CRISPR/Cas9 gRNA activity depends on free energy changes and on the target PAM context

Corsi, Giulia I; Qu, Kunli; Alkan, Ferhat; Pan, Xiaoguang; Luo, Yonglun; Gorodkin, Jan

Published in:
Nature Communications

DOI:
10.1038/s41467-022-30515-0

Publication date:
2022

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
Corsi, G. I., Qu, K., Alkan, F., Pan, X., Luo, Y., & Gorodkin, J. (2022). CRISPR/Cas9 gRNA activity depends on free energy changes and on the target PAM context. Nature Communications, 13, [3006]. https://doi.org/10.1038/s41467-022-30515-0
CRISPR/Cas9 gRNA activity depends on free energy changes and on the target PAM context

Giulia I. Corsi, Kunli Qu, Ferhat Alkan, Xiaoguang Pan, Yonglun Luo & Jan Gorodkin

A major challenge of CRISPR/Cas9-mediated genome engineering is that not all guide RNAs (gRNAs) cleave the DNA efficiently. Although the heterogeneity of gRNA activity is well recognized, the current understanding of how CRISPR/Cas9 activity is regulated remains incomplete. Here, we identify a sweet spot range of binding free energy change for optimal efficiency which largely explains why gRNAs display changes in efficiency at on- and off-target sites, including why gRNAs can cleave an off-target with higher efficiency than the on-target. Using an energy-based model, we show that local gRNA-DNA interactions resulting from Cas9 “sliding” on overlapping protospacer adjacent motifs (PAMs) profoundly impact gRNA activities. Combining the effects of local sliding for a given PAM context with global off-targets allows us to better identify highly specific, and thus efficient, gRNAs. We validate the effects of local sliding on gRNA efficiency using both public data and in-house data generated by measuring SpCas9 cleavage efficiency at 1024 sites designed to cover all possible combinations of 4-nt PAM and context sequences of 4 gRNAs. Our results provide insights into the mechanisms of Cas9-PAM compatibility and cleavage activation, underlining the importance of accounting for local sliding in gRNA design.
The bacterial CRISPR (clustered regularly interspaced short palindromic repeats)-Cas9 endonuclease has been transformed into a powerful genome-editing tool that is nowadays broadly applied in biological, agricultural and medical research. Cleavage by the widely studied Streptococcus pyogenes Cas9 ortholog (SpCas9), hereafter referred to as Cas9, is mediated by a 20-nt segment of a guide RNA (gRNA) complementary to a target DNA sequence preceding a protospacer adjacent motif (PAM), where the Cas9 is recruited. The canonical PAM of SpCas9 is 5′-3′ “NGG”. In this study we refer to this canonical PAM sequence as 5′-NGN1-3′, thus including both the GG binding motif and its 1-bp flanking sequence context (N−1 and N+1) in the definition of PAM. Once a stable gRNA–DNA heteroduplex is formed, double-strand breaks (DSBs) are produced on the DNA by the Cas9 nuclease domains. In the case of the most broadly used SpCas9 protein, DSBs are introduced by the HNH and RuvC nuclease domains 3 nt upstream from the PAM.

A large number of in silico methods for gRNA design report that the cleavage efficiency of Cas9 can largely vary due to the sequence- and structural properties of the gRNAs. However, despite the immense activity in the field, a comprehensive analysis linking gRNA binding patterns, free binding energy changes and unfolding free energy changes to cleavage efficiency is surprisingly still lacking.

A major advance in the understanding of Cas9-gRNA binding was given by Globyte et al., who revealed that DNA interrogation by Cas9 occurs not only by random collisions with the DNA, as previously reported, but also by lateral diffusion. During this process, Cas9 ‘slides’ along the DNA in a local region (≈20 nt) as part of its search for a target on which the Cas9-gRNA complex can bind firmly. The study of Globyte et al. was, however, limited to the Cas9 sliding dynamics on short-distanced PAMs. Hence, the effect of Cas9 binding at sites overlapping the on-target PAM on cleavage activity remains unexplained.

Previously, we introduced the first energy-based model for Cas9-gRNA-target binding. We applied it to predict Cas9 off-target and, recently, on-target activities, improving over other available methods. Here, we employ this energy-based model to systematically analyse the relationship between nucleotide bindings and cleavage potential on a large dataset of 11,602 experimentally validated gRNA efficiencies. Next, this energy-based model is applied to explain the cleavage activity of Cas9 at gRNA bindings that originate from local overlapping PAMs on which Cas9 slides (local sliding PAMs). Introducing local sliding PAMs in the computation of the CRISPRspec gRNA specificity score, which is a ‘competition score’ that measures the ability of Cas9 to bind at the on-target while accounting for possible off-targets in the genome, leads to a better identification of gRNAs with high efficiency and low off-target potential. We validated the effect of local sliding PAMs on the cleavage efficiency with public data and in-house generated data from HEK293T cells by altering the sequence and context of Cas9 binding sites at the targets of four gRNAs. Our results further show that local sliding broadens the PAM compatibility of both wild-type and Cas9-derived variants.

Results

Cas9 on-target efficiency modelled with binding free energy changes. Modelling gRNA–DNA bindings in terms of free energy changes is a convenient approach that can be applied to evaluate both on-targets and off-targets with mismatches or bulges. Sufficient gRNA-target-DNA complementarity is essential to trigger the HNH conformational changes that activate DNA cleavage by both HNH and RuvC. To test if binding free energy changes can explain this activation we consider the contributions from our previously established CRISPR/Cas9 binding model: 

\[ \Delta G = \delta_{\text{PAM}}(\Delta G_{\text{H}} - \Delta G_{\text{U}} - \Delta G_{\text{O}}) \] 

with \( \delta_{\text{PAM}} = 1 \) for 5′-NGGN1-3′ PAMs. We analysed the relationship between binding free energy changes and on-target cleavage efficiency, measured as indel frequency, in a dataset of 11,602 Cas9 gRNAs binding at targets with 5′-NGN1-3′ PAMs from the datasets of Kim et al. and Xiang et al., merged following our previously established protocol. During the pre-processing of the original 23,902 gRNAs, instances with low specificity were removed to avoid biases due to global off-targets competition, and part of the data was isolated as an independent test set for later usage, accounting for data similarity (dataset ‘Xiang 2021 test’, see Methods and Supplementary Table 1). The gRNA–DNA hybridisation free energy change (\( \Delta G_{\text{H}} \)) is calculated by using stacked gRNA–DNA base pairs weighted based on biochemical profiling of nuclease-dead Cas9 binding kinetics. The \( \Delta G_{\text{H}} \) of highly efficient gRNAs is mostly confined in a sweet spot narrow interval ranging between −64.53 and −47.09 kcal/mol, while more extreme values are observed for inefficient gRNAs (Fig. 1a and Supplementary Table 2). The relation of \( \Delta G_{\text{H}} \) with the cleavage efficiency is substantially more profound compared to that of the GC-content, despite the high correlation between these two properties (Supplementary Fig. 1). A similar but less pronounced trend emerges for the target DNA–DNA opening free energy change, \( \Delta G_{\text{O}} \), regarded as a penalty for unwinding the DNA (Fig. 1a and Supplementary Table 3). The gRNA self-folding free energy change, \( \Delta G_{\text{G}} \), is included in the binding model as a penalty for unfolding gRNA structures, a process required for subsequent target recognition. Our results show that more stable gRNA self-folding structures negatively affect cleavage activity (Fig. 1a and Supplementary Table 4), complementing previous observations in a large-scale scenario. Before binding to Cas9, the scaffold sequence attached to a gRNA can interact with the bases in the gRNA sequence. This can disrupt the optimal structure of the scaffold, whose correct folding is required to form a complex with Cas9. Hence, spacers that interact with the scaffold were removed during data pre-processing whenever the structure of the full gRNA, spacer and scaffold, displayed reduced accessibility of the loops and bulges necessary to bind with Cas9 (see Supplementary Fig. 2 and Methods).

