BMC Bioinformatics

SOFTWARE

Open Access

GLiDe: a web-based genome-scale CRISPRi sgRNA design tool for prokaryotes

Tongjun Xiang[1,2], Huibao Feng[1,3*], Xin-hui Xing[1,4,5] and Chong Zhang[1,4*]

Tongjun Xiang and Huibao Feng are joint authors.

*Correspondence: fhb_14@163.com; chongzhang@tsinghua.edu.cn

[1] MOE Key Laboratory for Industrial Biocatalysis, Institute of Biochemical Engineering, Department of Chemical Engineering, Tsinghua University, Beijing 100084, China
[2] Department of Chemical and Biomolecular Engineering, University of California Los Angeles, Los Angeles, CA 90095, USA
[3] Division of Biology and Bioengineering, California Institute of Technology, Pasadena, CA 91125, USA
[4] Center for Synthetic and Systems Biology, Tsinghua University, Beijing 100084, China
[5] Institute of Biopharmaceutical and Health Engineering, Tsinghua Shenzhen International Graduate School, Shenzhen 518055, China

Abstract

Background: CRISPRi screening has become a powerful approach for functional genomic research. However, the off-target effects resulting from the mismatch tolerance between sgRNAs and their intended targets is a primary concern in CRISPRi applications.

Results: We introduce Guide Library Designer (GLiDe), a web-based tool specifically created for the genome-scale design of sgRNA libraries tailored for CRISPRi screening in prokaryotic organisms. GLiDe incorporates a robust quality control framework, rooted in prior experimental knowledge, ensuring the accurate identification of off-target hits. It boasts an extensive built-in database, encompassing 1,397 common prokaryotic species as a comprehensive design resource. It also provides the capability to design sgRNAs for newly discovered organisms by accepting uploaded design resource. We further demonstrated that GLiDe exhibits enhanced precision in identifying off-target binding sites for the CRISPRi system.

Conclusions: We present a web server that allows the construction of genome-scale CRISPRi sgRNA libraries for prokaryotes. It mitigates off-target effects through a robust quality control framework, leveraging prior experimental knowledge within an end-to-end, user-friendly pipeline.

Keywords: CRISPRi sgRNA design, Off-target binding identification, Public webserver

Background

As the number of bacterial genomes continues to grow, genotype–phenotype mapping is emerging as a potent approach in functional genomics research. It yields valuable insights into the microbiology and engineering of microorganisms [1–3]. To investigate the function of a specific gene, the most straightforward strategy is to perturb the gene and then observe the resulting changes in phenotype [4]. In recent years, CRISPR system originates from bacterial immune system has provided an extraordinary tool for genome editing, which is programmable and applicable in a wide range of organisms [5–7]. By inactivating the endonuclease activity of Cas9 through mutations (D10A and H840A), nuclease-dead Cas9 (dCas9) is generated [8]. This can be adapted for transcriptional regulation, giving rise to the CRISPR interference (CRISPRi) system [9]. This innovative approach facilitates CRISPRi screening [10], a technique for genome-wide

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

gene perturbation. This method employs a pooled library of single-guide RNAs (sgRNAs), where each sgRNA is meticulously designed to guide dCas9 protein to a specific genetic locus flanked by 3'-NGG protospacer adjacent motif (PAM) via Watson–Crick base pairing [8], resulting in the inhibition of transcription of the target gene. Following the introduction of the sgRNA library, high-throughput screening techniques such as fluorescence-activated cell sorting [11, 12] or fitness screening [13, 14] are used to assess how cells with different sgRNAs behave. These behaviors are subsequently identified through next-generation sequencing. However, false positive results may arise occasionally in CRISPRi screening, primarily due to unexpected binding, a phenomenon commonly known as "off-targeting." This can occur due to the tolerance of mismatches between sgRNAs and their intended targets [15]. Consequently, off-target binding may lead to the unintended inhibition of additional genes beyond the originally targeted one. Hence, ensuring highly reliable sgRNA design is of paramount importance.

To date, a variety of tools have been established for sgRNA design [16–22]. It's worth noting that these applications are predominantly tailored for the CRISPR/Cas9 system in eukaryotes, where target DNA cleavage typically takes place following a conformational gating process [23]. This process requires sufficient base pairing between the sgRNA and the target DNA. Therefore, off-target binding is more prone to occur compared to off-target cleavage [24]. Additionally, it's crucial to consider the context of prokaryotic genomes when discussing sgRNA design tools. Prokaryotic genomes differ fundamentally from eukaryotic genomes in terms of genome accessibility due to a less organized chromosome structure [25, 26] and variations in intracellular conditions. These unique characteristics may impact the applicability of existing sgRNA design tools to prokaryotic systems.

To tackle the challenges outlined above, our research focuses on the development of an accurate off-target identification algorithm tailored specifically for the CRISPRi system in prokaryotes. Drawing upon insights gained from extensive pooled screening assays evaluating binding affinities of diverse sequences in prokaryotes [10, 27], we have conducted a comprehensive analysis to discern the impact of mismatches on binding affinity. From these derived principles, we have developed GLiDe, a web-based tool that allows users to easily and rapidly design genome-scale sgRNA libraries for CRISPRi screening in prokaryotes. GLiDe offers a built-in database of 1,397 common microorganisms and accepts uploaded reference files for less common species. Additionally, we illustrated that sgRNAs generated by GLiDe display a reduced propensity for off-target binding. Our in-depth analysis, supported by quantitative data, unequivocally demonstrated that GLiDe showcased substantial performance improvement in designing sgRNA libraries for CRISPRi screening. Consequently, through rigorous design principles and meticulous experimentation, we have firmly established GLiDe as a highly promising tool poised to substantially promote functional genomic studies in prokaryotes.

Methods

Data prerequisites for the design of sgRNA library

To Design an sgRNA library, GLiDe requires two types of files: a FASTA file containing the entire DNA sequence, and an annotation file offering comprehensive details including genomic names, coding types, locations, strands, phases, and functional attributes.

