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

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

[1] MOE Key Laboratory for Industrial Biocatalysis, Institute of Biochemical Engineering, Department of Chemical Engineering, Tsinghua University, Beijing 100084, China

[2] Division of Biology and Bioengineering, California Institute of Technology, Pasadena, CA 91125, USA

[3] Center for Synthetic and Systems Biology, Tsinghua University, Beijing 100084, China

[4] Institute of Biopharmaceutical and Health Engineering, Tsinghua Shenzhen International Graduate School, Shenzhen 518055, China

[†] Joint Authors

ABSTRACT

CRISPRi screening has become a powerful approach for functional genomic research. However, the issue of off-target effects resulting from the mismatch tolerance between sgRNAs and their intended targets is a primary concern in CRISPRi applications. To address this issue, 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 1,426 common prokaryotic species as a comprehensive design resource. In addition, GLiDe provides the capability to design sgRNAs for newly discovered organisms. Our comparative analysis demonstrated that GLiDe exhibits enhanced precision in identifying off-target binding sites for the CRISPRi system.

Availability and Implementation: Freely available on the web at: https://www.thu-big.net/sgRNA_design/.

Contact: Huibao Feng: fhb_14@163.com, Chong Zhang: chongzhang@tsinghua.edu.cn

Supplementary information:

INTRODUCTION

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. An established technique is CRISPR interference (CRISPRi) screening [4]. This method employs a pooled library of single-guide RNAs (sgRNAs), where each sgRNA is meticulously designed to guide nuclease-dead Cas9 (dCas9) protein to a specific genetic locus flanked by 3'-NGG protospacer adjacent motif (PAM) via Watson-Crick base pairing [5], 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 (FACS) [6,7] or fitness screening [8,9] are used to assess how cells with different sgRNAs behave. These behaviors are subsequently identified through next-generation sequencing (NGS). However, false positive results can occasionally arise due to unexpected binding, a phenomenon commonly referred to as "off-targeting." This occurs because the binding process can tolerate several mismatches [10]. Hence, ensuring highly reliable sgRNA design is of paramount importance.

To date, a variety of tools have been established for sgRNA design [11–17]. It's worth noting that

these applications are predominantly tailored for the CRISPR/Cas9 system, where target DNA cleavage typically takes place following a conformational gating process [18]. This process requires sufficient base pairing between the sgRNA and the target DNA. In contrast, the CRISPRi system, which we are focusing on, involves inhibiting gene expression rather than cleaving DNA. In CRISPRi, off-target binding is more prone to occur compared to off-target cleavage [19], presenting challenges in accurately pinpointing off-target binding occurrences. 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 [20,21] 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 [4,22], we have conducted a comprehensive analysis to discern the impact of mismatches on binding affinity. From these derived principles, we have developed CRISPRi Guide Library Designer (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 1,426 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). These files can be easily obtained from a variety of publicly available databases such as NCBI [23], EMBL [24], DDBJ [25], etc. In the case of a newly sequenced organism, they can also be generated using genome annotation pipeline like PGAP [26].

sgRNA library design workflow

As shown in Figure 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 [27,28] and are grouped into a cluster. This is achieved by employing BLASTN [29] 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,

corresponding to canonical and non-canonical PAM sequences [30]. 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 [9]. An illustrative example is presented in Figure S1. Following this, candidate sgRNAs undergo a rigorous quality control process to eliminate those with potential off-target hits (see "Quality control of sgRNA library" section). 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.

Quality control of sgRNA library

The quality control of sgRNAs (Figure 2B) involves three alignments according to the user defined target strand (template or non-template) using SeqMap [31]. For instance, if the non-template strand is selected, FASTA1 would be aligned with FASTA1, FASTA2, and FASTA3. In each alignment, sequence pairs with two or fewer mismatches are picked for subsequent scoring evaluations, considering that two or more mismatches in the seed region substantially weaken the binding affinity [32]. Scoring metrics are defined based on the distance of mismatches to the PAM, considering that mismatches are generally better tolerated at the 5′ end of the sgRNA than at the 3′ end [13]. The sgRNA regions are categorized from the 3’ end to the 5’ end as follows: region I (7 nt), region II (5 nt), and region III (the remainder of the sgRNA) [10,33]. For NGG PAM, mismatch penalties for region I, II, and III are 8, 4.5, and 2.5, respectively. Conversely, for NAG PAM, the penalties are 10,

7, and 3, respectively (Figure S2). An off-target site is identified when the penalty score falls below the user-defined threshold, with a recommended range of 18–21 for designing genome-scale libraries.

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, where each sgRNA is associated with its target gene and ranked according to its proximity to the start codon. This ranking is because sgRNAs targeting closer to the transcription start sites have shown higher knock-down activity [34]. In addition to the table, GLiDe also generates an interactive graphical interface using D3GB [35]. 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. 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 [4]. The other plasmid was derived from pN20test-114mCherry-r0-m1 [22], 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 (Figure S3). 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 hours. 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 OD_{600} = 0.02 and then grown for 4 hours to the logarithmic phase. After cultivation,

5 μL of culture medium from each well was diluted into 200 μL of phosphate-buffered saline (PBS). 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 [36] 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 (Equation 1).

$$f(x) = \lambda N(\mu_1, \sigma_1^2) + (1-\lambda) N(\mu_2, \sigma_2^2) \quad (1)$$

The mean expression strength was calculated with Equation 2.