The \( \Delta G_{\text{B}} \) binding free energy change combines \( \Delta G_{\text{H}} \), \( \Delta G_{\text{O}} \) and \( \Delta G_{\text{G}} \) to estimate the residual gRNA–DNA binding free energy change after accounting for the DNA unwinding and gRNA unfolding penalties. The \( \Delta G_{\text{B}} \) significantly differs between low and high-efficient gRNAs, with the latter obtaining greater benefit in terms of free energy change by binding to the target (Fig. 1a and Supplementary Table 5). However, despite having extremely low (favourable) \( \Delta G_{\text{B}} \) gRNAs highly rich in GC content remain disadvantageous.

The \( \Delta G_{\text{H}} \) was further detailed in the position-specific free energy change contributions of stacking gRNA–DNA base pairs and this free energy profile was compared with a sequence logo highlighting the positional nucleotide preferences of highly efficient gRNAs. We observe that the 3′-1 seed region of highly efficient gRNAs is characterised by more stable interactions (lower free energy change) with the DNA. When ignoring the estimated impact of Cas9 as the weight on the stacking free energy changes, we notice bigger discrepancies in the 5′ part of the gRNA despite the general lack of sequence variation (Supplementary Fig. 3 and Fig. 1b), supporting that these weights are important in the energy-based model. This profile of free energy changes is consistent with the notion that the differences between efficient and inefficient gRNAs are attributable to the actual binding properties near the PAM, which otherwise are ‘only’ observable as position-wise single nucleotide variations. The preferred nucleotides in the 3′ seed end of highly
efficient gRNAs are guanine at N19-N20 and cytosine at N18-N19, where NX refers to the position in the spacer from the 5′ end. The aversion to uracil (U) at the gRNA 3′ seed end was previously explained as a transcription deficiency, as multiple Ts on the DNA, combined with the downstream T-rich sequence in the DNA sequence of the gRNA scaffold, might trigger transcription termination by polymerase III\textsuperscript{27}. Considering that stacking base pairs containing uracil present the lowest binding free energy change benefit (Supplementary Table 6) an additional explanation for this negative effect is possibly the poor hybridisation stability of U-rich gRNA seeds. Supporting this, the presence of up to 2 Us at the gRNA 3′ seed end, not
contiguous to the gRNA scaffold and not sufficient to trigger polymerase III termination, is also linked to low efficiency (Supplementary Fig. 4). Altogether, these results reveal that binding and folding interactions between nucleic acids have a significant role in Cas9 cleavage activation, which requires energetically favourable interactions with tight binding at the gRNA 3′ seed region.

Low gRNA-target hybridisation free energy change results in increased activity at off-targets. In a study investigating the impact of bulges in gRNA–DNA interactions, Lin et al. revealed that DNA bulges are better tolerated at off-target sites with high GC content and that bulged off-targets can even surpass fully complementary interactions in terms of efficiency. To demonstrate this, Lin et al. tested the indel frequency of four gRNAs with different GC content after creating DNA bulges at each position of the target sequences by systematically removing one base at a time in the gRNAs. To explain the efficiency increase recorded at bulged interactions, we examined the affinity between the key free energy change component $\Delta G_{GH}$ and the sweet spot. The remaining two components, $\Delta G_{GU}$ and $\Delta G_{GO}$, presented only minor variations. There were 19 gRNA-target bindings with at least 10% cleavage frequency, which we analysed, i.e., interactions with DNA bulges present at positions at which they are tolerated. The gRNA with the highest GC content (100%) has an indel frequency of 30% at the fully complementary target, and the $\Delta G_{GH}$ of this binding (~66.85 kcal/mol) falls outside of the sweet spot. Strikingly, six bulged off-targets of this gRNA have increased efficiency (indel frequency = 44.77 ± 8.33%) and in five of these the free energy change penalty caused by the bulge (between +6.58 and +15.38 kcal/mol) is sufficient to enter the preferential $\Delta G_{GH}$ range (Supplementary Table 7). On the contrary, a gRNA with medium GC content (~50%) and $\Delta G_{GH}$ in the sweet spot exhibit lower efficiency in the presence of DNA bulges that push $\Delta G_{GH}$ outside of the desired range (Supplementary Table 7). Nothing can be stated for the remaining two gRNAs in the dataset, with 65 and 35% GC content. The former maintains its position in the range of preferential $\Delta G_{GH}$ even in the presence of bulges, which reduce the cleavage efficiency. The latter is instead too unstable to tolerate any bulge (high $\Delta G_{GH}$).

Inspired by this result, we analysed the GUIDE-seq off-target dataset of Tsai et al. to understand if mismatches can favour the binding of gRNAs with high GC content by shifting $\Delta G_{GH}$ into the sweet spot. Notably, the four gRNAs with the highest GC-content in the dataset (67–80%) have off-targets with mismatches that surpass the on-target in terms of cleavage activity, measured as GUIDE-seq read counts. We focused our analysis on off-targets with no 3-nt PAM (NGG) variation to the on-target and with up to three mismatches to the gRNA, none of which should be in the four PAM-proximal nucleotides, as this could interfere with sequence composition preferences (Fig. 1b). On- and off-target sites (Fig. 2a) were classified by their cleavage activity compared to that of their on-target and by the number of GUIDE-seq read counts, setting a threshold of 300 as low cleavages (Fig. 2b). Of the four on-target sites, three are within the preferential $\Delta G_{GH}$ range, while all four off-targets with higher efficiency compared to the on-target reside in the $\Delta G_{GH}$ sweet spot. These off-targets are better centered in the preferential $\Delta G_{GH}$ range compared to the on-targets, except for the off-target of the gRNA ‘VEGFA site 1 tru-gRNA’. This is a shortened 18-nt gRNA, and as such, it is expected to have higher $\Delta G_{GH}$ (less stable binding) compared to the 20-nt gRNAs employed in the rest of the analyses. Instead, all except one of the 18 off-targets with efficiency lower than the on-target fall out of the sweet spot due to their increased $\Delta G_{GH}$ (Fig. 2c). Notably, the efficiency at these off-target sites cannot be explained by the GC content of the matching bases at the target site (Fig. 2d). This shows that the position-specific weighted evaluation of matching and mismatching stacking interactions of $\Delta G_{GH}$ gives a major benefit in the assessment of gRNA–DNA binding potential compared to a mere nucleotide content calculation.

These results suggest that the sweet spot affinity is a suitable criterion for defining cleavage efficiency at target sites with perfect complementarity, while the combination of the sweet spot range with bulges and mismatches interestingly explains why some off-targets are more efficient than their corresponding on-targets.

The impact of local sliding over canonical PAMs on cleavage efficiency. Based on the finding that Cas9 can search for PAMs by sliding on the DNA, we devised an energy-based model to explain the cleavage activity of Cas9 binding at sites with GG motifs overlapping the on-target PAM (Fig. 3a). We reasoned that Cas9 can bind at juxtaposed upstream PAMs forming a gRNA–DNA interaction with a bulge on the first PAM-proximal gRNA nucleotide, and spontaneously slide towards the intended on-target PAM to maximise gRNA–DNA complementarity and increase stability. Conversely, binding at downstream PAMs can anchor the complex in sub-optimal configurations with a DNA bulge between the gRNA–DNA hybrid and the PAM. Except for this single bulge, the hybrid is tightly coupled by 20 base pairs and can thus prevent Cas9 from sliding away or dissociating from the PAM, while keeping the site inaccessible to other Cas9 complexes. However, because a single PAM-proximal DNA bulge can be tolerated, cleavage cannot be excluded at these sub-optimal bindings. Considering the flexibility of nucleic acid bindings, we also do not exclude the possibility of sliding from the on-target PAM to an adjacent downstream PAM at which the gRNA and DNA remain fully bound (Fig. 3a). In line with this model, previous studies have shown that the presence of guanine immediately downstream from the on-target PAM is unfavourable for cleavage efficiency. In the merged dataset of Xiang et al. used to study free energy change properties (of 11,602 gRNAs), the mean efficiency at targets with a downstream PAM (DNA
bulge) is 12.64% lower compared to the efficiency at targets with isolated non-overlapping PAMs (Fig. 3b). Instead, a mean increase in efficiency of 7.24% characterises targets with upstream PAMs (gRNA bulge). Finally, targets with both down- and upstream sliding PAMs display a percentage decrease in mean efficiency of 3.96%. These observations show that the context of Cas9 binding sites significantly impacts gRNA efficiency.