Regarding the annotation file, two formats are acceptable: (i) generic feature format files (GFF) and (ii) protein/RNA table files (PTT/RNT). GLiDe has employed a built-in database containing reference files for 1,397 common organisms based on NCBI RefSeq release 225 (updated on July 12, 2024). This database is scheduled to be updated every six months. For newly sequenced organisms, these files can be generated using genome annotation pipeline like PGAP [28].

sgRNA library design workflow

As shown in Fig. 2A, upon receiving the files, GLiDe constructs a coding list that connect each gene feature with its corresponding sequence. Genes with high similarity are considered as the same function [29, 30] and are grouped into a cluster. This is achieved by employing BLASTN [31] with default parameters for sequence alignment, applying a strict threshold (<0.001 evalue, $>95\%$ identity, $>95\%$ hit coverage, and $>95\%$ query coverage). Each cluster may encompass one or more genes, and genes within the same cluster are considered to be multiple copies of the same gene. Consequently, potential hits across the cluster are not considered as off-target. For example, Table S1 provides a list of all clusters of multi-copy coding genes in *E. coli* MG1655. Candidate sgRNAs are identified using regular expressions, targeting two categories of guides: N20NGG and N20NAG (N = A, T, C or G), corresponding to canonical and non-canonical PAM sequences [9]. These candidate sgRNAs are categorized into three groups: (i) sgRNAs with an NGG PAM targeting the non-template strand; (ii) sgRNAs with an NGG PAM targeting the template strand; and (iii) sgRNAs with an NAG PAM. Three FASTA files (FASTA1, FASTA2, and FASTA3) are generated, corresponding to group (i), (ii), and (iii), respectively. These files contain the PAM-proximal 12-bp regions (seed regions) of all candidate sgRNAs, as these regions are pivotal for binding [14]. An illustrative example is presented in Fig. S1. Following this, candidate sgRNAs undergo a rigorous quality control process to eliminate those with potential off-target hits (see "Quality control and output" section in Results). Subsequently, only sgRNAs adhering to the user-specified upper and lower limits for GC content are retained. Finally, the designed library is presented to the user, as detailed in the "Visualization of results" section.

Design of negative control sgRNAs

Negative control sgRNAs are designed to have no specific targets across the genome, which can be used to assess the influence of external factors on cellular phenotype. GLiDe designs these sgRNAs by generating random N20 sequences and subsequently removing those with notable target sites. For both NGG and NAG PAMs, a penalty score of 25 is applied, and the GC content limits match those of the sgRNA library. Additionally, GLiDe ensures that there are no five or more consecutive identical bases in these negative control sgRNAs.

Visualization of results

GLiDe provides the final sgRNA library in a table, with each sgRNA linked to its target gene. The sgRNAs closer to the start codon are listed first because those targeting near the transcription start site (TSS) are more likely to have higher knock-down activity [9]. Although a recent study showed some genes exhibit a different trend in the relationship

between repression rate and distance to the TSS, guides near the TSS still show high efficiency [32]. In addition to the table, GLiDe also generates an interactive graphical interface using D3GB [33]. This interface provides users with a comprehensive overview of the entire genome and the designed sgRNA library.

DNA manipulation and reagents

Plasmid extraction and DNA purification procedures were carried out employing kits provided by Vazyme. PCR reactions were performed utilizing KOD One™ PCR Master Mix from TOYBO Life Science. The PCR primers were ordered from Azenta (Table S2). Plasmids were constructed by Gibson Assembly, with a mixture comprising 10 U/μL T5 exonuclease, 2 U/μL Phusion High-Fidelity DNA polymerase, and 40 U/μL Taq DNA ligase, all sourced from New England Biolabs. The antibiotic concentrations for kanamycin and ampicillin were maintained at 50 and 100 mg/L, respectively. All experiments were conducted in *Escherichia coli* MCm [10], a comprehensive list of all strains and plasmids utilized in this study can be found in Table S3.

Plasmid construction

The reporting system was established using two separate plasmids: one harbored dCas9, named pdCas9-J23111, was previously constructed as described in prior work [10]. The other plasmid was derived from pN20test-114mCherry-r0-m1 [27], which was responsible for the expression of sgRNA and *mcherry*. In this particular plasmid, sgRNA expression was regulated by the J23119 promoter, while the N20 sequence was inserted upstream of the −35 region of the J23114 promoter, controlling *mcherry* expression. A total of 21 distinct pN20test-114mCherry plasmids were constructed (Fig. S2). The plasmids were assembled using the Gibson Assembly method from PCR products (primers listed in Table S2), using the original pN20test-114mCherry-r0-m1 plasmid as the template. All constructed plasmids were confirmed through Sanger sequencing.

Cell cultivation

Strains were initially cultured overnight at 37 °C and 220 rpm in a 48-well deep-well plate, each contained 1 mL of LB medium with kanamycin and ampicillin. The grown cells were transferred to fresh LB medium with a 0.5% dilution and grown again under the same conditions as above for 10 h. This subculture process was repeated to ensure the stability of mCherry expression and avoid cell adhesion. In preparation for cytometry assays, cultures were next diluted in fresh LB medium with antibiotics to OD600 = 0.02 and then grown for 4 h to the logarithmic phase. After cultivation, 5 μL of culture medium from each well was diluted into 200 μL of phosphate-buffered saline. Three independent biological replicates were prepared for each strain.

Flow cytometry assay and data processing

The flow cytometry assay was performed on an LSRFortessa flow cytometer (BD Biosciences) using a 96-well plate. Gating based on the FSC area and SSC area was carried out to exclude non-cell particles. To ensure accurate measurements, autofluorescence was quantified using the MCm/PdCas9-J23111 strain and subsequently subtracted during the data analysis process.

In the cytometry analysis, the fluorescence intensity distribution was log10-transformed and fitted to a two-component Gaussian mixture model [34] with parameters $(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2)$ through the expectation–maximization algorithm. Here, λ and $1-\lambda$ represent the mixing coefficients of the two Gaussian components, μ_1, μ_2, σ_1 and σ_2 represent the mean and standard deviation of the first and second Gaussian component, respectively (Eq. 1).

$$f(x) = \lambda N\left(\mu_1, \sigma_1^2\right) + (1-\lambda)N\left(\mu_2, \sigma_2^2\right) \tag{1}$$

The mean expression strength was calculated with Eq. 2.

$$Mean = Mean_1 \times Mean_2 = exp(m_1 + V_1/2) \times exp(m_2 + V_2/2) \tag{2}$$

where $m_1 = \lambda_i \mu_{1i} log(10)$, $V_1 = \left(\lambda_i \sigma_{1i} log(10)\right)^2$, $m_2 = (1-\lambda_i)\mu_{2i} log(10)$ and $V_2 = \left((1-\lambda_i)\sigma_{2i} log(10)\right)^2$.

