CRISPRi-FGP: web-based genome-scale CRISPRi sgRNA design and validation tool in prokaryotes

Tongjun Xiang,[1#] Huibao Feng,[1#] Xin-hui Xing,[2] Chong Zhang[1*]

[1]MOE Key Laboratory for Industrial Biocatalysis, Institute of Biochemical Engineering, Department of Chemical Engineering, Tsinghua University, Beijing 100084, China; Center for Synthetic and Systems Biology, Tsinghua University, Beijing 100084, China.

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

[#]These authors contributed equally

ABSTRACT

Pooled CRISPRi with high-throughput sequencing is emerging as a novel functional genomic study method in prokaryotes. Due to the diversity of sgRNA activity and off-target effect, it is crucial to have a robust sgRNA library designing tool. Here we propose a web tool named CRISPRi-FGP (http://www.thu-big.net/sgRNA_design/), to assist sgRNA library design in prokaryotes. CRISPRi-FGP applies a build-in database containing 1,300 common prokaryotes as standard designing source, while also accepts customized designing. It employs a scoring metrics based on experiment data, together with specially trimed algorithm for off-target selection both in high quality and speed. CRISPRi-FGP exhibited high accuracy and robustness on sgRNA library designing in multiple prokaryotes. Finally, an experiment comparing CRISPRi-FGP and existing tool was conducted and showed CRISPRi-FGP has improved ability in recognizing off-target sites. All these results showed CRISPRi-FGP with the potential to be a precise and robust tool for high-quality sgRNA library design in prokaryotes.

INTRODUCTION

With the boosting accumulation of bacterial genomes, functional genomic screening is drawing more attention as a powerful strategy for genotype-phenotype mapping, which provides insights into microbiology and engineering of microorganisms (1–3). To investigate the function of a specific gene, the straightforward way is to perturb it and assay the corresponding phenotypic consequence subsequently. Recently, CRISPRi-seq method has been established for associating transcriptional perturbations with corresponding phenotypic changes in a massively parallel manner (4). It starts with a pooled single guided RNA (sgRNA) library, in which each sgRNA is designed to guide dead Cas9 (dCas9) protein to a specific genetic locus, resulting in transcriptional inhibition of the target gene. After transformation of the library, different sgRNAs would show varying performances in subsequent high-throughput screening such as fluorescence-activated cell sorting (FACS)

and fitness screening, which can be further identified through next generation sequencing (NGS). sgRNA guide the dCas9 to the genomic region flanked by a 3'NGG protospacer adjacent motif (PAM) via 20-bp Watson–Crick base pairing (5), however, false positive results sometimes occur owing to unexpected binding. This phenomenon, knowing as "off-target", is because binding process can tolerant several mismatches (6). Hence sgRNA design must avoid other potential off-target hits in the whole genome.

To avoid false positive in CRISPRi-seq, a high accuracy and robustness sgRNA library designing tool is needed in pooled CRISPRi. Previously, multiple tools have been established for sgRNA design, including web-based single genome sgRNA library designing tools (7,8), localized software for whole-genome or concentrated sgRNA design (9,10), and web-based CRISPR/Cas9 sgRNA library designing tool (11,12). However, existing tools only use off-target search rules learned from previous studies performed in eukaryotic cells. Considering the fundamental differences of prokaryotic and eukaryotic genomes, such as less organized chromosome structure (13,14), which enhances the accessibility of genome, or the variety in intracellular conditions, the applicability of existing tools on prokaryotes is problematic. Furthermore, recent researches reveal that Cas protein has impact on off-target activity by conformational change in the HNH nuclease domain while recognition of sgRNA-DNA which enables productive engagement with the scissile phosphate (15,16). While existing tools concentrate more on off-target cleavage (for CRISPR/Cas), conformation gating is present when used for off-target binding (for CRISPR/dCas), indicating that off-target searching rules used by present tools may not have good performance in CRISPRi.