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

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

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

$$repression\ rate = \frac{Mean_{Ri} - Mean_{Ri-PC}}{Mean_{Ri-NC} - Mean_{Ri-PC}} \quad (3)$$

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 (Figure 1).

Input and configuration

GLiDe includes an extensive built-in database harboring over 1,426 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 (see **Methods**); (iii) GC limits, which represent the upper and lower boundaries for the GC content of each sgRNA; (iv) spacer length, which defines the length of the sgRNA being designed; 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 [30], we offer the alternative option for other applications, such as genome editing, base editing [37,38], or dynamic imaging [39], where there may be less strand preference.

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.

Process and output

Upon configuration, a server request is initiated to commence the workflow (Figure 2). Anticipated processing time for a standard genome-scale library design range from 5 to 20 minutes, varying based on the size of the genome. The result sgRNA library is thoughtfully presented in a tabular format (Figure 1B), 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 [35], 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 [18], 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 [14] 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 [22]. In this system, we inserted 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 (Figure 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 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 (Figure 3B, $P < 1.36 \times 10^{-5}$ for all groups, two-tailed *t*-test; Figure S4). The repression rates ranged from 51.12% (R1) to 90.70% (R4), with an average of 69.47% (see **Methods**). The blank control groups exhibited similar mCherry fluorescence intensities (Figure 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 is designed to create genome-scale CRISPRi sgRNA libraries, with the goal of supporting functional genomic studies for prokaryotes. It has an intuitive web interface and requires minimal input. In addition, it incorporates a quality control mechanism grounded in experimental insights, to avoid off-target bindings. We 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 [40,41], bioprocesses optimization in industrially relevant microorganisms [42,43], or identification of complex genetic interactions [44].

However, although GLiDe showed a high accuracy for the identification of potential off-target hits, it's important to acknowledge a potential limitation in GLiDe's quality control. While it is widely accepted that the PAM-proximal region plays a substantial role in determining sgRNA binding affinity, the current scoring system primarily relies on empirical data, lacking a comprehensive theoretical foundation. Furthermore, the model assumes equal contributions from different mismatched nucleotide pairs, despite known variations in stability among these pairs [45]. Certain mismatches, such as rU/dG, rG/dT, and rG/dG, exhibit stabilities similar to Watson-Crick pairs [46], a nuance not fully accounted for in the existing framework. 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) [22]. This approach allows us to make rational predictions about binding affinities. However, especially for multi mismatches, the

nearest-neighbor parameters for RNA/DNA mismatches are still insufficient [47,48]. Therefore, a comprehensive measurement of parameters might be required to help predict the binding affinities of sgRNAs.

DATA AVAILABILITY

Raw data of the flow cytometry assays and plasmid maps related to this work can be accessed via GitHub (https://github.com/falconxtj/GLiDe).

ACKNOWLEDGEMENT

We thank Dr. T. Wang (Tsinghua University) for building the original demo of the sgRNA design tool. We also thank members from the C. Zhang and H. Xing laboratories for critical discussions of this work.

FUNDING

This work was supported by the National Key R&D Program of China [2023YFC3402400 to C.Z.]; the National Natural Science Foundation of China [21938004 to X.X.]; the National Natural Science Foundation of China [U2032210 to C.Z.].

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this article.

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; 6:767–772
2. Garst AD, Bassalo MC, Pines G, et al. Genome-wide mapping of mutations at single-nucleotide resolution for protein, metabolic and genome engineering. Nat Biotechnol 2017; 35:48–55