Multiple overlapping binding sites distanced more than 1 nt from the on-target one while being interspaced from it only by guanines are rare in the dataset (Supplementary Fig. 5). We analysed adjacent PAMs created by G stretches up to 3 nt upstream or downstream of the on-target PAM binding site. The limit of 3 nt is imposed by the size of the target surrogates used to evaluate efficiency, which are integrated randomly in the genome and have unknown context. Multiple downstream sliding PAMs do not penalise gRNA efficiency to a greater extent than single ones. Instead, multiple upstream sliding PAMs are more favourable than single ones (Supplementary Fig. 5). However, these involve position $-2$ or $-3$ from the on-target binding site (N19, N20), which are part of the seed region, and therefore the increase in efficiency may be related to the nucleotide preferences in the seed in addition to the sliding effects. This bias is not present at sliding PAMs immediately upstream from the on-target site (position $-1$), which do not overlap the gRNA-target. Therefore, in the following, we focus on overlapping PAMs with up to 1 nt from the on-target Cas9 binding site.

**Fig. 2 Cleavage activity at off-targets modelled with binding free energy changes.**

- **a** Sequences and properties of off-targets identified by GUIDE-seq for four gRNAs that have at least one off-target cleaved more efficiently than the on-target. The sequence of the on-target site is reported on top of a list of the profiles of targets detected by GUIDE-seq. Each sequence in the listed targets is shown as follows: dots indicate bases that are the same as the on-target site, while differences between the on-target sequence are displayed with coloured nucleotides. Off-targets were filtered to avoid sites with >3 mismatches to the gRNA, or mismatches located in the four PAM-proximal nucleotides or the PAM itself.

- **b** Histogram of GUIDE-seq read counts at target sites in the full dataset of Tsai et al. A red vertical bar indicates the arbitrary threshold of 300, used to distinguish sites with low off-target activity.

- **c** gRNA–DNA hybridisation free energy change $\Delta G_h$ of target sites grouped by cleavage activity. The sweet spot preferential range of hybridisation free energy change $\Delta G_h$, which contains 80% of the highly efficient gRNAs (top 20% efficient) in the merged dataset of Xiang et al., used for free energy change profiling, is highlighted in grey.

- **d** Percentage of G and C bases at target positions complementary to the gRNA. Target sites are grouped by cleavage activity. The preferential range of GC%, defined as in (c), is highlighted in grey. Source data are provided as a Source Data file.
Local sliding PAMs allow gRNAs to bind with their targets forming bulged interactions. Bulged bindings are often not accounted for by off-target scoring tools, as their computation is highly demanding. However, the bulged interactions at local sliding PAMs are quick to identify and evaluate, and because of their direct accounting for bulged interactions at local sliding PAMs as pseudo-off-targets, only the most favourable one (lowest CRISPRspec) is used. Given a gRNA the effect of local sliding PAMs is integrated by adding to CRISPRspec the gain or penalty in efficiency attributable to the local sliding. To do so, CRISPRspec is firstly linearly re-scaled to the efficiency (see Methods). The parameters that define the local sliding addends are obtained by linearly fitting the deviation of the ∆G_B measured at a local sliding PAM from the median ∆G_B computed at all local sliding PAMs of the same type in the dataset (up- or down-stream) to the efficiency. The deviation is used as extreme ∆G_B is unfavourable (Supplementary Fig. 6). To fit the linear model, the merged dataset of Xiang et al. (excluding the test portion) is used with no prior filter on CRISPRspec (14,981 gRNAs, see Supplementary Table 1). The CRISPRspec score extended with local sliding PAMs is integrated into our previously defined CRISPRspec global gRNA specificity score²⁰. The higher the CRISPRspec score, the less binding competition (off-targets) affects a gRNA. In general, a score above 5 empirically delineates low off-target potential²⁰. In brief, CRISPRspec is calculated as −log₁₀ of the fraction between the sum of the Boltzmann-weighted ∆G_B computed at all targets in the genome with up to a certain number of mismatches excluding (numerator) or including (denominator) the on-target²⁰. Among competing PAMs, only the most favourable one (lowest ∆G_B) is used. Given a gRNA the effect of local sliding PAMs is integrated by adding to CRISPRspec the gain or penalty in efficiency attributable to the local sliding.
gRNA efficiency before and after the addition of local sliding considerations on two independent test sets. Only gRNAs (20mer) with >3 nt differences to those used for fitting were used (6361 gRNAs for ‘Xiang 2021 test’, 4026 gRNAs in ‘Hart 2015’, see Supplementary Table 1). To focus the test on the impact of sliding, the evaluation was performed on gRNAs targeting sites where sliding is possible, i.e., with a downstream or an upstream PAM (2219 gRNAs for ‘Xiang 2021 test’, 1241 gRNAs in ‘Hart 2015’). The gRNAs were partitioned into three equally sized sets of low, medium and high CRISPRspec or CRISPRspecExt. Binning values were estimated on the dataset employed for parameter fitting. In ‘Xiang 2021 test’ the efficiencies of gRNAs with CRISPRspec medium or high are not different prior to the inclusion of local sliding (Fig. 3c). Other groups, instead, are significantly distinct. After introducing local sliding, the separation between the three groups becomes more profound and all three present significant differences in terms of efficiency (Fig. 3c). On the external independent test set ‘Hart 2015’, CRISPRspec does not significantly separate gRNAs with different efficiencies, while this task is accomplished by CRISPRspecExt. This analysis shows that CRISPRspecExt can isolate well highly specific, and thus efficient, gRNAs. The positive correlation between our competition score CRISPRspecExt and gRNA efficiency in the ‘Xiang 2021 test’ and ‘Hart 2015’ test sets (Pearson’s r = 0.29 and r = 0.26 respectively) is more substantial compared to that of other relevant properties taken individually, such as the AGG and the GC content (r = −0.26 and r = 0.15 respectively in ‘Xiang 2021 test’, no significant correlation in ‘Hart 2015’, Supplementary Fig. 8). These results support the concept that strong binding competition at local and/or global off-target sites impairs cleavage activity.

Validation of sliding effects at canonical PAMs in HEK293T cells. As revealed from the analysis above, the gRNA efficiency is affected by the context flanking the on-target PAM binding site. To further validate this effect, we measured the efficiency of four gRNAs for the genes PAX5, BMP3, ADRA2C and CHRNA2. These gRNAs displayed high cleavage efficiency in our previous surrogate-based evaluation of gRNAs and exhibited no trace of enrichment or depletion in the edited cells17 (see Methods). For each gRNA, target DNA surrogate sites were designed to carry all possible 16 variations of N_1 and N_1 in the 5’-N_1GGN_1-3’ PAM (Supplementary Table 8). The efficiency was measured as indel frequency 6 and 10 days after lentiviral transduction in HEK293T cells, following our previously established protocol17 (see Methods). The efficiencies obtained at the two time points were strongly correlated (Pearson’s r = 0.98, Spearman’s R = 0.79) and were therefore averaged (Supplementary Fig. 9 and Supplementary Data 1). In agreement with the local PAM-sliding model presented above, the cleavage at targets with G_1GGH_1 and H_1GGG_1 PAMs displays, in this order, percentage increase and decrease in mean efficiency of 11.31 and 12.13% compared to sites with H_1GGH_1 PAM, where H symbolises any of A, T or C. Clustering targets based on the N_1N_1 context produces an exclusive cluster for G_1H_1, which comprises the top three contexts with highest mean efficiency across the four gRNAs, while the lowest 3 belong to the H_1G_1 contexts (Fig. 4a).

