37**, 64–72 (2019).
28. Li, V. R., Zhang, Z. & Troyanskaya, O. G. CROTON: an automated and variant-aware deep learning framework for predicting CRISPR/Cas9 editing outcomes. *Bioinformatics* **37**, i342–i348 (2021).
29. Zhu, L. J., Holmes, B. R., Aronin, N. & Brodsky, M. H. CRISPRseek: a Bioconductor package to identify target-specific guide RNAs for CRISPR-Cas9 genome-editing systems. *PLoS One* **9**, e108424 (2014).
30. Xie, S., Shen, B., Zhang, C., Huang, X. & Zhang, Y. sgRNAcas9: a software package for designing CRISPR sgRNA and evaluating potential off-target cleavage sites. *PLoS One* **9**, e100448 (2014).
31. 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).
32. Xiao, A. et al. CasOT: a genome-wide Cas9/gRNA off-target searching tool. *Bioinformatics* **30**, 1180–1182 (2014).
33. Heigwer, F., Kerr, G. & Boutros, M. E-CRISP: fast CRISPR target site identification. *Nat. Methods* **11**, 122–123 (2014).
34. 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).
35. 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).
36. Liu, H. et al. CRISPR-ERA: a comprehensive design tool for CRISPR-mediated gene editing, repression and activation. *Bioinformatics* **31**, 3676–3678 (2015).
37. Peng, D. & Tarleton, R. EuPaGDT: a web tool tailored to design CRISPR guide RNAs for eukaryotic pathogens. *Microb. Genom.* **1**, e000033 (2015).
38. Oliveros, J. C. et al. Breaking-Cas—interactive design of guide RNAs for CRISPR-Cas experiments for ENSEMBL genomes. *Nucleic Acids Res.* **44**, W267–W271 (2016).
39. Pulido-Quetglas, C. et al. Scalable design of paired CRISPR guide RNAs for genomic deletion. *PLOS Comput. Biol.* **13**, e1005341 (2017).
40. Perez, A. R. et al. GuideScan software for improved single and paired CRISPR guide RNA design. *Nat. Biotechnol.* **35**, 347–349 (2017).
41. Liu, H. et al. CRISPR-P 2.0: an improved CRISPR-Cas9 tool for genome editing in plants. *Mol. Plant* **10**, 530–532 (2017).
42. Xie, X. et al. CRISPR-GE: a convenient software toolkit for CRISPR-based genome editing. *Mol. Plant* **10**, 1246–1249 (2017).
43. Concordet, J.-P. & Haeussler, M. CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. *Nucleic Acids Res.* **46**, W242–W245 (2018).
44. McKenna, A. & Shendure, J. FlashFry: a fast and flexible tool for large-scale CRISPR target design. *BMC Biol.* **16**, 74 (2018).
45. Peng, H., Zheng, Y., Zhao, Z., Liu, T. & Li, J. Recognition of CRISPR/Cas9 off-target sites through ensemble learning of uneven mismatch distributions. *Bioinformatics* **34**, i757–i765 (2018).
46. Jacquin, A. L. S., Odom, D. T. & Lukk, M. Crisflash: open-source software to generate CRISPR guide RNAs against genomes annotated with individual variation. *Bioinformatics* **35**, 3146–3147 (2019).
47. Labun, K. et al. CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing. *Nucleic Acids Res.* **47**, W171–W174 (2019).
48. Minkenberg, B., Zhang, J., Xie, K. & Yang, Y. CRISPR-PLANT v2: an online resource for highly specific guide RNA spacers based on improved off-target analysis. *Plant Biotechnol. J.* **17**, 5–8 (2019).
49. Bao, X. R., Pan, Y., Lee, C. M., Davis, T. H. & Bao, G. Tools for experimental and computational analyses of off-target editing by programmable nucleases. *Nat. Protoc.* **16**, 10–26 (2021).
50. Newman, A., Starrs, L. & Burgio, G. Cas9 cuts and consequences; detecting, predicting, and mitigating CRISPR/Cas9 on- and off-target damage. *BioEssays* **42**, 2000047 (2020).
51. Sledzinski, P., Nowaczyk, M. & Olejniczak, M. Computational tools and resources supporting CRISPR-Cas experiments. *Cells* **9**, 1288 (2020).
52. Wang, J., Zhang, X., Cheng, L. & Luo, Y. An overview and metanalysis of machine and deep learning-based CRISPR gRNA design tools. *RNA Biol.* **17**, 13–22 (2020).
53. Konstantakos, V., Nentidis, A., Krithara, A. & Paliouras, G. CRISPR–Cas9 gRNA efficiency prediction: an overview of predictive tools and the role of deep learning. *Nucleic Acids Res.* **50**, 3616–3637 (2022).
54. Alipanahi, R., Safari, L. & Khanteymoori, A. CRISPR genome editing using computational approaches: a survey. *Front. Bioinforma.* **2**, 1001131 (2023).
55. Liu, G., Zhang, Y. & Zhang, T. Computational approaches for effective CRISPR guide RNA design and evaluation. *Comput. Struct. Biotechnol. J.* **18**, 35–44 (2020).
56. Buccitelli, C. & Selbach, M. mRNAs, proteins and the emerging principles of gene expression control. *Nat. Rev. Genet.* **21**, 630–644 (2020).