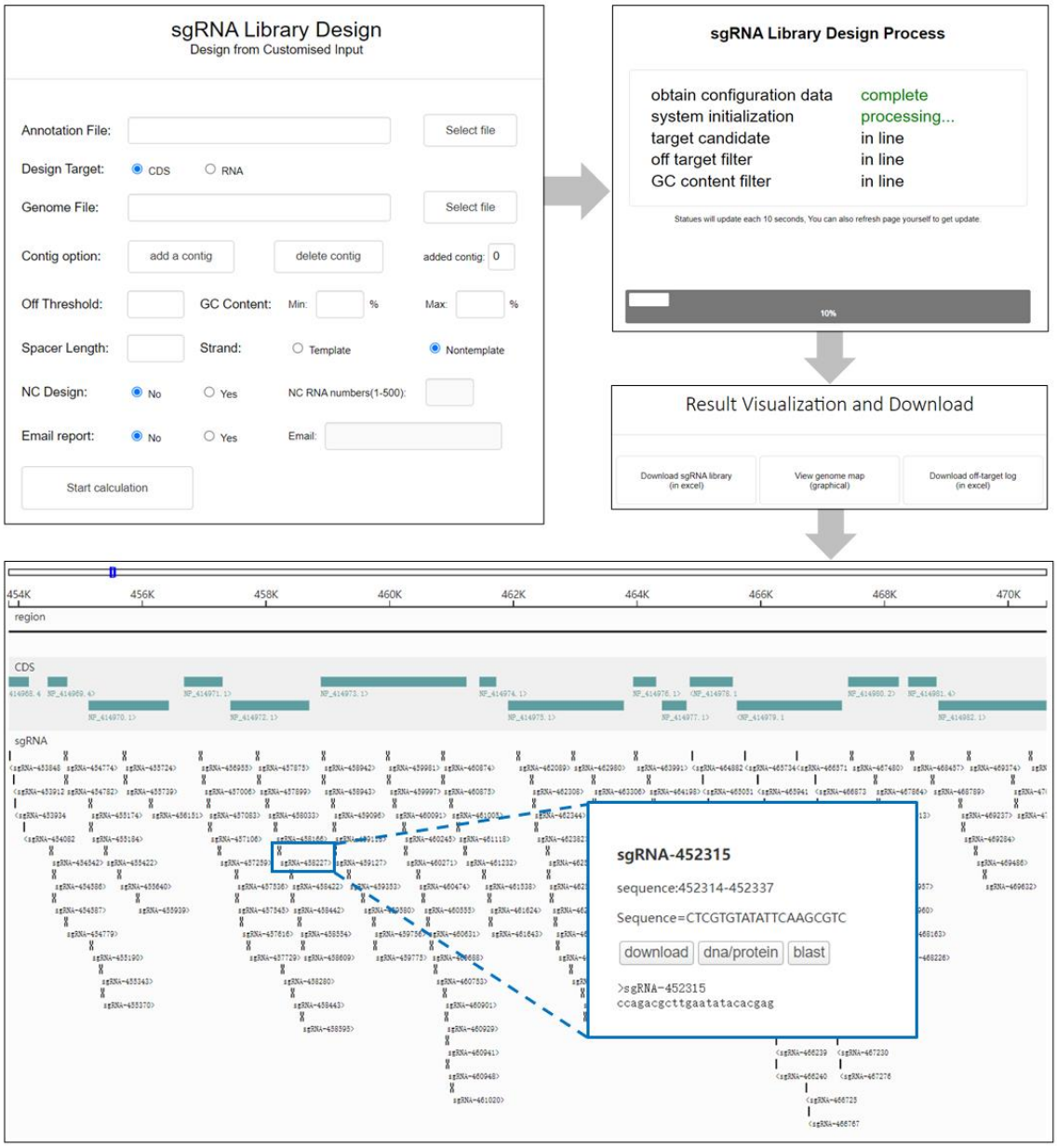
3. Freed EF, Winkler JD, Weiss SJ, et al. Genome-Wide Tuning of Protein Expression Levels to Rapidly Engineer Microbial Traits. ACS Synth Biol 2015; 4:1244–1253
4. Wang T, Guan C, Guo J, et al. Pooled CRISPR interference screening enables genome-scale functional genomics study in bacteria with superior performance. Nat Commun 2018; 9:2475
5. Jinek M, Chylinski K, Fonfara I, et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science (1979) 2012; 337:816–821
6. Hawkins JS, Silvis MR, Koo BM, 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
7. Lian J, Schultz C, Cao M, et al. Multi-functional genome-wide CRISPR system for high throughput genotype–phenotype mapping. Nature Communications 2019 10:1 2019; 10:1–10
8. Chen P, Michel AH, Zhang J. Transposon insertional mutagenesis of diverse yeast strains suggests coordinated gene essentiality polymorphisms. Nature Communications 2022 13:1 2022; 13:1–15
9. de Bakker V, Liu X, Bravo AM, et al. CRISPRi-seq for genome-wide fitness quantification in bacteria. Nature Protocols 2021 17:2 2022; 17:252–281
10. Hsu PD, Scott DA, Weinstein JA, et al. DNA targeting specificity of RNA-guided Cas9 nucleases. Nat Biotechnol 2013; 31:827–832
11. Shalem O, Sanjana NE, Zhang F. High-throughput functional genomics using CRISPR–Cas9. Nat Rev Genet 2015; 16:299–311
12. Heigwer F, Kerr G, Boutros M. E-CRISP: fast CRISPR target site identification. Nat Methods 2014; 11:122–123
13. Sander JD, Joung JK. CRISPR-Cas systems for editing, regulating and targeting genomes. Nat Biotechnol 2014; 32:347–355
14. Montague TG, Cruz JM, Gagnon JA, et al. CHOPCHOP: a CRISPR/Cas9 and TALEN web tool for genome editing. Nucleic Acids Res 2014; 42:W401–W407
15. Concordet J-P, Haeussler M. CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. Nucleic Acids Res 2018; 46:W242–W245
16. 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; 30:1473–1475
17. Liu H, Wei Z, Dominguez A, et al. CRISPR-ERA: a comprehensive design tool for CRISPR-mediated gene editing, repression and activation: Fig. 1. Bioinformatics 2015; 31:3676–3678
18. Sternberg SH, LaFrance B, Kaplan M, et al. Conformational control of DNA target cleavage by CRISPR–Cas9. Nature 2015; 527:110–113
19. Boyle EA, Andreasson JOL, Chircus LM, et al. High-throughput biochemical profiling reveals sequence determinants of dCas9 off-target binding and unbinding. Proceedings of the National Academy of Sciences 2017; 114:5461–5466
20. Struhl K. Fundamentally Different Logic of Gene Regulation in Eukaryotes and Prokaryotes. Cell 1999; 98:1–4
21. Kuzminov A. The Precarious Prokaryotic Chromosome. J Bacteriol 2014; 196:1793–1806
22. Feng H, Guo J, Wang T, et al. Guide-target mismatch effects on dCas9-sgRNA binding activity in living bacterial cells. Nucleic Acids Res 2021; 49:1263–1277
23. Sayers EW, Bolton EE, Brister JR, et al. Database resources of the National Center for Biotechnology Information in 2023. Nucleic Acids Res 2023; 51:D29–D38
24. Thakur M, Bateman A, Brooksbank C, et al. EMBL's European Bioinformatics Institute (EMBL-EBI) in 2022. Nucleic Acids Res 2023; 51:D9–D17