If local sliding PAMs are not considered, CRISPRspec is constant for different PAM contexts of the gRNA targets. In contrast, CRISPRspecExt is positively correlated to the gRNA efficiency, thanks to the addition of the sliding PAM contributions (Fig. 4b). Thus, our energy-based binding model allows improved identification of highly specific and efficient gRNAs based on the context of their targets. The portion of the variance associated with the differences in the PAM binding site context explained by the sliding activity on adjacent canonical ‘GG’ binding motifs was assessed as the sum of squares to the grand mean relative to each condition (sliding upstream, downstream, on both sides or on none) and weighted by the size of each group, divided by the total sum of squares (see Methods). The sliding activity explains >50% of the variance in the targets designed for the gRNAs ADRA2C (53.49%), PAX5 (52.44%) and BMP3 (50.58%) while this proportion is lower, 32.44%, in the case of gRNA CHRNA2. Other features of the target sites may be responsible for the remaining, unexplained, variance. For instance, the random integration of the surrogate sequences carrying the targets may impact their accessibility and, therefore, the ability of Cas9 to cleave these sites, which is a limitation of the current surrogate-based approaches17. In addition, the effect of the different context nucleotides on PAM-probing and the possible sliding activity of Cas9 on non-canonical PAMs, described in the next section, may also contribute to the remaining variance (i.e., binding and sliding at G_1GGH_1 may vary for different H_1). Additional data will be necessary to evaluate this feature and eventually include it in our sliding model.

Local sliding broadens non-canonical PAM compatibility. Cleavage by Cas9 can also verify at target sites flanked by non-canonical PAMs, of which N_1AGN_1 is reported as the most active10. To test if the concept of local sliding is applicable to non-canonical PAMs, we evaluated the efficiency, measured as indel frequency, of the same four gRNAs described above for target DNA sites bearing all possible variations of the PAM binding motif and its context. The activity of Cas9 at non-canonical PAMs is generally low, therefore we additionally measured gRNA efficiency in HEK293T cells with Doxycycline-induced overexpression of Cas9 (Dox+). The gRNA efficiencies measured in Dox+ and Dox− cells are well correlated (Supplementary Fig. 10). In untreated cells (Dox−) the mean efficiency at targets flanked by the canonical GG PAM binding site is 65.49% while at the non-canonical AG and GA it is respectively 21.18 and 14.93%. At other non-canonical binding sites, the mean efficiency is below 10% (TG = 5.29%, GC = 5.06%, CG = 3.69%, GT = 2.00%, AA = 1.12%, TA = 0.90%, CC = 0.88%, CA = 0.85%, TC = 0.81%, TT = 0.72%, AT = 0.71%, CT = 0.62%, AC = 0.61%) (Fig. 4c). In the Dox+ group (Supplementary Fig. 11), extremely high efficiency is obtained at bindings with a canonical PAM regardless of their context (mean = 96.79%). The efficiency at the non-canonical AG binding site in Dox+ reaches levels close to those of GG in the Dox− group (mean = 59.54%). The over-expression of Cas9 increases the efficiency at sites flanked by other non-canonical PAMs as well, such as those with binding sites GA, TG, GC and CG, with mean efficiencies of 41.50, 15.09, 12.75 and 8.86%, respectively. Cleavage at targets with other PAMs remains instead rare or null (mean efficiency <5% for all).

Among all possible alternative PAMs those with at least one G in the binding motif are the most efficient ones (Fig. 4c). These sites can incorporate local sliding PAMs with a canonical GG binding motif if a G is present in the context upstream (G_1GGN_1) or downstream (N_1GGG_1). Following Cas9 binding at such PAMs, gRNAs may form bulged interactions with their targets in lack or excess of 1 nt (Fig. 4d). Indeed, the cleavage efficiency measured in Dox+ HEK293T cells at alternative PAMs with a canonical GG binding motif upstream (G_1GGH_1 and G_1GTH_1) or downstream (H_1TGG_1 and H_1CGG_1) is significantly higher compared to that of targets with the same alternative PAM binding sites but no G in the N_1N_1 context (Fig. 4d, see also Supplementary Figs. 12–15 for results on all PAM binding sites in Dox− and Dox+). Alternative
PAMs with binding motif AG and GA, which are the closest to the canonical GG in terms of efficiency, do not display a preference for local sliding toward canonical PAMs. This suggests that fully complementary gRNA–DNA interactions at a non-canonical AG or GA are equally tolerated as bulged bindings at the canonical GG PAM binding site.

The efficiencies related to target sites with $G_{-1}AGH_{+1}$ PAMs in Dox+ cells are significantly more efficient (mean increase 33.45%) compared to $H_{-1}AGH_{+1}$ PAMs (Fig. 4e, Supplementary Fig. 14). This trend is similar to the one observed for the canonical $G_{-1}GGH_{+1}$ PAM in Dox−. $G_{-1}$ generates a non-canonical GA binding site, less stable than GG or AG. This could facilitate...
Fig. 4 Indel frequency of gRNAs at targets with different PAM sequences in HEK293T cells. a Heatmap of indel frequencies (Dox- cells) of four gRNAs targeting sites with 5′-N,GNN,−3′ PAMs clustered by Euclidean distance (missing values were linearly interpolated from targets with the same N,GNN context). Averages of columns and rows are indicated. Leaves in the dendrogram are colourised as in Fig. 3b. Pearson’s correlation between indel frequency of four gRNAs and their specificity extended with local sliding (CRISPRspExtx, blue colour). *p < 0.05, **p < 0.01, ns non-significant (gRNA PAX5 p = 7.05E-03, ADAR2C p = 0.02, CHRN8B2 p = 0.11, BMP3 p = 8.73E-3). The specificity score CRISPRspexc, not influenced by PAM contexts, is indicated with a dashed line. c Indel frequency (Dox- cells) of gRNAs binding at targets with different PAM binding sites (X-axis). The swarm plot details the indel frequencies of each gRNA. Boxes represent the first and third quartiles (Q1 and Q3). The median is shown as a line; whiskers extend up (or down) to the last (or first) value lower (or greater) than Q3+1.5*(Q3-Q1) (or Q1-1.5*(Q3-Q1)). Number of gRNAs per box, left to right: 59; 56; 53; 60; 54; 62; 60; 58; 61; 59; 62; 58; 57; 62; 63. d Indel frequency (Dox+ cells) at targets with alternative PAMs but canonical GG local sliding PAM upstream (top) or downstream (bottom). Boxes and swarm plots are defined as in (c). Binding contexts represented in less than three of four gRNAs or with G1 (top) or G1 (bottom) were excluded. See Supplementary Figs. 12-15 for results on all PAMs and cell treatments. One-sided t-test p values from left to right, top to bottom: 0.17; 9.72E-04; 1.69E-07; 0.94; 1.40E-03; 2.59E-07 (*p < 0.05, **p < 0.01, ***p < 0.001, ****p < 1E-04). The alternative hypothesis is that efficiencies at targets with G1 (top) or G1 (bottom) have a larger mean than, respectively, H1 or H1. Number of elements per box, from left to right, top to bottom: 30; 12; 35; 11; 36; 11; 35; 12; 36; 11; 35; 12. e Heatmap of indel frequencies (Dox+ cells) of four gRNAs at targets with 5′-N,AGN,−3′ PAMs. See (a) for further details. Source data are provided as a Source Data file.