The repression rate of each group was calculated using Eq. 3.

$$repression\ rate = 1 - \frac{Mean_{Ri} - Mean_{Ri-PC}}{Mean_{Ri-NC} - Mean_{Ri-PC}} \tag{3}$$

where $i = 1$–4.

Results

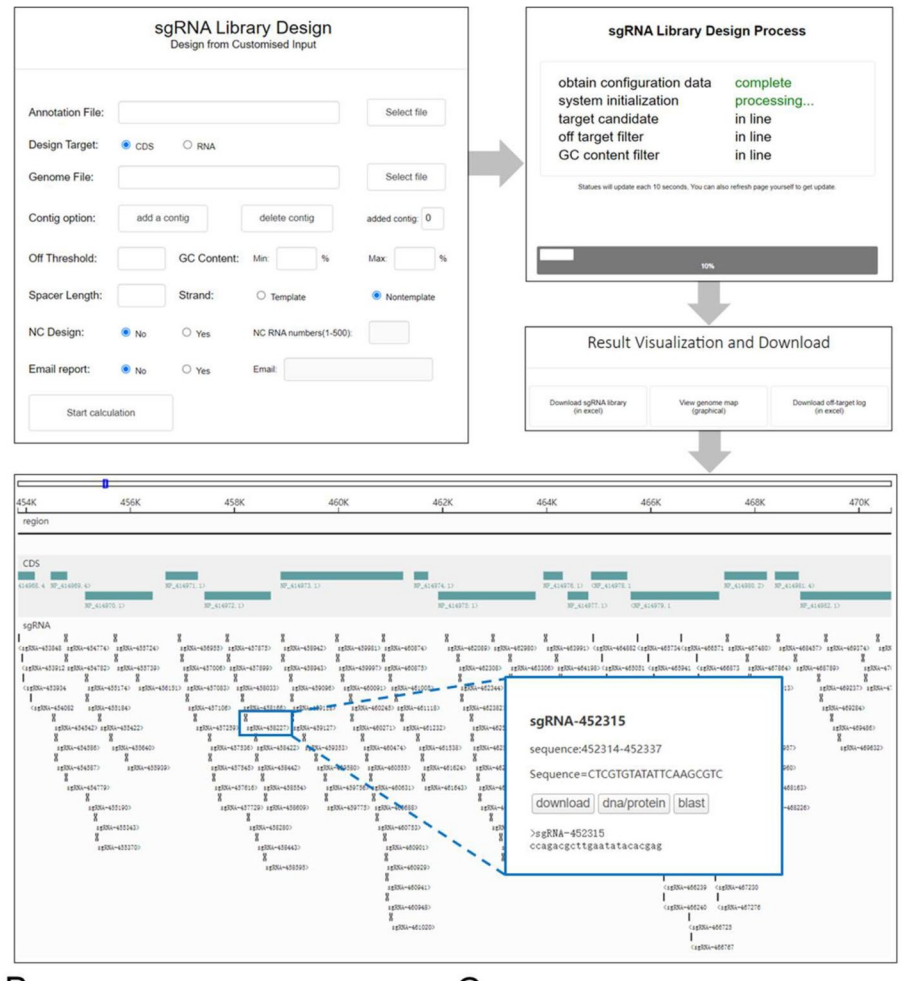
Implementation of the GLiDe web tool

We have developed GLiDe, a web-based tool for genome-scale CRISPRi sgRNA library design in prokaryotes, which is available at https://www.thu-big.net/sgRNA_design/. It employs a powerful algorithm for identifying off-target hits and outputs the results on an interactive page (Fig. 1A). This result can also be easily downloaded from the website in tabular format (Fig. 1B).

Input and configuration

GLiDe includes an extensive built-in database harboring over 1,397 prokaryotic organisms and allows users to upload their reference files (see Methods). To optimize sgRNA design, five essential parameters are required: (i) design target, which can be either coding sequence (CDS) or RNA coding genes (RNA); (ii) off-target threshold, which is the penalty score used in the sgRNA quality control, with a default value of 20. Based on our experience, a recommended range for designing genome-scale libraries is 16–21; (iii) GC limits, which represent the upper and lower boundaries for the GC content of each sgRNA, with default limits of 30% and 85%. sgRNAs with very high or very low GC content are reported to be less active [35]; (iv) spacer length, which defines the length of the sgRNA being designed, with a default value of 20; and (v) target strand, which can be either template or non-template, referring to the strand that the sgRNAs target. It's important to note that, although for common CRISPRi systems, the non-template strand is the preferred choice to ensure effective gene silencing [9], we offer the alternative option for other applications, such as genome editing, base editing [36, 37], or dynamic imaging [38], where there may be less strand preference.

A



B

Location	Strand	Length	Gene	Guide sequence	Guide position	Guide number
20815..21078	-	87	*rpsT*	AAGTACGCATCATAGAGCGA	20988	1
190..255	+	21	*thrL*	CGCACCGTTACCTGTGGTAA GGTGGTGCTAATGCGTTTCA	249 210	2

C

Guide sequence	Potential off-target hit(s)	penalty Score	Off-target number
TTAAAACCAGTCAGTTGCTC	TGAAGGCGTGCCAGTTGCTC TAAAGAACAGTCAGATGCTC	19.0 15.5	2
ATCTCGGCATTCAATTTCTT	ATAGAGGAAGTGAATTTCTT CAAACGCCAGTCAATTTCTT	19.0 17.0	2

Fig. 1 Work flow and results of a GLiDe query. **A** The main design process starts when a "start calculation" request is received, and the website automatically redirects to the progress page that displays the real-time progress of the calculation. After the calculation is complete, GLiDe redirects again to the results page. Users are free to download the result tables and view the sgRNA library on the interactive page. **B** The designed sgRNA library is grouped by its targeted gene. **C** The list of sgRNAs which are excluded from quality control, along with their off-target hits and penalty scores

GLiDe can also design library for multi-contig sequences, such as strains containing plasmids. For the second and subsequent contigs, the sequence file is necessary, while the annotation file is optional. This feature is designed for contigs where there is no need for sgRNA design, but the prevention of sgRNAs targeting these specific contigs is desired, like plasmid vectors used in experiments. If users choose to upload the annotation file for these additional contigs, GLiDe will design sgRNAs accordingly.