25. Tanizawa Y, Fujisawa T, Kodama Y, et al. DNA Data Bank of Japan (DDBJ) update report 2022. Nucleic Acids Res 2023; 51:D101–D105
26. Li W, O'Neill KR, Haft DH, et al. RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. Nucleic Acids Res 2021; 49:D1020–D1028
27. McTavish H, LaQuier F, Arciero D, et al. Multiple copies of genes coding for electron transport proteins in the bacterium Nitrosomonas europaea. J Bacteriol 1993; 175:2445–2447
28. Schrider DR, Hahn MW. Gene copy-number polymorphism in nature. Proceedings of the Royal Society B: Biological Sciences 2010; 277:3213–3221
29. Johnson M, Zaretskaya I, Raytselis Y, et al. NCBI BLAST: a better web interface. Nucleic Acids Res 2008; 36:W5–W9
30. Qi LS, Larson MH, Gilbert LA, et al. Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. Cell 2013; 152:1173–1183
31. Jiang H, Wong WH. SeqMap: mapping massive amount of oligonucleotides to the genome. Bioinformatics 2008; 24:2395–2396
32. Boyle EA, Andreasson JOL, Chircus LM, et al. High-throughput biochemical profiling reveals sequence determinants of dCas9 off-target binding and unbinding. Proceedings of the National Academy of Sciences 2017; 114:5461–5466
33. Gilbert LA, Horlbeck MA, Adamson B, et al. Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. Cell 2014; 159:647–661
34. Guo J, Wang T, Guan C, et al. Improved sgRNA design in bacteria via genome-wide activity profiling. Nucleic Acids Res 2018; 46:7052–7069
35. Barrios D, Prieto C. D3GB: An Interactive Genome Browser for R, Python, and WordPress. J Comput Biol 2017; 24:447–449
36. Feng H, Li F, Wang T, et al. Deep-learning–assisted Sort-Seq enables high-throughput profiling of gene expression characteristics with high precision. Sci Adv 2023; 9:
37. Gaudelli NM, Komor AC, Rees HA, et al. Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. Nature 2017; 551:464–471
38. Komor AC, Kim YB, Packer MS, et al. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. Nature 2015 533:7603 2016; 533:420–424
39. Chen B, Gilbert LA, Cimini BA, et al. Dynamic Imaging of Genomic Loci in Living Human Cells by an Optimized CRISPR/Cas System. Cell 2013; 155:1479–1491
40. Peters JM, Colavin A, Shi H, et al. A Comprehensive, CRISPR-based Functional Analysis of Essential Genes in Bacteria. Cell 2016; 165:1493–1506
41. Rousset F, Cabezas-Caballero J, Piastra-Facon F, et al. The impact of genetic diversity on gene essentiality within the Escherichia coli species. Nat Microbiol 2021; 6:301–312
42. Li S, Jendresen CB, Landberg J, et al. Genome-Wide CRISPRi-Based Identification of Targets for Decoupling Growth from Production. ACS Synth Biol 2020; 9:1030–1040
43. Donati S, Kuntz M, Pahl V, 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
44. Jaffe M, Dziulko A, Smith JD, et al. Improved discovery of genetic interactions using CRISPRiSeq across multiple environments. Genome Res 2019; 29:668–681
45. Sugimoto N, Nakano S, Katoh M, et al. Thermodynamic Parameters To Predict Stability of RNA/DNA Hybrid Duplexes. Biochemistry 1995; 34:11211–11216
46. Sugimoto N, Nakano M, Nakano S. Thermodynamics−Structure Relationship of Single Mismatches

in RNA/DNA Duplexes. Biochemistry 2000; 39:11270–11281
47. Watkins NE, Kennelly WJ, Tsay MJ, et al. Thermodynamic contributions of single internal rA·dA, rC·dC, rG·dG and rU·dT mismatches in RNA/DNA duplexes. Nucleic Acids Res 2011; 39:1894–1902
48. Xiang T, Feng H, Xing X, et al. Thermodynamic Parameters Contributions of Single Internal Mismatches In RNA/DNA Hybrid Duplexes. bioRxiv 2022; 2022.11.25.517909

TABLE AND FIGURES LEGENDS

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

Figure 1. Workflow 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.