Evidence for local sliding in Cas9 variants. To corroborate our observations on the local sliding activity of Cas9, we analysed the data from Kim et al.16 which measured indel frequencies of gRNAs using 13 SpCas9 variants16, consisting of the wild-type SpCas9, five high-fidelity variants (eSpCas9(1.1)31, SpCas9-HF132, HypaCas933, evoCas934 and Sniper-Cas935), five variants with altered PAM compatibility (VQR36, VRER36, VRQR32, QRQR37 and SpCas9-NG38), and two variants with both such properties (VRQR-HF132 and xCas939). We and others previously reported that the data presented in Kim et al.16, in which a modified gRNA scaffold is employed, is not entirely compatible with that of similar recent studies17,40. A closer look into the library dedicated to the analysis of PAM contexts (for which cleavage at surrogate targets followed by all combinations of 4-nt PAMs was examined for 30 gRNAs) reveals that in the case of SpCas9, gRNAs with a U at the 3′ seed end (U20) are more efficient (indel frequency mean ± std = 52.46 ± 5.44%, n = 110 targets of seven gRNAs) compared to gRNAs with a C20 (mean ± std = 46.23 ± 6.90%, n = 239 targets of 15 gRNAs) or an A20 (mean ± std = 47.44 ± 7.64%, n = 121 targets of eight gRNAs) (two-sided t-test p value C20 compared to U20 = 1.66E-15, A20 compared to U20 = 3.61E-08). Of note, no gRNA with a G20 is present in the dataset. The higher efficiency observed in gRNAs with a U20 is in contrast with our results of nucleotide preferences based on the merged data from Kim et al.14 and Xiang et al.17 (Fig. 1b), as well as with other independent reports48,15. The nucleotide N20 constitutes the gRNA bulge in the upstream sliding and thus plays a central role in the analysis of sliding dynamics. Because of this incongruence, we do not validate local sliding on the canonical PAM using the Kim et al.16 database. The PAM preferences in such dataset are, instead, as expected, with GG, AG and GA being the preferred binding sites (in this order), and other PAM binding sites showing little or no activity (Supplementary Fig. 17). As the efficiency linked to different PAM binding sites is much higher compared to that of sliding dynamics on canonical PAMs, we can reliably use the dataset from Kim et al.16 to further explore our hypothesis on Cas9 sliding from non-canonical to adjacent canonical PAMs. In the case of the wild-type SpCas9, targets followed by alternative PAMs G1,GCH1, G1,GTH1, are cleaved more efficiently compared to targets with the same alternative binding site but no G at position −1 (Fig. 5a). Similarly, targets followed by alternative PAMs H1,TGG1 and H1,CGG1 are cleaved more efficiently than H1,TG1 and H1,CG1 (Fig. 5a). No increase in efficiency is observed for the PAMs G1,GAH1 and H1,AGG1 compared to H1,GAH1 and H1,AGG1, due to the higher efficiency at these alternative PAM binding sites (Fig. 5a). These results are congruent with those obtained with the data we generated. Moreover, we examined high-fidelity Cas9 variants that do not contain variations in the PAM-recognition domains after sorting them by efficiency and inverse specificity, previously assessed16. The same preferences for up- and downstream sliding toward canonical PAMs are visible in Sniper-Cas9, which is the variant closest to the wild-type in terms of efficiency and specificity, and gradually disappear in more specific variants (eSpCas9(1.1), SpCas9-HF1, HypaCas9), less tolerant to imperfect gRNA-site. The possible gRNA–DNA hybrid formed at this site contains a 1-nt gRNA bulge and therefore has limited efficiency observed in gRNAs with imperfect PAMs, while H1,TGG1 and H1,CGG1 are cleaved more efficiently than H1,TG1 and H1,CG1 (Fig. 5a). No increase in efficiency is observed for the PAMs G1,GAH1 and H1,AGG1 compared to H1,GAH1 and H1,AGG1, due to the higher efficiency at these alternative PAM binding sites (Fig. 5a). This results are congruent with those obtained with the data we generated. Furthermore, we analysed the efficiency of the Cas9 variant SpCas9-NG, which shows good compatibility with N1-GNN1, PAMs, and xCas9, which has increased binding specificity and tolerates N1,GNN1, PAMs, although N1,GNN1 remains highly preferred16. In these variants, upstream sliding is expected at G1,HHH1-PAMs, while H1,HGG1 and H1,HHG1 may result in downstream sliding of, respectively, 1 or 2 nucleotides (Fig. 5b). Consistently with our model, targets followed by GHGGH are cleaved more efficiently than those followed by HHHH, in both variants (Fig. 5b top). Downstream sliding shows increased efficiency at targets with either of the PAMs HHGH and HHGH compared to HHHH in SpCas9-NG, while in xCas9 only HHGH is more efficient than HHHH (Fig. 5b bottom). This is expected considering the increased specificity of xCas9, that may tolerate one but not two DNA bulges between the gRNA–DNA binding and the xCas9 PAM binding site. The consistent efficiency increase observed at PAMs that allow for up- or downstream sliding to canonical binding sites and the
Discussion

In this work, we describe CRISPR/Cas9 cleavage as an energy-driven process in which efficiency significantly depends on nucleotide hybridisation and folding free energy changes. Our analysis of the free energy change preferences of efficient and inefficient gRNAs underlines the importance of designing gRNAs with low self-folding stability, balanced hybridisation of free energy change and firm target binding at the seed region. In regard to the ongoing discussion on whether Cas9 binding can be modelled in equilibrium, here we show that this is the case. Notably, our sweet spot of gRNA–DNA hybridisation free energy change provides an explanation for the higher cleavage activity previously measured at bulged or imperfectly matching off-targets compared to on-targets. Recent studies report that the mismatch tolerance patterns of different gRNAs can highly vary, and that sequence composition features such as enrichment of guanine and depletion of uracil characterise gRNAs with high guide-intrinsic mismatch permissiveness. Our results suggest that the fitness to the sweet spot and the free energy changes related to imperfect matches to the gRNA are suitable criteria to evaluate guide-intrinsic mismatch tolerance. By expanding the concept of gRNA-target-DNA hybridisation with the inclusion of imperfect matches containing gRNA or DNA bulges, we explain the activity of gRNAs at targets with adjacent PAMs on which Cas9 can locally slide. In the merged dataset of Xiang et al. used to examine
binding free energy change properties, such local sliding activity affects ~35% of the gRNAs that have a target site in hg38. We show that sliding activities can affect cleavage efficiency both positively and negatively, depending on the binding free energy change at bulged interactions. The inclusion of local sliding effects, which can be interpreted as local off-targets, improves the definition of gRNA specificity, previously based solely on global off-targets. Although interactions at sliding PAMs can result in DNA cleavage and produce indels near the target site, fulfilling the editing purpose in knockout experiments, their activity near the on-target site can lead to undesired effects in applications that require Cas9 to be positioned precisely at the target, such as in base editing. We validate the effect of local sliding on gRNA cleavage efficiency by analysing independent data (Fig. 3b, c) and by generating efficiency data for 4 gRNAs targeting DNA sites that carry all possible alterations of the Cas9 binding site and its immediate context (Fig. 4 and Supplementary Figs. 12–16). Such evaluation of different PAMs and contexts for the same gRNAs eliminates the bias that would derive from measuring the effect of contexts and PAMs using distinct gRNAs. Furthermore, we reveal that cleavage can be obtained more efficiently at targets followed by non-canonical PAMs whenever sliding on adjacent canonical PAMs is possible. This result, supported by our data and by a recent gRNA efficiency data from wild-type and engineered Cas9 variants, implies that local sliding broadens Cas9-PAM compatibility. This can be highly useful in those situations in which the choice of sites to be targeted for cleavage is limited (e.g., knock-in of genomic variants). While our study is centred on Cas9-gRNA interactions, the concepts we describe can be extensively applied in relation to other RNA-guided complexes. In conclusion, the local context of Cas9 binding sites can strongly impact Cas9 binding and cleavage efficiency, and this can largely be explained by our binding energy-based model. We also show that the assessment of gRNA specificity is enhanced once local sliding PAMs are considered and that this helps to identify gRNAs with higher specificity and efficiency.