Quality control and output

Upon configuration, a server request is initiated to commence the workflow (Fig. 2A). The core process is the off-target hit identification process, which relies on the strand invasion model grounded in the natural binding process [39–41] and previous experimental findings [8, 42]. The fundamental concept involves identifying sequences that are less likely to result in off-target hits, with a focus on those exhibiting fewer mismatches in the PAM-proximal region compared to the target sequence. This process involves three alignments according to the user defined target strand (template or non-template) using SeqMap [43]. For instance, if the non-template strand is selected, FASTA1 would be aligned with FASTA1, FASTA2, and FASTA3. We implement two general rules to evaluate the impact of each off-target hit (Fig. 2B).

One is seed region rule, where the seed region refers to the PAM-proximal 7–12 bp region. Mismatches within the seed region exert a notable impact on the binding affinity [14, 44]. Our previous study revealed that two mismatches in the seed region substantially weaken binding affinity [27]. Despite its effectiveness, this principle has not been integrated into existing tools. Consequently, in our design, off-target binding is not deemed to occur when there are more than two mismatches within the 12 bp PAM-proximal region.

The second rule is the penalty scoring rule, which considers the influence of mismatches based on their distance to the PAM, considering that mismatches are generally better tolerated at the 5' end of the sgRNA than at the 3' end [18]. The sgRNA regions are categorized from the 3' end to the 5' end as region I (7 nt), region II (5 nt), and region III (the remaining sgRNA sequence). This division is informed by the experimental findings

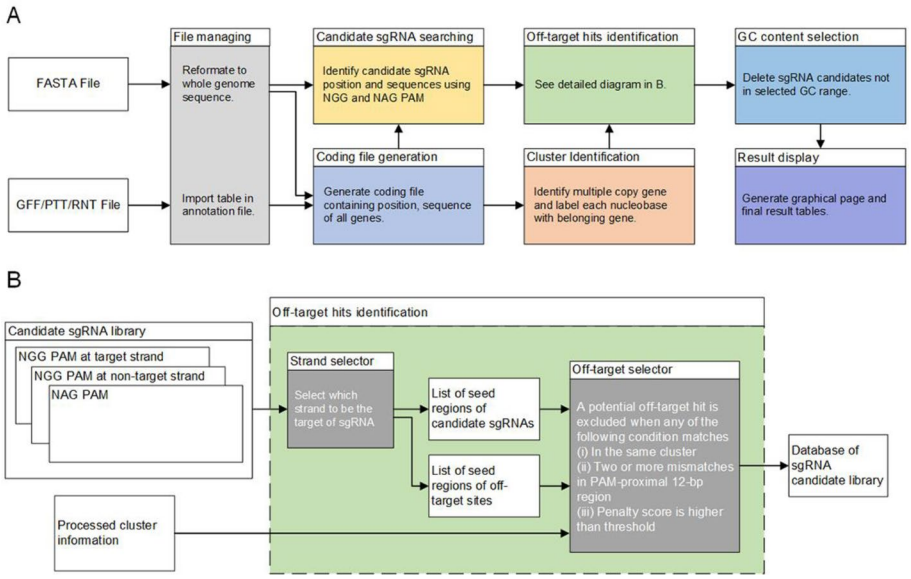


Fig. 2 Functional Functional blocks of GLiDe. **A** The GLiDe flowchart encompasses the processing of the FASTA file and the annotation file through multiple modules. These modules conduct candidate sgRNA searches, quality control (identifying off-target hits and selecting GC content), and ultimately present the results. **B** The off-target hits identification module generates three FASTA files containing 12 bp PAM-proximal sequences of sgRNA candidates. Alignments are performed among these three files to identify potential off-target hits

[15, 24, 35]. The mismatch penalties for region I, II, and III are 8, 4.5, and 2.5 (for NGG PAM), and 10, 7, and 3 (for NAG PAM), respectively (Fig. S3). This principle has been used in previous studies [10, 45]. An off-target site is identified and relevant sgRNAs are eliminated from further processing when the penalty score, Σ (penalty $\times$ mismatch), falls below the user-defined threshold. For example, if the threshold is set at 21, an sgRNA identified to have potential off-target effects with two mismatches in Region I and one mismatch in Region II would have a penalty score of $4.5+(8\times2)=20.5$, which is less than 21. Consequently, this sgRNA would be removed from the library. An illustrated example of the quality control process is presented in Fig. S4.

The anticipated processing time for a standard genome-scale library design range from 5 to 20 min, varying based on the size of the genome. The result sgRNA library is thoughtfully presented in a tabular format, where sgRNAs are meticulously organized by their respective targeted genes and annotated with their corresponding start positions (see Methods). Additionally, an interactive page is generated using the D3GB genome browser [33], offering an intuitive visualization of the sgRNA library. This page accurately maps all genes and sgRNAs to their natural positions within the genome.

GLiDe is capable of designing CRISPRi sgRNA libraries

It is important to note that instead of CRISPR/Cas9 system that existing sgRNA design tools are focusing on, GLiDe is specially tailored for CRISPRi system. CRISPR/Cas9 system requires conformational gating process for formatting an R-loop structure before DNA cleavage [23], which requires sufficient base pairing between the sgRNA and the target DNA, reducing the likelihood of off-target cleavage. Consequently, sgRNAs without off-target cleavage may have potential off-target binding effects. GLiDe implements a more rigorous off-target selection principle, enabling the identification of these sgRNAs. To illustrate this capability, we compared sgRNA library of *Corynebacterium glutamicum* ATCC 13032 designed by GLiDe and sgRNA libraries generated by two widely used tools: CHOPCHOP [19] and Synthego (https://design.synthego.com/). We discovered that some of the highly ranked sgRNAs intended for CRISPR/Cas9 systems were indeed identified as having the potential for off-target binding. Among these identified off-target sequences, we randomly chose four (R1–R4, Table 1) for further validation.