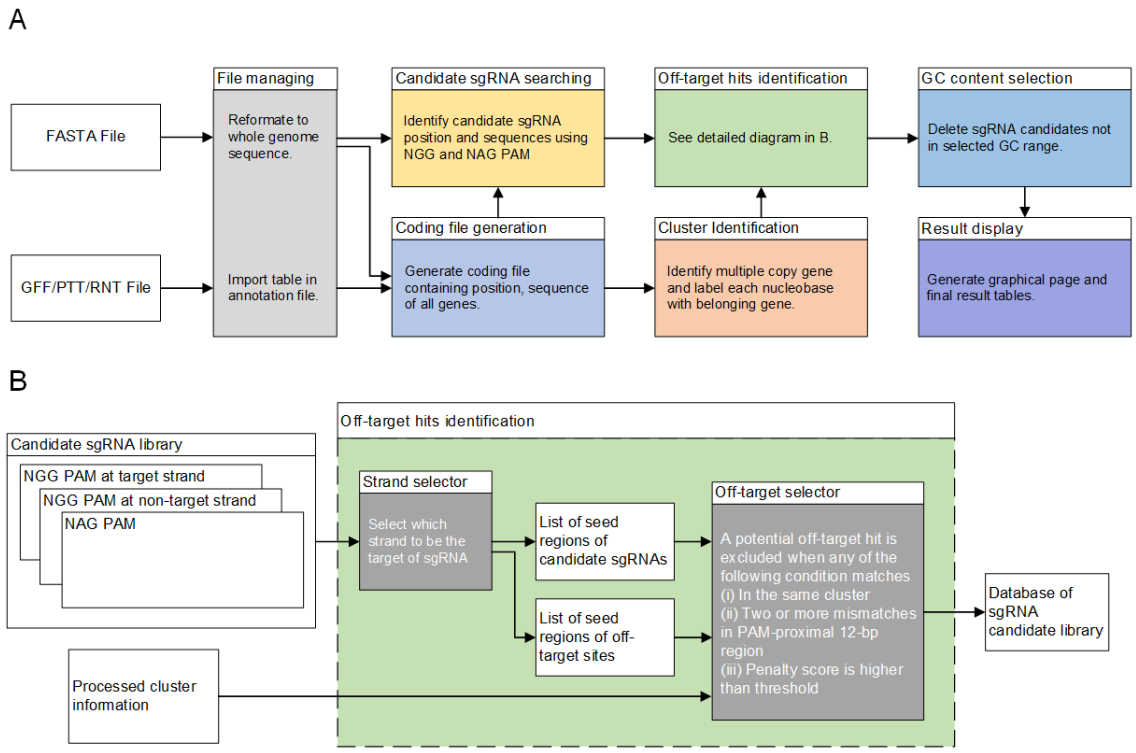


Figure 2. Functional blocks of GLiDe. **(A)** The GLiDe flowchart encompasses the processing of the FASTA file and the annotation file through multiple modules. These modules conduct candidate sgRNA searches, quality control (identifying off-target hits and selecting GC content), and ultimately present 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.