Methods

Cell culture. Human embryonic kidney cells (HEK293T, originally purchased from ATCC catalogue num. CRL-3216) were cultured in DMEM media containing 10% foetal bovine serum (FBS) and 1% penicillin-streptomycin in a tissue culture incubator at 37 °C with 5% CO2. PCR mycoplasma detection kit (catalogue num. PM008, Shanghai Yise Medical Technology) was routinely used to test the mycoplasma contamination. The cells used in this study gave negative results in the mycoplasma contamination test. SpCas9-expressing HEK293T (HEK293T–SpCas9) cells were generated by a PiggyBac transposon system followed by selection in the presence of 50 µg/ml hygromycin to ensure high Cas9 activity. HEK293T cells were generated by a PiggyBac transposon system followed by selection in the presence of 50 µg/ml hygromycin to ensure high Cas9 activity. HEK293T cells were transiently transduced with pPB-TRE-spCas9-Hygrovector and pGMV-hybase vector with a 9:1 ratio to generate SpCas9-expressing HEK293T.

Surrogate library design and plasmid library preparation. To construct the gRNA-target-DNA surrogate library we selected four gRNAs with indel efficiency >90% in our previous CRISPRon chip (Dox+) and designed complementary targets flanked by 4-nm amplicons sites variated in all possible ways (PAM=NNNNN). This resulted in 4 × 256 gRNA-target-DNA surrogates, plus 10 positive sequence surrogates. The total number of sequences we generated was 1062. All the sequences were amplified by PCR, subjected to PCR using BSA restriction enzymes followed by selection in the presence of 50 µg/ml hygromycin to ensure high Cas9 activity. HEK293T cells were transiently transduced with pPB-TRE-spCas9-Hygrovector and pGMV-hybase vector with a 9:1 ratio to generate SpCas9-expressing HEK293T.

PCR ampiclons from a surrogate library from cells. The genomic DNA was extracted with the phenol-chloroform method. To remove DNA contamination, the genomic DNA was digested with RNase A (OMEGA). Then the genomic DNA was subjected to PCR using PrimeSTAR polymerase (Takara, R045Q). The PCR primers were 5′-GGACAACATCATGCATGCAAT-3′ and 5′-ACTCCTTCTAAAGGACTCTAGTAG-3′. The PCR products were purified by 1.5% gel and mixed with equal amounts and deep sequenced. The amplicons were subjected to deep sequencing on the MiSeq sequencer (Illumina) platform. All the samples were subjected to pair-ended 150 bp deep-sequencing.

Data pre-processing. Raw sequencing reads were processed to obtain read counts (indel or total) for each surrogate target following our previously published method. Briefly, the method consists of the following passes: quality assessment and cleaning of raw reads with FastQC v.0.11.3 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), paired reads merge with FLASH v.1.2.12 alignment; to the reference library with BWA-MEM v.0.7.17 (http://bio-bwa.sourceforge.net/), and reads quantification by exploiting the aligned reads and identifying variations in the expected mapping pattern using pysam v.0.15.0 (https://github.com/pysam-developers/pysam). For each target site, g + gRNA (20 bp) + scaffold (82 bp) - barcode (10 bp) + GTT (terminal sequence) are guaranteed to remain unchanged when extracting reads at each random PAM site. To distinguish the PAM sites efficiently, we exploited the 10 bp barcode sequence at the beginning of each surrogate target sequence. Relative indel frequencies, expressed in percentage, were obtained separately from the read counts measured on day 6.

Lentivirus production and lentivirus titre quantification. The lentivirus was produced in HEK293T cells by co-transfection with packaging plasmids. Briefly, for lentivirus production in 10 cm dish, the DNA/PEI (Polyethylenimine Linear, MW 40000) mixture contains 13 µg plenti-TRAPseq vectors, 3 µg pRSV REV, 3.75 µg pMD.2 G, 13 µg pMDGP-Lg/p-RRE, 100 µg PEI 40000 solution (1 µg/ml in sterilised cell culture supernatant supplemented with opti-MEM without phenol red (Invitrogen)) to a final volume of 1 ml. The transfection mixture was pipetted up and down several times gently, and further incubated for keeping at room temperature (RT) for 20 min. Then the transfection complex was added to HEK293T cells in a 10-cm dish containing 10 ml of culture medium. After 48 h, the supernatant was harvested, filtered through a 0.45 µm filter, and polybrene solution (Sigma-Aldrich) was added to the crude virus to a final concentration of 8 µg/ml. The crude virus was aliquoted into 15 ml tubes (5 ml/tube) and stored in a –80 °C freezer. Vial titre was estimated by counting the number of GFP positive cells in the virus-treated population by flow cytometry (FCM) as follows: (1) HEK293T cells were split and seeded to a 24-well plate on day 1. Generally, gradient volumes of 5, 10, 20, 40, 80, and 160 µl of crude lentivirus were added to the cells, and each volume was assayed in duplicate (Supplementary Fig. 18); (2) On day 2, lentivirus transduction was conducted when cells were ~60–80% confluent. Therefore transduction, the last two wells of cells were detached using 0.05% EDTA-Trypsin to determine the total number of cells in one well (Ntr). Then the gradient volume of the crude virus was added to each well and swirled gently; (3) On day 3, we changed to a fresh medium; (4) On day 4, cells were harvested and washed twice in PBS. The cells were fixed in 4% formalin solution at RT for 30 min, then washed with PBS twice. FCM was performed using a BD LSRSortest™ cell analyzer and the FACSDiva v.9.0 software with at least 50,000 events collected for each sample in replicates. The FCM output data was analysed with the FCM analysis software NovoExpress v1.5.6. The percentage of GFP-positive cells, for all samples, was calculated as:

$$\% = \frac{\text{Num. GFP + positive cells}}{\text{Total cells}} \times 100$$

(1)

For accurate titre determination, there should be a linear relationship between the FCM-positive percentages and crude volume. The titre (transducing units (TU/mL)) was calculated according to the following formula in which V represents the crude volume (µl) used for initial transduction:

$$\text{TU} = \frac{\text{Total cells} \times \%}{1000}$$

(2)
and day 10 (Dox+ or Dox−) respectively as:

\[
\text{indel frequency} (\%) = \frac{\text{num reads with indels} \times 100}{\text{Num total reads}}
\]

The drop-out of our protocol was minimal, and we obtained the following total unique gRNA count for BMPs n = 253; ADRA2C n = 253; CHRNA2 n = 250; PAX5 n = 255. A threshold on the minimum number of total reads was defined by iteratively removing samples with less than n total reads (n from 0 to 200, step size of 5). The value of n with the lowest associated Spearman’s correlation p value between the indel frequencies measured on day 6 and day 10 was selected. This threshold was performed separately for each time point Dox+ or Dox−, for which the resulting minimum reads thresholds were 90 and 35, respectively. Indel frequencies from day 6 and day 10 were then averaged. This data are available as Supplementary Data 1. For each gRNA, multiple targets of one selected context were evaluated using different barcodes. These presented a mean normalised standard deviation of 0.03 and 0.11 in removal of single bases in the gRNAs. Sequences and efficiencies included the crRNA:tracrRNA fusion loop and three tail loops49. The comparison scaffold, computed by solely folding the scaffold with RNAfold. This structure consists in measuring the Euclidean distance between binding probabilities for the 5% sgRNAs with the highest distance to the optimal structure is less efficient than the on-target.