To quantitatively assay the off-target binding effects, we applied a reporter system to connect binding affinity with mCherry expression [27]. In this system, we inserted

Table 1 An example of four high rank sgRNAs designed by existing tools and their potential off-target hits identified by GLiDe

No	Design tool	Rank	Target gene	sgRNA sequence	Off-target gene	Potential off-target hits discovered byGLiDe
R1	chopchop	1	*Cgl0025*	T***AA***TTC***GAA***TGGTTC CACGG	*Cgl2364*	T***CC***TTC***CTC***TGGTTC CACGG
R2	chopchop	3	*Cgl0044*	***GCACGAT***GCCCACCAC ACCG	*Cgl1968*	***CGGATGG***GCCCACCAC ACCG
R3	Synthego	3	*Cgl0005*	***TTCT***G***CGC***ATGGCTTCG GCG	*Cgl0369*	***GCGC***G***GCG***ATGGCTTCG GCG
R4	Synthego	2	*Cgl0059*	***GAG***CT***CAG***ATCGCT CAACAT	*Cgl1680*	***TCT***CT***TGA***ATCGCTCAA CAT

The bold italic letters are the mismatched bases

the GLiDe identified off-target hit (N20 sequence) upstream of the *mcherry* promoter. Upon binding of the dCas9-sgRNA complex to the N20 sequence, the expression of *mcherry* would be repressed, resulting in altered fluorescence intensity. For each sgRNA (EXP) and its potential off-target hit, two additional sgRNAs were designed as positive (PC) and negative control (NC). The PC was the reverse-complement of the N20 sequence and was used to represent full repression; meanwhile, the NC had no binding affinity with the N20 sequence and was used to represent the fluorescence intensity without repression (Fig. 3A). Since sgRNAs may bind to the endogenous genome to impact fluorescence intensity, experiments were conducted in *Escherichia coli* MCm, and all sgRNAs were cross-referenced with the bacterial genome to ensure the absence of potential off-target hits. Moreover, a blank control N20 sequence (blank) was designed, which had no binding affinity to any sgRNA, to assess whether

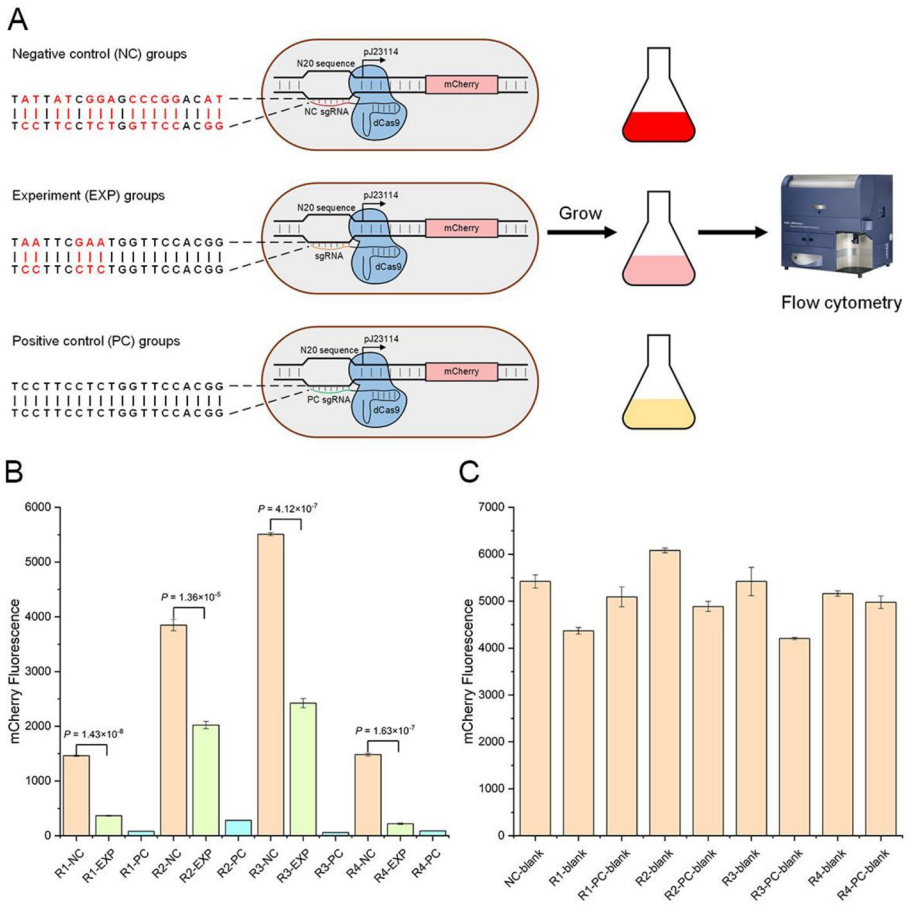


Fig. 3 Assessment of off-target binding affinities using flow cytometry assay. **A** The binding affinity of each mismatched sgRNA was compared with positive controls and negative controls. Fluorescence intensities of the three groups were measured by flow cytometry. **B** All sgRNAs from the EXP groups exhibited significant off-target binding affinities ($P < 10\text{-}4$for all groups, two tailed *t*-test). **C** The fluorescence intensities of the blank control groups were similar, demonstrated that the effects of sgRNA expression on *mcherry* expression are negligible. Therefore, the fluorescence intensity variations among EXP, PC, and NC groups were originated from the differences in binding affinities of sgRNAs. In both (**B**) and (**C**), error bars represent the standard deviations of biological replicates ($n = 3$)

the sgRNAs themselves had any influence on mCherry fluorescence. The sgRNA and N20 sequences used in all experiment groups are listed in Table 2.

The results revealed that all four sgRNAs bound to the off-target hits identified by GLiDe, leading to reduced fluorescence intensity compared to the negative control groups (Fig. 3B, $P < 1.36 \times 10^{-5}$ for all groups, two-tailed *t*-test; Fig. S5). The repression rates ranged from 51.12% (R2) to 90.70% (R4), with an average of 69.47% (see Methods). This rate is capable of causing unexpected outcomes and potentially result in false positives in the screening experiment [11]. The blank control groups exhibited similar mCherry fluorescence intensities (Fig. 3C), indicating that the observed repressions were primarily attributed to off-target effects. These findings provide strong evidence that GLiDe excels in the design of CRISPRi sgRNA libraries.