57. Bergman, S. & Tuller, T. Widespread non-modular overlapping codes in the coding regions. *Phys. Biol.* **17**, 31002 (2020).
58. Bahiri-Elitzur, S. & Tuller, T. Codon-based indices for modeling gene expression and transcript evolution. *Comput. Struct. Biotechnol. J.* **19**, 2646–2663 (2021).
59. Schmid-Burgk, J. L. et al. Highly parallel profiling of Cas9 variant specificity. *Mol. Cell* **78**, 794–800.e8 (2020).
60. 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).
61. Martin, F. J. et al. Ensembl 2023. *Nucleic Acids Res.* **51**, D933–D941 (2023).
62. Moreno, P. et al. Expression Atlas update: gene and protein expression in multiple species. *Nucleic Acids Res.* **50**, D129–D140 (2022).
63. Diament, A. et al. ChimeraUGEM: unsupervised gene expression modeling in any given organism. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btz080 (2019).
64. Pechmann, S. & Frydman, J. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nat. Struct. Mol. Biol.* **20**, 237–243 (2013).
65. Roymondal, U., Das, S. & Sahoo, S. Predicting gene expression level from relative codon usage bias: an application to Escherichia coli genome. *DNA Res.* **16**, 13–30 (2009).
66. 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).
67. Kwon, K. H. et al. SpCas9 activity prediction by DeepSpCas9, a deep learning–based model with high generalization performance. *Sci. Adv.* **5**, eaax9249 (2022).
68. Lundberg, S. M. et al. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, 56–67 (2020).
69. Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C. & Doudna, J. A. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507**, 62–67 (2014).
70. Sharp, P. M. & Li, W. H. The codon Adaptation Index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15**, 1281–1295 (1987).
71. Reis, M. D., Savva, R. & Wernisch, L. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res.* **32**, 5036–5044 (2004).
72. Tuller, T. et al. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell* **141**, 344–354 (2010).
73. Luo, Y. et al. New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res.* **48**, D882–D889 (2020).
74. Lorenz, R. et al. ViennaRNA package 2.0. *Algorithms Mol. Biol.* **6**, 26 (2011).

Acknowledgements

S.B. is supported by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel Aviv University. The study was also supported by CRISPR-IL consortium. The funders played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

Author contributions

S.B. and T.T. conceived the study and designed the methodologies. S.B. collected the data and performed all the analyses under the guidance of T.T.; S.B. and T.T. wrote the manuscript. Both authors have read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41540-024-00431-8.

Correspondence and requests for materials should be addressed to Tamir Tuller.

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

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