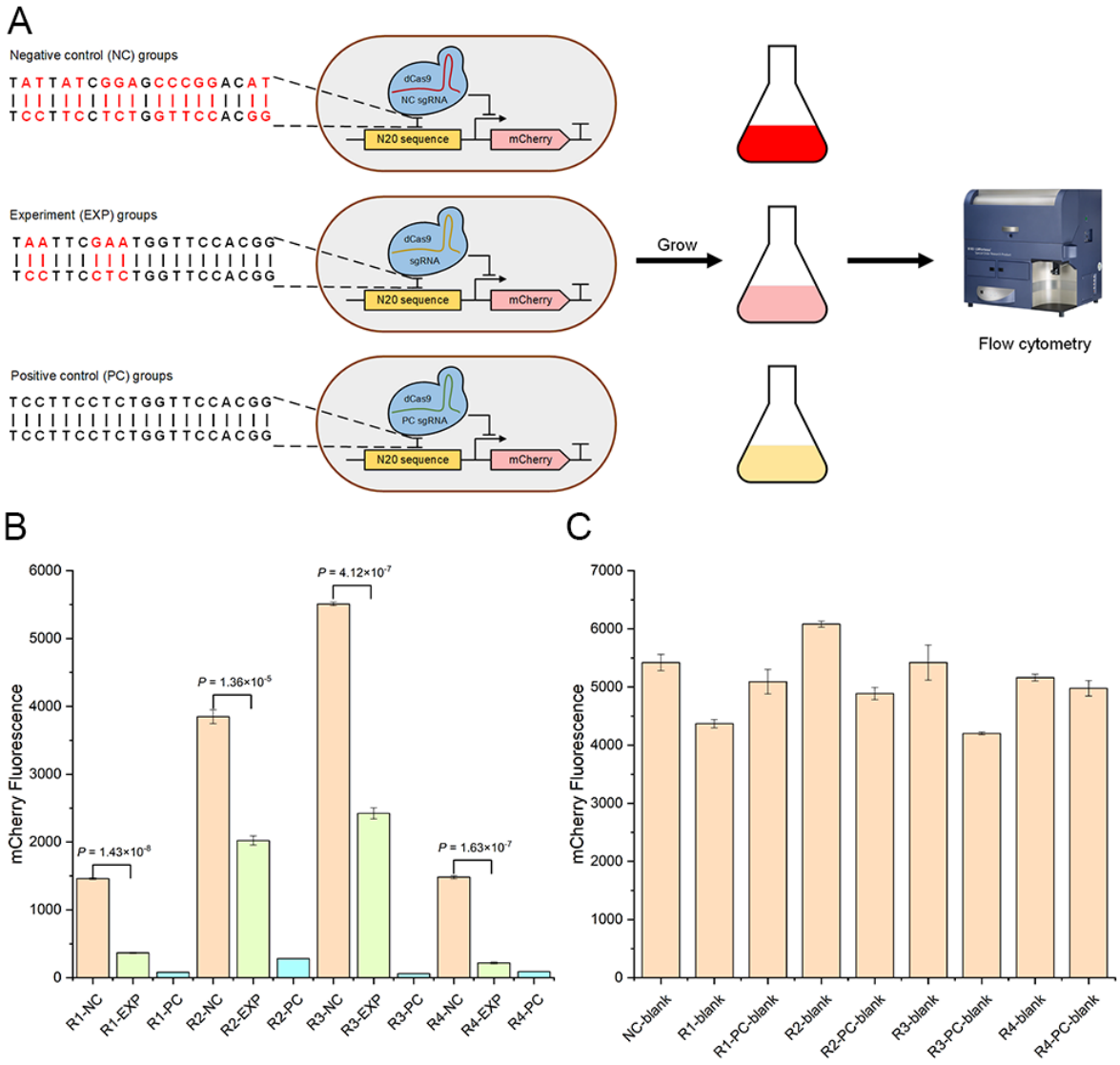


Figure 3. Assessment of off-target binding affinities using flow cytometry assay. **(A)** The binding affinity of each mismatched sgRNA was compared with positive controls and negative controls. Fluorescence intensities of the three groups were measured by flow cytometry. **(B)** All sgRNAs from the EXP groups exhibited significant off-target binding affinities ($P < 10^{-4}$ 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).

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

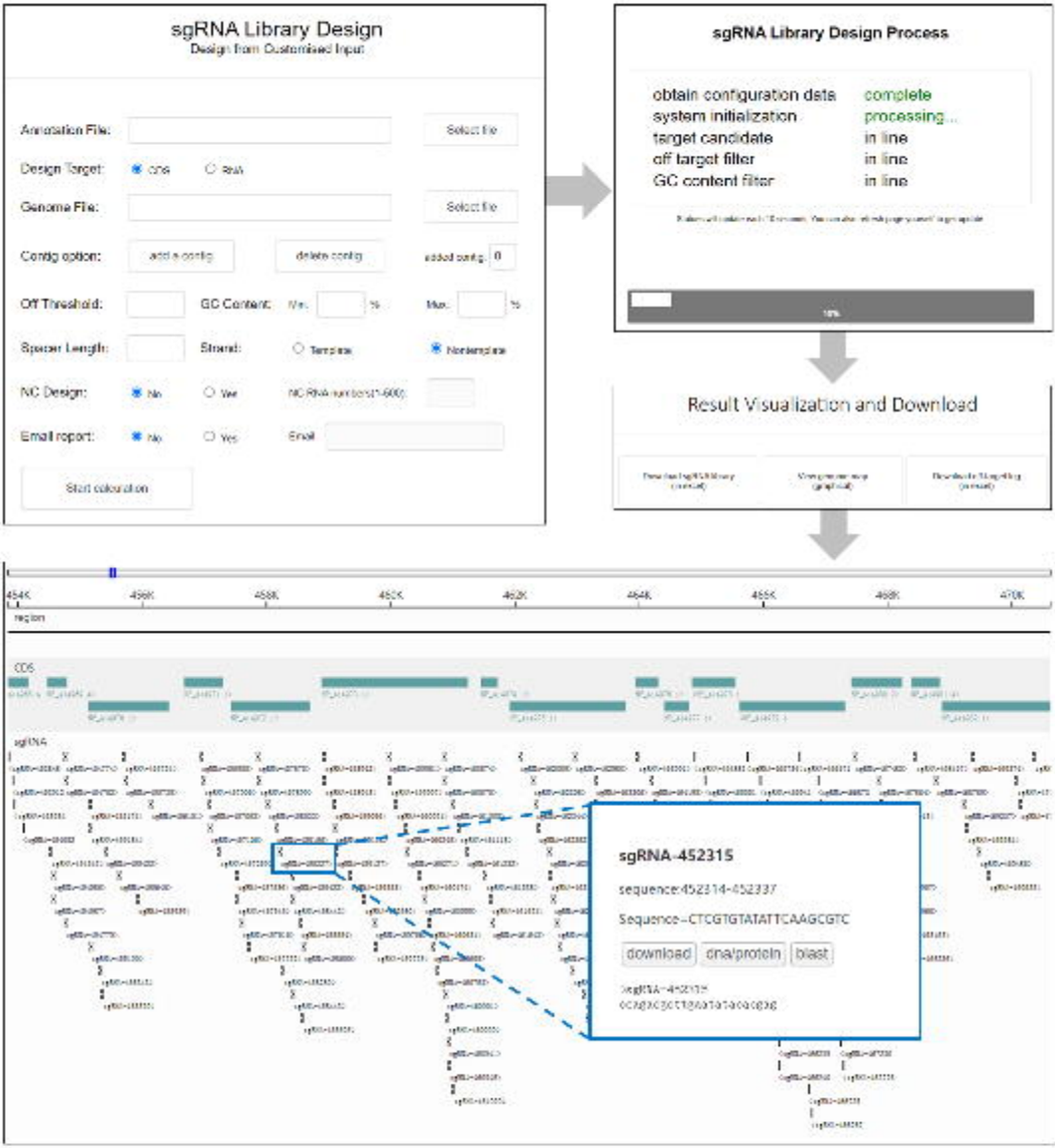
No.	Design tool	Rank	Target gene	sgRNA sequence	Off-target gene	Potential off-target hits discovered by GLiDe
R1	chopchop	1	*Cgl0025*	T***AA***TTC***GAA***TGGTTCCACGG	*Cgl2364*	T***CC***TTC***CTC***TGGTTCCACGG
R2	chopchop	3	*Cgl0044*	***GCACGAT***GCCCACCACACCG	*Cgl1968*	***CGGATGG***GCCCACCACACCG
R3	Synthego	3	*Cgl0005*	***TTCT***G***CGC***ATGGCTTCGGCG	*Cgl0369*	***GCGC***G***GCG***ATGGCTTCGGCG
R4	Synthego	2	*Cgl0059*	***GAG***CT***CAG***ATCGCTCAACAT	*Cgl1680*	***TCT***CT***TGA***ATCGCTCAACAT