Discussion

GLiDe presents several distinctive advantages. Firstly, it concentrates on designing genome-scale CRISPRi sgRNA libraries, a focus notably lacking in prior web-based platforms like CHOPCHOP [19] and synthego (https://design.synthego.com/). Secondly, in contrast to existing scripts for bacterial CRISPRi sgRNA design [14, 46, 47], GLiDe provides a more convenient and flexible way for a wide range of user groups. As a web-based tool, the entire calculation process occurs online, eliminating the need for installation or local dependencies. Its compact, user-friendly interface simplifies the process, requiring researchers to set only a few parameters and optionally submit the sequence and annotation of the genome of interest. Finally, GLiDe's quality control mechanism is built on a solid theoretical foundation. Two key design rules,

Table 2 sgRNAs and N20 sequences of each test group

Test group	sgRNA sequence	N20 sequence
R1-NC	TATTATCGGAGCCCGGACAT	TCCTTCCTCTGGTTCCACGG
R1-EXP	TAATTCGAATGGTTCCACGG	
R1-PC	TCCTTCCTCTGGTTCCACGG	
R2-NC	TATTATCGGAGCCCGGACAT	CGGATGGGCCCACCACACCG
R2-EXP	GCACGATGCCCACCACACCG	
R2-PC	CGGATGGGCCCACCACACCG	
R3-NC	TATTATCGGAGCCCGGACAT	GCGCGGCGATGGCTTCGGCG
R3-EXP	TTCTGCGCATGGCTTCGGCG	
R3-PC	GCGCGGCGATGGCTTCGGCG	
R4-NC	TATTATCGGAGCCCGGACAT	TCTCTTGAATCGCTCAACAT
R4-EXP	GAGCTCAGATCGCTCAACAT	
R4-PC	TCTCTTGAATCGCTCAACAT	
NC-blank	TATTATCGGAGCCCGGACAT	CATGTATTATACACGAAGTT
R1-blank	TAATTCGAATGGTTCCACGG	
R1-PC-blank	TCCTTCCTCTGGTTCCACGG	
R2-blank	GCACGATGCCCACCACACCG	
R2-PC-blank	CGGATGGGCCCACCACACCG	
R3-blank	TTCTGCGCATGGCTTCGGCG	
R3-PC-blank	GCGCGGCGATGGCTTCGGCG	
R4-blank	GAGCTCAGATCGCTCAACAT	
R4-PC-blank	TCTCTTGAATCGCTCAACAT	

the seed region rule (not yet considered by existing tools) and the penalty scoring rule, have been introduced based on extensive experimental evidence. This robust theoretical underpinning enhances the reliability of GLiDe, contributing to the accuracy and effectiveness of the tool. Our further experiments have demonstrated GLiDe's superiority over existing tools when employed to design CRISPRi sgRNA libraries. Therefore, we believe GLiDe has extensive applicability in functional genomic studies such as essential genes study [48, 49], bioprocesses optimization in industrially relevant microorganisms [50, 51], or identification of complex genetic interactions [52].

The quality control methodology employed by GLiDe is based on the strand invasion model. During binding, mismatches in PAM-proximal region, particularly when there are two mismatches in seed region, could create a substantial energy barrier that cannot be adequately compensated. This barrier effectively impedes the strand invasion from the beginning, increasing the likelihood of dCas9 disengaging from the target. However, there is still potential limitation in the estimation of the impact of PAM-proximal mismatches. The current penalty scoring system primarily relies on experimental data, lacking a robust theoretical explanation. Furthermore, the current model assumes equal contributions from different mismatched nucleotide pairs, yet variations in stability exist among these mismatches [53]. Certain mismatches, such as rU/dG, rG/dT, and rG/dG, exhibit similar stabilities to Watson–Crick pairs [54]. One potential avenue for improvement is to incorporate thermodynamics. By treating the combination process as a Markov chain model, we can quantitatively calculate the thermodynamics of the sgRNA/DNA binding process based on fundamental thermodynamic parameters (nearest-neighbor parameters) [27]. We have further gathered nearest-neighbor parameters for the remaining 12 RNA/DNA single internal mismatches [55] and incorporated these data into the quantitative CRISPRi design tool we previously introduced (https://www.thu-big.net/sgRNA_design/Quantitative_CRISPRi_Design/). Nevertheless, the current tool faces limitations in computing binding activities when mismatches are adjacent or located at the ends due to a lack of corresponding thermodynamic parameters. The challenge arises because adjacent mismatches are too numerous to be directly measured by experiments. Substantial work is still needed to complete the thermodynamic data for these scenarios.

It should be noted that GLiDe does not consider the on-target activity of sgRNAs, which has been proved to associate with sequence [56] and target location, including the distance from the TSS [9] and start codon [32]. Therefore, the designed library could include some less effective sgRNAs. However, this limitation is generally acceptable, as the major concern in CRISPRi screening experiments is off-target effects, which can lead to false positive results. Less effective sgRNAs can still achieve target gene knockdown, and the inclusion of multiple sgRNAs per gene helps reduce potential impact of less effective ones. Moreover, apart from the above-mentioned factors, the optimal window for active sgRNA positioning varies across different organisms [57, 58]. Currently, some models have been constructed based on machine learning [32, 56]. While these studies offer valuable proof-of-concept results, a unified model for predicting on-target guide efficiency in prokaryotes has not yet been established. Therefore, creating a comprehensive approach remains an important direction for future research. Overall, owing to the usability, solid basis, and high

precision, we believe GLiDe can serve as a powerful tool that provides a wide range of novel research opportunities.

Conclusions

GLiDe is a web-based tool for genome-scale CRISPRi sgRNA library design for prokaryotic organisms. It has a powerful quality control principle based on large-scale empirical data to enhance the accuracy of identifying off-target hits. It also has a substantial database containing 1,397 common prokaryotes and supports sgRNA design for new organisms using uploaded reference files.

Availability and requirements

Project name: GLiDe. Project home page: https://www.thu-big.net/sgRNA_design/. Operating system(s): Platform independent. Programming language: Python. Other requirements: None. License: MIT. Any restrictions to use by non-academics: None.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-024-06012-0.