Collection and pre-processing of external datasets

Merced Xiang et al. dataset (2021). This dataset employed to study the relationship between binding free energy changes and on-target cleavage efficiency was obtained by merging the data of Kim et al.14 and Xiang et al.17 as previously described in Xiang et al.17. This dataset consists of 25,902 gRNA sequences and corresponding Cas9 efficiencies, measured as indel frequencies, at targets flanked by 5′-NGG, 3′-S PAMs. The specificity of gRNAs (CRISPRspec) was evaluated with the CRISPRprof pipeline v.1.1.220 on the human genome hg38. The following two gRNAs failed to be evaluated by the pipeline within 2 weeks due to the extreme saturated gRNAs. Next, the dataset was filtered to remove gRNAs likely to form sub-optimal structures with the scaffold. For this, pairwise binding probabilities of the bases in the sgRNAs (gRNAs + scaffold) were evaluated with RNAfold v.2.2.548. The sgRNAs were compared to the optimal secondary structure of the scaffold, computed by solely folding the scaffold with RNAfold. This structure includes the crRNA:tracrRNA fusion loop and three tail loops49. The comparison consisted in measuring the Euclidean distance between binding probabilities for pairs of bases that bind the optimal structure, or the probability to be unbound for those bases that do not pair with any other in the optimal structure (1 – probability of pairing with either of their bases). For each of the two distance measures, we tested if the 5% sgRNAs with the highest distance to the optimal structure is less efficient than other sgRNAs (Mann-Whitney one-sided test, see Supplementary Fig. 2). A significant result was obtained for the unpaired bases (p = 3.68E-15) while for paired bases the distance was not highly pronounced (p = 5.31E-02). Hence, we removed gRNAs with indel frequencies at target sites <2% (n = 539). The resulting 21,402 gRNAs were split into a training set (n = 14,981) and a test set (n = 6,421), keeping sequences with the same target in the 30mer (gRNA + target context) in the same subset. The analysis of the free energy change properties was carried out on the human genome hg38 from Ricearch2, which includes the pre-compiled file containing the array of 19 weights for stacking base-pair contributions defined in CRISPRprof29. All gRNA–DNA interactions were forced to end at the PAM. The optimal start position was searched by iteratively shortening the DNA target sequence by 1 nt from an initial length of 24 nt upstream of the PAM to a minimum of 2 nt. For targets at PAMs overlapping while sliding 1 nt upstream/downstream only up to 21 nt were effectively utilised by the algorithm to find optimal gRNA–DNA interactions. This number varies in the case of the data from Lin et al. and GUIDE-seq. The DNA–DNA opening free energy change at bindings with bulges or mismatches was calculated for the sequence stretch involved in the gRNA–DNA interaction, which may be longer or shorter compared to the fully complementary case. For the function for DNA–DNA binding free energy change calculation of CRISPRprof was used. In the data from Lin et al. and GUIDE-seq, in which gRNA–DNA bindings can have different lengths, out of all possible sub-optimal interactions the one with the lowest ΔG was considered the optimal one. For all RNA–RNA, RNA–DNA and DNA–DNA stacking interactions, free energy change parameters were obtained from Ricearch2, which includes parameters for RNA–DNA56,57, RNA–RNA58 and DNA–DNA59,61 stacking interactions, as previously described60. Missing parameters, such as those for stacked bulges for RNA–DNA interactions, were obtained by averaging the corresponding RNA–RNA and DNA–DNA parameters, as in previous studies60,62.

Integration of local sliding PAMs in CRISPRspec

Local sliding PAMs were included in the CRISPRspec score measuring binding competition by considering bindings flanking sliding PAMs as pseudo-off-targets and by including them in the partition function calculation of CRISPRspec. By linearly fitting to the efficiency, we estimate the weighted averages of bindings attributed to the local sliding PAMs and added it on top of the original CRISPRspec measure obtained from the CRISPRprof pipeline v.1.1.220. Off-targets with up to six mismatches in the human genome hg38 were searched with Ricearch2 (v.2.1.12) as explained above. Given the set S of all gRNAs, let Ssup and Sdown be the subsets of gRNAs in S that have respectively upstream and downstream local sliding PAMs. For a gRNA x, let M be the median function, [x] the absolute value of x, \(\Delta G_{\text{up}}[x]\) and \(\Delta G_{\text{down}}[x]\) the \(\Delta G\) free energy changes computed at up- and down-stream binding sites of x, and γ(x) the experimentally measured efficiency of x. Then, the extended CRISPRspec, CRISPRspecExt, of gRNA x takes the local sliding PAMs into account as follows:

\[
\text{CRISPRspecExt}[x] = a_0 + a_1 \text{CRISPRspec}[x] + \lambda_{\text{up}}[x] \left( \left| \frac{\Delta G_{\text{up}}[x]}{M \left( \text{CRISPRspec}[x], \forall i \in S_{\text{sup}} \right) } \right| \right) + \lambda_{\text{down}}[x] \left( \left| \frac{\Delta G_{\text{down}}[x]}{M \left( \text{CRISPRspec}[x], \forall i \in S_{\text{down}} \right) } \right| \right)
\]

With:

\[
\lambda_{\text{up}}[x] = \begin{cases} 0 & \text{if } x \notin S_{\text{sup}} \\ 1 & \text{if } x \in S_{\text{sup}} \end{cases}
\]

\[
\lambda_{\text{down}}[x] = \begin{cases} 0 & \text{if } x \notin S_{\text{down}} \\ 1 & \text{if } x \in S_{\text{down}} \end{cases}
\]
The optimal least squares coefficients \( \alpha_0, \alpha_1, \beta_0, \beta_1, \gamma_0, \gamma_1 \) were estimated as:

\[
\alpha_0, \alpha_1 = \arg\min_{\alpha_0, \alpha_1} \sum_{i=3}^{7} \left( x_i - \alpha_0 - \alpha_1 \text{CRISPRRep}_i \right)^2
\]

\[
\beta_0, \beta_1 = \arg\min_{\beta_0, \beta_1} \sum_{i=3}^{7} \left( \left( \gamma_i - \beta_0 \right) \gamma_i - \beta_1 \text{CRISPRRep}_i \right)^2
\]

\[
\gamma_0, \gamma_1 = \arg\min_{\gamma_0, \gamma_1} \sum_{i=3}^{7} \left( \left( \gamma_i - \gamma_0 \right) \gamma_i - \gamma_1 \text{CRISPRRep}_i \right)^2
\]

Hence, parameters for local sliding PAMs are estimated from the absolute deviation of \( \text{AG} \) to the median \( \text{AG} \) of all local sliding PAMs of the same type (up/down) in the training set.

**Statistical analysis.** The SciPy v.1.5.0-1.6.3, NumPy v.1.18.5, pandas v.1.2, and scikit-learn v.0.23.2 modules were applied for data analysis in Python v.3.8.3. Plots were generated using Matplotlib v.3.2.2 and seaborn v.0.11.1. The impact of gRNA free energy change properties in defining gene knockout efficiency was analysed after sorting and separating gRNAs into two groups, high (top 20%) and low (bottom 20%) efficient, based on their reported indel frequencies (Supplementary Fig. 19). The intervals of preferential values identified for each free energy change property were set to contain 80% of the total highly efficient gRNAs. The portion of efficiency variance associated to the differences in the context of the ‘GG’ PAM binding site that could be explained by the sliding model was computed as the sum of squares (SSQsliding) between the efficiencies of each sliding condition and the grand mean (GM) divided by the total sum of squares between the efficiencies and the grand mean (SSQGM):

\[ \text{Explained variance} = \frac{\text{SSQsliding}}{\text{SSQGM}} \]

With:

\[ \text{SSQsliding} = \sum_{i=3}^{7} (x_i - \text{GM})^2 \]

\[ \text{SSQGM} = \sum_{i=3}^{7} n_i \text{VAR} \text{GM} \]

\[ \text{GM} = \frac{\sum_{i=3}^{7} n_i x_i}{\sum_{i=3}^{7} n_i} \]

Where \( S \) is the set of all gRNAs with a ‘GG’ PAM binding site, \( x_i \) is the efficiency of a gRNA, and \( M \) and \( n \) are the mean (M) of respectively the data subsets split by sliding categories: upstream (up); downstream (down); upstream and downstream (up/down); no sliding (none).

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

**Data availability**

The raw sequencing data generated in this study have been deposited in the NCBI Sequence Read Archive under accession codes PRJNA732366 and in the China National Genbank under accession code CRM000001405.38. Additional data used in this study are available as supplementary material in the following publications: dataset by Xiang et al.17 [https://doi.org/10.1038/s41467-019-0874-0], dataset by Kim et al.14 [https://doi.org/10.1126/sciadv.aax9249], dataset by Lin et al.29 [https://doi.org/10.1093/nar/gku402], dataset by Tzai et al.30 [https://doi.org/10.1038/s41588-019-0093-x], dataset by Kim et al.38 [https://doi.org/10.1038/s41587-020-0420-z]. Source data for the figures and supplementary figures are provided as a Source Data file. Source data are provided with this paper.