Table 2. sgRNAs and N20 sequences of each test group

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

R4-PC	TCTCTTGAATCGCTCAACAT	
NC-blank	TATTATCGGAGCCCGGACAT	
R1-blank	TAATTCGAATGGTTCCACGG	
R1-PC-blank	TCCTTCCTCTGGTTCCACGG	
R2-blank	GCACGATGCCCACCACACCG	
R2-PC-blank	CGGATGGGCCCACCACACCG	CATGTATTATACACGAAGTT
R3-blank	TTCTGCGCATGGCTTCGGCG	
R3-PC-blank	GCGCGGCGATGGCTTCGGCG	
R4-blank	GAGCTCAGATCGCTCAACAT	
R4-PC-blank	TCTCTTGAATCGCTCAACAT	

A



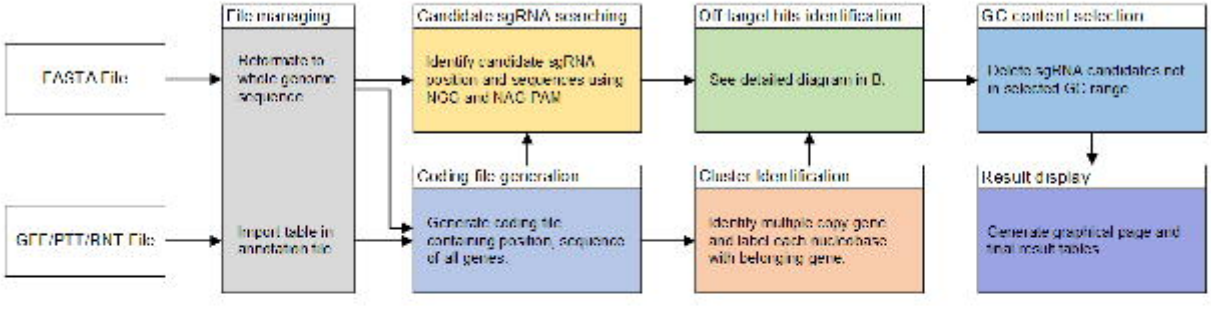
B

Location	Strand	Length	Gene	Guide sequence	Guide position	Guide number
20815 21078	-	37	gwT	AACTACCCATCATACACCCA	20983	1
190..255	+	21	thrL	CGCACCGTACCTGTGGTAA GGTGGTGCTAATGCGTTTCA	249 210	2