Additional file1

Acknowledgements
We thank the support of the National Key R&D Program of China [2023YFC3402300 to CZ], the National Natural Science Foundation of China [21938004 to XX], and the National Natural Science Foundation of China [U2032210 to CZ]. We also thank Dr. T. Wang (Tsinghua University) for building the original demo of the sgRNA design tool. We thank members from the C. Zhang and H. Xing laboratories for critical discussions of this work.

Author contributions
TX and HF designed GLiDe, conducted all experiments and data analysis, wrote the manuscript and prepared figures. XX and CZ conceived GLiDe, obtained funding, supervised software design and development. All authors edited and approved the manuscript. All authors read and approved the final manuscript.

Funding
This work was supported by the National Key R&D Program of China [2023YFC3402300 to CZ]; the National Natural Science Foundation of China [21938004 to XX]; the National Natural Science Foundation of China [U2032210 to CZ].

Availability of data and materials
The datasets generated and analysed during the current study are available in the GitHub repository, https://github.com/falconxtj/GLiDe. Including an offline version of GLiDe with tutorial, sample data, raw data of the flow cytometry assays and plasmid maps.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 9 April 2024 Accepted: 9 December 2024
Published online: 03 January 2025

References

1. van Opijnen T, Bodi KL, Camilli A. Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. Nat Methods. 2009;6:767–72.
2. Garst AD, Bassalo MC, Pines G, Lynch SA, Halweg-Edwards AL, Liu R, et al. Genome-wide mapping of mutations at single-nucleotide resolution for protein, metabolic and genome engineering. Nat Biotechnol. 2017;35:48–55.

3. Freed EF, Winkler JD, Weiss SJ, Garst AD, Mutalik VK, Arkin AP, et al. Genome-wide tuning of protein expression levels to rapidly engineer microbial traits. ACS Synth Biol. 2015;4:1244–53.
4. Feng H, Yuan Y, Yang Z, Xing X, Zhang C. Genome-wide genotype-phenotype associations in microbes. J Biosci Bioeng. 2021;132:1–8.
5. Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, et al. Multiplex genome engineering using CRISPR/Cas systems. Science. 2013;339:819–23.
6. Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, et al. RNA-guided human genome engineering via Cas9. Science. 2013;339:823–6.
7. Bikard D, Jiang W, Samai P, Hochschild A, Zhang F, Marraffini LA. Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system. Nucleic Acids Res. 2013;41:7429–37.
8. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. Science. 2012;337:816–21.
9. Qi LS, Larson MH, Gilbert LA, Doudna JA, Weissman JS, Arkin AP, et al. Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. Cell. 2013;152:1173–83.
10. Wang T, Guan C, Guo J, Liu B, Wu Y, Xie Z, et al. Pooled CRISPR interference screening enables genome-scale functional genomics study in bacteria with superior performance. Nat Commun. 2018;9:2475.
11. Hawkins JS, Silvis MR, Koo BM, Peters JM, Osadnik H, Jost M, et al. Mismatch-CRISPRi reveals the co-varying expression-fitness relationships of essential genes in *Escherichia coli* and *Bacillus subtilis*. Cell Syst. 2020;11:523-535.e9.
12. Lian J, Schultz C, Cao M, HamediRad M, Zhao H. Multi-functional genome-wide CRISPR system for high throughput genotype–phenotype mapping. Nat Commun. 2019;10:1–10.
13. Chen P, Michel AH, Zhang J. Transposon insertional mutagenesis of diverse yeast strains suggests coordinated gene essentiality polymorphisms. Nat Commun. 2022;13:1–15.
14. de Bakker V, Liu X, Bravo AM, Veening JW. CRISPRi-seq for genome-wide fitness quantification in bacteria. Nat Protoc. 2022;17:252–81.
15. Hsu PD, Scott DA, Weinstein JA, Ran FA, Konermann S, Agarwala V, et al. DNA targeting specificity of RNA-guided Cas9 nucleases. Nat Biotechnol. 2013;31:827–32.
16. Shalem O, Sanjana NE, Zhang F. High-throughput functional genomics using CRISPR–Cas9. Nat Rev Genet. 2015;16:299–311.
17. Heigwer F, Kerr G, Boutros M. E-CRISP: fast CRISPR target site identification. Nat Methods. 2014;11:122–3.
18. Sander JD, Joung JK. CRISPR-Cas systems for editing, regulating and targeting genomes. Nat Biotechnol. 2014;32:347–55.
19. Montague TG, Cruz JM, Gagnon JA, Church GM, Valen E. CHOPCHOP: a CRISPR/Cas9 and TALEN web tool for genome editing. Nucleic Acids Res. 2014;42:W401–7.
20. Concordet J-P, Haeussler M. CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. Nucleic Acids Res. 2018;46:W242–5.
21. 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. 2014;30:1473–5.
22. Liu H, Wei Z, Dominguez A, Li Y, Wang X, Qi LS. CRISPR-ERA: a comprehensive design tool for CRISPR-mediated gene editing, repression and activation. Bioinformatics. 2015;31:3676–8.
23. Sternberg SH, LaFrance B, Kaplan M, Doudna JA. Conformational control of DNA target cleavage by CRISPR–Cas9. Nature. 2015;527:110–3.
24. Boyle EA, Andreasson JOL, Chircus LM, Sternberg SH, Wu MJ, Guegler CK, et al. High-throughput biochemical profiling reveals sequence determinants of dCas9 off-target binding and unbinding. In: Proceedings of the National Academy of Sciences. 2017;114:5461–6.
25. Struhl K. Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. Cell. 1999;98:1–4.
26. Kuzminov A. The precarious prokaryotic chromosome. J Bacteriol. 2014;196:1793–806.
27. Feng H, Guo J, Wang T, Zhang C, Xing X. Guide-target mismatch effects on dCas9–sgRNA binding activity in living bacterial cells. Nucleic Acids Res. 2021;49:1263–77.
28. Li W, O'Neill KR, Haft DH, Dicuccio M, Chetvernin V, Badretdin A, et al. RefSeq: expanding the prokaryotic genome annotation pipeline reach with protein family model curation. Nucleic Acids Res. 2021;49:D1020–8.
29. McTavish H, LaQuier F, Arciero D, Logan M, Mundfrom G, Fuchs JA, et al. Multiple copies of genes coding for electron transport proteins in the bacterium Nitrosomonas europaea. J Bacteriol. 1993;175:2445–7.
30. Schrider DR, Hahn MW. Gene copy-number polymorphism in nature. In: Proceedings of the Royal Society B: Biological Sciences. 2010;277:3213–21.
31. Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL. NCBI BLAST: a better web interface. Nucleic Acids Res. 2008;36:W5-9.
32. Yu Y, Gawlitt S, de Andrade Sousa LB, Merdivan E, Piraud M, Beisel CL, et al. Improved prediction of bacterial CRISPRi guide efficiency from depletion screens through mixed-effect machine learning and data integration. Genome Biol. 2024;25:13.
33. Barrios D, Prieto C. D3GB: an interactive genome browser for R, python, and WordPress. J Comput Biol. 2017;24:447–9.
34. Feng H, Li F, Wang T, Xing X, Zeng A, Zhang C. Deep-learning–assisted sort-seq enables high-throughput profiling of gene expression characteristics with high precision. Sci Adv. 2023. https://doi.org/10.1126/sciadv.adg5296.
35. Gilbert LA, Horlbeck MA, Adamson B, Villalta JE, Chen Y, Whitehead EH, et al. Genome-scale CRISPR-mediated control of gene repression and activation. Cell. 2014;159:647–61.
36. Gaudelli NM, Komor AC, Rees HA, Packer MS, Badran AH, Bryson DI, et al. Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. Nature. 2017;551:464–71.
37. Komor AC, Kim YB, Packer MS, Zuris JA, Liu DR. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. Nature. 2016;533:420–4.
38. Chen B, Gilbert LA, Cimini BA, Schnitzbauer J, Zhang W, Li GW, et al. Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. Cell. 2013;155:1479–91.