**References**

1. Doudna, J. A. & Charpentier, E. Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science* **346**, 1258096 (2014).
2. Jinek, M. et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
3. Huai, C. et al. Structural insights into DNA cleavage activation of CRISPR-Cas9 system. *Nat. Commun.* **8**, 1375 (2017).
4. Doudna, J. A. & Charpentier, E. Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Sci. Adv.* **1**, e1500088 (2015).
5. Wang, D. et al. Optimized CRISPR guide RNA design for two high-fidelity Cas9 variants by deep learning. *Nat. Commun.* **10**, 4284 (2019).
6. Xiang, X. et al. Enhancing CRISPR-Cas9 gRNA efficiency prediction by data integration and deep learning. *Nat. Commun.* **9**, 3328 (2018).
7. Sternglass, S. H., Redding, S., Jinek, M., Greene, E. C. & Doudna, J. A. RNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507**, 62–67 (2014).
8. Goh, B. Y., Lee, S. H., Bae, T., Kim, J. S. & Joo, C. CRISPR/Cas9 searches for a protosilencer adjacent motif by lateral diffusion. *EMBO J.* **38**, e99629 (2019).
9. Wang, D. et al. Optimized CRISPR guide RNA design for two high-fidelity Cas9 variants by deep learning. *Nat. Commun.* **10**, 4284 (2019).
10. Kim, N. et al. Prediction of the sequence-specific cleavage activity of Cas9 variants. *Nat. Biotechnol.* **38**, 1328–1336 (2020).
11. Xiang, X. et al. Enhancing CRISPR-Cas9 gRNA efficiency prediction by data integration and deep learning. *Nat. Commun.* **9**, 3328 (2018).
12. Sternglass, S. H., Redding, S., Jinek, M., Greene, E. C. & Doudna, J. A. RNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Science* **343**, 805–809 (2014).
13. Thyme, S. B., Akhmetova, L., Montague, T. G., Valen, E. & Schier, A. F. Internal guide RNA interactions interfere with Cas9-mediated cleavage. *Nat. Commun.* **7**, 11750 (2016).
14. Jensen, K. T. et al. Chromatin accessibility and guide sequence secondary structure affect CRISPR-Cas9 gene editing efficiency. *FEBS Lett.* **591**, 1892–1901 (2017).
15. Nishimasu, H. et al. Crystal structure of Cas9 in complex with guide RNA and DNA. *Cell* **156**, 935–949 (2014).
16. Wu, X. et al. Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. *Nat. Biotechnol.* **32**, 670–676 (2014).
17. Gao, Z., Herrera-Carrillo, E. & Berkhour, B. Determination of the exact transcription termination signal for type 3 polymerase III. *Mol. Ther. Nucleic Acids* **10**, 36–44 (2018).
18. 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).
19. Tzai, S. Q. et al. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat. Biotechnol.* **33**, 187–197 (2015).
20. Slaymaker, I. M. et al. Rationally engineered Cas9 nucleases with improved specificity. *Science* **351**, 84–88 (2016).
21. Kleinstiver, B. P. et al. High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects. *Science* **349**, 80–84 (2015).
22. Chen, J. S. et al. Enhanced proofreading governs CRISPR-Cas9 targeting accuracy. *Nature* **550**, 407–410 (2017).
34. Casini, A. et al. A highly specific SpCas9 variant is identified by in vivo screening in yeast. Nat. Biotechnol. 36, 265–271 (2018).
35. Lee, J. K. et al. Directed evolution of CRISPR-Cas9 to increase its specificity. Nat. Commun. 9, 3048 (2018).
36. Kleinstiver, B. P. et al. Engineered CRISPR-Cas9 nucleases with altered PAM specificities. Nature 523, 481–485 (2015).
37. Anders, C., Bargsten, K. & Jinek, M. Structural plasticity of PAM recognition by engineered variants of the RNA-guided endonuclease Cas9. Mol. Cell 61, 895–902 (2016).
38. Nishimatsu, H. et al. Engineered CRISPR-Cas9 nuclease with expanded targeting space. Science 361, 1259–1262 (2018).
39. Hu, J. H. et al. Evolved Cas9 variants with broad PAM compatibility and high DNA specificity. Nature 556, 57–63 (2018).
40. Moreb, E. A. & Lynch, M. D. Genome dependent Cas9/gRNA search time underlies sequence dependent gRNA activity. Nat. Commun. 12, 5034 (2021).
41. Klein, M., Eslami-Mossallam, B., Arroyo, D. G. & Depek, M. Hybridization kinetics explains CRISPR-Cas off-targeting rules. Cell Rep. 22, 1413–1423 (2018).
42. Huston, N. C. et al. Identification of guide-intrinsic determinants of Cas9 specificity. CRISPR J. 2, 172–185 (2019).
43. Fu, R. et al. Systematic decomposition of sequence determinants governing CRISPR/Cas9 specificity. Nat. Commun. 13, 474 (2022).
44. Chen, S.; Zhou, Y.; Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34, 1884–1890 (2018).
45. Magoc, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27, 2957–2963 (2011).
46. Heng, L. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at arXiv 1103.3997 (2013).
47. Li, H. et al. The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).
48. Lunter, G. & ViennaRNA package 2.0. Algorithms Mol. Biol. 6, 26 (2011).
49. Haeseler, M. & Concordet, J.-P. Genome editing with CRISPR-Cas9: can it get any better? J. Genet. Genomics 43, 239–250 (2016).
50. Hart, T. et al. High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. Cell 163, 1515–1526 (2015).
51. Haeseler, 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).
52. Frankish, A. et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 47, D766–D773 (2019).
53. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods 17, 261–272 (2020).
54. Alkan, F. et al. RSearch2: suffix array-based large-scale prediction of RNA–RNA interactions and siRNA off-targets. Nucleic Acids Res. 45, e60 (2017).
55. Wenzel, A., Abbaali, E. & Gorodkin, J. RSeach: fast RNA–RNA interaction search using a simplified nearest-neighbor energy model. Bioinformatics 28, 2738–2746 (2012).
56. John Santalucia, J. & Hicks, D. The thermodynamics of RNA structural motifs. Annu. Rev. Biophys. Biomol. Struct. 33, 415–440 (2004).
57. Allawi, H. T. & Santalucia, J. Thermodynamics and NMR of internal G·T mismatches in DNA. Biochemistry 36, 10581–10594 (1997).
58. Turner, D. H. & Mathews, D. H. NNDB: the nearest-neighbor energy model. Nucleic Acids Res. 38, D280–D282 (2009).
59. Watkins, N. E. Jr et al. Thermodynamic contributions of single internal rA·dA, rC·dG, rG·dG and rU·dT mismatches in RNA/DNA duplexes. Nucleic Acids Res. 39, 1894–1902 (2010).
60. Sugimoto, N., Nakano, M. & Nakano, S.-I. Thermodynamics—structure relationship of single mismatches in RNA/DNA duplexes. Biochemistry 39, 11270–11280 (2000).
61. Sugimoto, N. et al. Thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes. Biochemistry 34, 11211–11216 (1995).
62. Lorenz, R., Hofacker, I. L. & Bernhart, S. H. Folding RNA/DNA hybrid duplexes. Bioinformatics 28, 2530–2531 (2012).
63. Harris, C. R. et al. Array programming with NumPy. Nature 585, 357–362 (2020).
64. McKinney, W. Proc. 9th Python in Science Conference (SciPy, 2010).
65. Pedregosa, F. et al. Scikit-learn: machine learning in python. J. Mach. Learn. Res. 12, 2825–2830 (2011).
66. Van Rossum, G. & Drake, F. L. Python 3 Reference Manual (CreateSpace, 2009).
67. Gorodkin, J., Heyer, L. J., Brunak, S. & Stormo, G. D. Displaying the information contents of structural RNA alignments: the structure logos. Comput. Appl. Biosci. 13, 583–586 (1997).