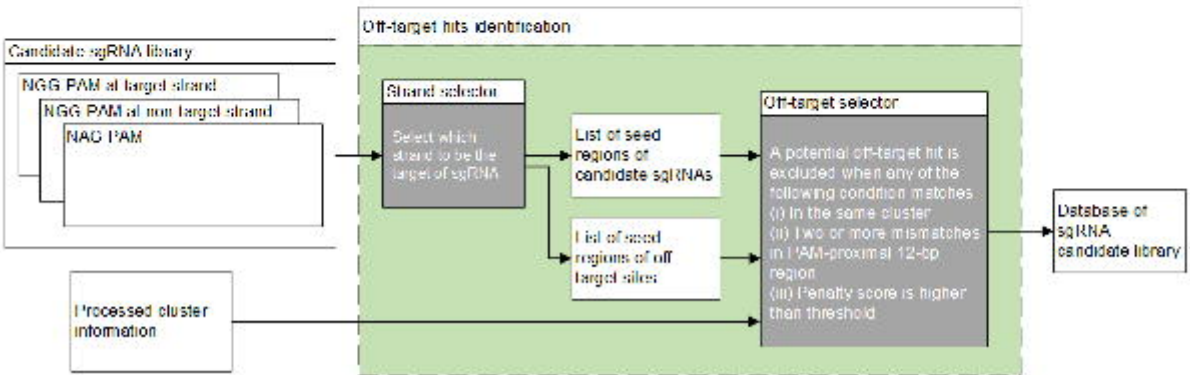
C

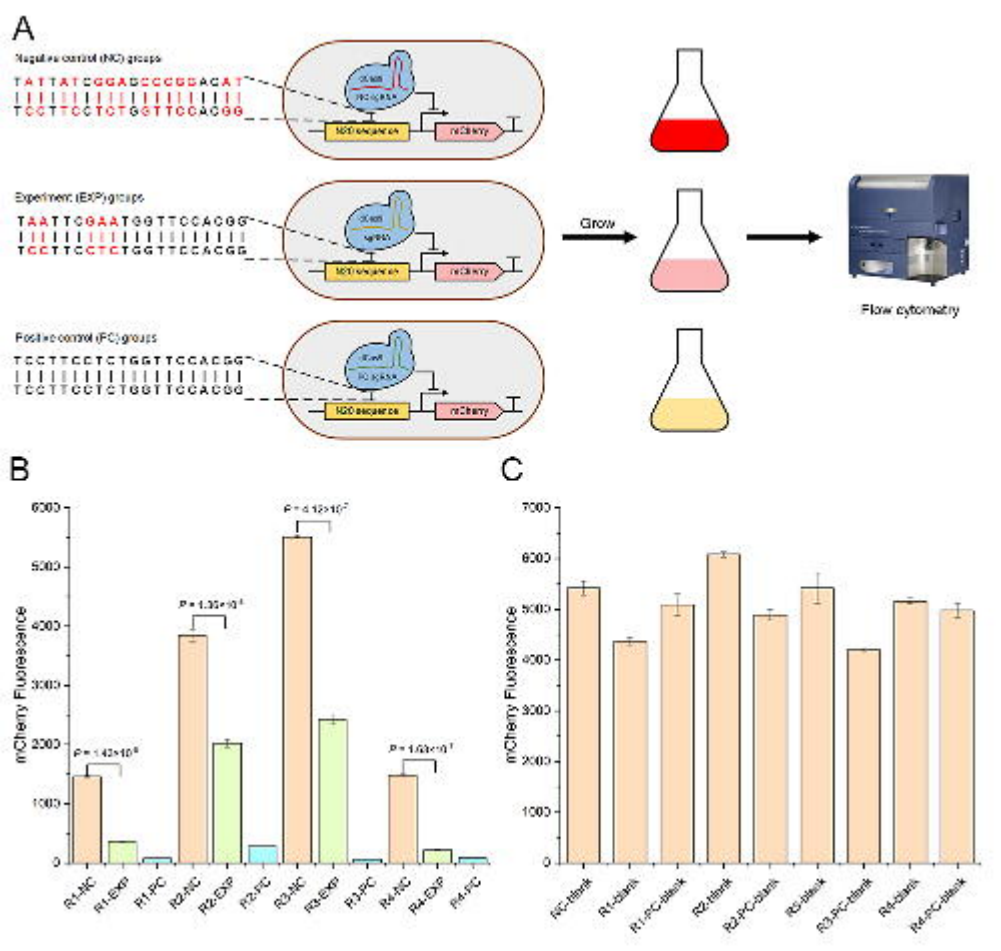
Guide sequence	Potential off target hit(s)	penalty Score	Off target number
TTAAAACCACTCACTTCCTC	TCAACCCCTCCCACTTCCTC TAAACAACACTCACATCCTC	19.0 15.5	2
ATCTCGGCATTCAATTTCTT	ATAGAGGAAGTGAATTTCTT CAAACGCCAGTCAATTTCTT	19.0 17.0	2

A


FASTA File
GFF/PTT/RNT File
File managing
Relormate to whole genome sequence
Import table in annotation file
Candidate sgRNA searching
Identify candidate sgRNA position and sequences using NGG and NAG PAM
Coding file generation
Generate coding file containing position, sequence of all genes.
Off target hits identification
See detailed diagram in B.
Cluster identification
Identify multiple copy gene and label each nucleotide with belonging gene.
GC content selection
Delete sgRNA candidates not in selected GC range
Result display
Generate graphical page and final result tables

B


Candidate sgRNA library
NGG PAM at target strand
NGG PAM at non target strand
NAG PAM
Off-target hits identification
Strand selection
Select which strand to be the target of sgRNA
List of seed regions of candidate sgRNAs
List of seed regions of off target sites
Off-target selector
A potential off-target hit is excluded when any of the following condition matches
(i) In the same cluster
(ii) Two or more mismatches in PAM-proximal 12-bp region
(iii) Penalty score is higher than threshold
Database of sgRNA candidate library
Processed cluster information


A
Negative control (NC) groups
TATTATCGGAGCCCGGACAT
TCCTTCCTCTGGTTCCACGG
Experiment (EXP) groups
TAATTCGAATGGTTCCACGG
TCCTTCCTCTGGTTCCACGG
Positive control (PC) groups
TCCTTCCTCTGGTTCCACGG
TCCTTCCTCTGGTTCCACGG
N20 sequence
mCherry
Grow
Flow cytometry
B
mCherry Fluorescence
R1-NC
R1-EXP
R1-PC
R2-NC
R2-EXP
R2-PC
R3-NC
R3-EXP
R3-PC
R4-NC
R4-EXP
R4-PC
C
mCherry Fluorescence
NC-blank
R1-blank
R1-PC-blank
R2-blank
R2-PC-blank
R3-blank
R3-PC-blank
R4-blank
R4-PC-blank