39. Szczelkun MD, Tikhomirova MS, Sinkunas T, Gasiunas G, Karvelis T, Pschera P, et al. Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes. In: Proceedings National Academy Science U S A. 2014;111:9798–803.
40. Nishimasu H, Ran FA, Hsu PD, Konermann S, Shehata SI, Dohmae N, et al. Crystal structure of Cas9 in complex with guide RNA and target DNA. Cell. 2014;156:935–49.
41. Cofsky JC, Soczek KM, Knott GJ, Nogales E, Doudna JA. CRISPR–Cas9 bends and twists DNA to read its sequence. Nat Struct Mol Biol. 2022;29:395–402.
42. Jones SK, Hawkins JA, Johnson NV, Jung C, Hu K, Rybarski JR, et al. Massively parallel kinetic profiling of natural and engineered CRISPR nucleases. Nat Biotechnol. 2020;39:84–93.
43. Jiang H, Wong WH. SeqMap: mapping massive amount of oligonucleotides to the genome. Bioinformatics. 2008;24:2395–6.
44. Semenova E, Jore MM, Datsenko KA, Semenova A, Westra ER, Wanner B, et al. Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. In: Proceedings National Academy Science U S A. 2011;108:10098–103.
45. Yu X, Li S, Feng H, Liao X, Xing XH, Bai Z, et al. CRISPRi-microfluidics screening enables genome-scale target identification for high-titer protein production and secretion. Metab Eng. 2023;75:192–204.
46. Liu X, Kimmey JM, Matarazzo L, de Bakker V, Van Maele L, Sirard JC, et al. Exploration of bacterial Bottlenecks and *Streptococcus pneumoniae* pathogenesis by CRISPRi-Seq. Cell Host Microb. 2021;29:107-120.e6.
47. Banta AB, Ward RD, Tran JS, Bacon EE, Peters JM. Programmable gene knockdown in diverse bacteria using mobile-CRISPRi. Curr Protoc Microbiol. 2020;59:e130.
48. Peters JM, Colavin A, Shi H, Czarny TL, Larson MH, Wong S, et al. A comprehensive, CRISPR-based functional analysis of essential genes in bacteria. Cell. 2016;165:1493–506.
49. Rousset F, Cabezas-Caballero J, Piastra-Facon F, Fernández-Rodríguez J, Clermont O, Denamur E, et al. The impact of genetic diversity on gene essentiality within the *Escherichia coli* species. Nat Microbiol. 2021;6:301–12.
50. Li S, Jendresen CB, Landberg J, Pedersen LE, Sonnenschein N, Jensen SI, et al. Genome-wide CRISPRi-based identification of targets for decoupling growth from production. ACS Synth Biol. 2020;9:1030–40.
51. Donati S, Kuntz M, Pahl V, Farke N, Beuter D, Glatter T, et al. Multi-omics analysis of CRISPRi-knockdowns identifies mechanisms that buffer decreases of enzymes in *E. coli* metabolism. Cell Syst. 2021;12:56-67.e6.
52. Jaffe M, Dziulko A, Smith JD, St. Onge RP, Levy SF, Sherlock G. Improved discovery of genetic interactions using CRISPRiSeq across multiple environments. Genome Res. 2019;29:668–81.
53. Sugimoto N, Nakano S, Katoh M, Matsumura A, Nakamuta H, Ohmichi T, et al. Thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes. Biochemistry. 1995;34:11211–6.
54. Sugimoto N, Nakano M, Nakano S. Thermodynamics−structure relationship of single mismatches in RNA/DNA duplexes. Biochemistry. 2000;39:11270–81.
55. Xiang T, Feng H, Xing X, Zhang C. Thermodynamic parameters contributions of single internal mismatches In RNA/DNA Hybrid Duplexes. bioRxiv. 2022;2022.11.25.517909.
56. Calvo-Villamañán A, Ng JW, Planel R, Ménager H, Chen A, Cui L, et al. On-target activity predictions enable improved CRISPR–dCas9 screens in bacteria. Nucleic Acids Res. 2020;48:e64–e64.
57. Guo J, Wang T, Guan C, Liu B, Luo C, Xie Z, et al. Improved sgRNA design in bacteria via genome-wide activity profiling. Nucleic Acids Res. 2018;46:7052–69.
58. Smith JD, Suresh S, Schlecht U, Wu M, Wagih O, Peltz G, et al. Quantitative CRISPR interference screens in yeast identify chemical-genetic interactions and new rules for guide RNA design. Genome Biol. 2016;17:1–16.

Publisher's Note

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