To address the limitations of existing tools in sgRNA library design, we have developed CRISPRi-Functional Genomics in Prokaryotes (CRISPRi-FGP), a web-based tool that allows user to easily and rapidly designs the genome-scale or single gene sgRNA library for pooled CRISPRi screening in prokaryotes. CRISPRi-FGP employs a rigorous off-target search algorithm, which specially optimized for prokaryotes, to predict the specificity of each target site in genome (7,17–19). Furthermore, CRISPRi-FGP has a user-friendly graphical interface, together with a build-in database of 1,300 common microorganism genome. It requires no programming skill or any supporting software or package, making CRISPRi-FGP a handy tool for experimental biologists with no or limited programming expertise. Moreover, experiment showed that CRISPRi-FGP has improved performance in predicting potential off-target sites containing mismatches comparing with other existing sgRNA design tools. This tool, with a solid experimental demonstration, achieves the state-of-the-art performances in predicting off-target activities of sgRNA while features. It has the potential for defining gene function for potentially any prokaryotic organism.

MATERIALS AND METHODS

Reference files and data pre-processing

Two reference files are needed for CRISPRi-FGP: sequence file and annotation file. Sequence file is a single contig genome file for designing, for both protein and RNA coding gene, a FASTA file (genomes as single contig) is required. Once the FASTA file containing the genomic sequence is imported, page breaks are removed, and the text is reformatted to be a complete genome sequence for further progress. Another needed file is annotation

file, which contains the features of all gens in microorganism including genomic name, coding type, location, strand, phase, functional attributes, etc. CRISPR-FGP accepts two types of standard file formats: (i) GFF3 file and (ii) PTT or RNT files, both are wildly used. Both files can be obtained from a variety of sources like publicly available databases (such as NCBI, EMBL and DDBJ), sequencing results or artificially genome edited microorganisms. Once the two files are uploaded, a coding sequence list is constructed for correlation of genomic features with the DNA sequence. CRIPSRi-FGP goes through the annotation file to get position of each gene, and go to genome sequence file for gene sequence. This table contains sequences of all genes required to design sgRNAs and label each gene with a unique ID.

Workflow of sgRNA library design in CRISPR-FGP

CRISPR-FGP has four major steps in workflow, functional block workflow is shown in Figure 2. The first step is Cluster identification, some genes have multiple copies in the genome (20,21), thus off-target within those genes should be filtered. In CRISPR-FGP, sequences have high similarity would be considered as the same function and be identified as cluster. So that sgRNAs targeting to these cluster genes could be kept in off-target selection. CRISPRi-FGP uses BLASTN (22) with default parameters for sequences comparation. We used a strict threshold of 0.001 evalue, >95% identity, >95% hit coverage and >95% query coverage. A list of all clusters in *E. coli* MG1655 is listed in Supplementary Table S1. Each nucleobase is then labeled with the gene ID of its belonging gene. If one nucleobase is not belonged to any gene, it would be labeled as 'none'. The second step is candidate sgRNAs selection, candidate sgRNAs and potential off-target locations are identified using regular expressions. Two types of guiding sequences: N20NGG and N20NAG, corresponding to canonical and non-canonical PAM sequences of CRISPR-dCas9 system (23), are identified from both chains of genome sequence. The candidate sgRNAs are categorized into three types based on both PAM sequence and its targeted strain: (i) sgRNAs with NGG PAM sequence and targeting to non-template strand; (ii) sgRNAs with NGG PAM sequence and targeting to template strand, and (iii) sgRNAs with NAG PAM sequence. An example of categorize is shown in Supplementary Figure S1. Three separate FASTA files are created to save the above three kinds of sgRNAs. Because of mismatches in seed region, the PAM-proximal 12-bp of each sgRNA, have a severe impact on binding or cleavage (9), only this region is saved in files. The third step is off-target selection, including mismatched N20 sequences and non-canonical PAM sequence. In off-target selection step, each seed sequences from the mentioned-above FASTA files are compared with all other sequences in all three FASTA files using Seqmap (18). Previous experiments showed that two or more mismatches in the seed region significantly weaken the binding activity (7,16). When two sgRNAs in different cluster have lower than two mismatches in their seed regions, a second selection is conducted using the full 20bp sequence of two sgRNAs. Scoring metrics inferred from literature were designed to evaluate off-target activity (6,19). Briefly, the protospacer region of potential off-target sites are divided into three parts from 5'- end to 3'- end: region III (8 nt), region II (5 nt) and region I (7 nt), respectively, according to the distance to the PAM site. Mismatches from different region are considered to have different effect on off-target activity because

mismatches are generally better tolerated at the 5′ end of sgRNA than at the 3′ end (10). If the PAM site of the off-target 23-mer is “NGG”, the mismatch penalty of region I, II and III were set as 2.5, 4.5, and 8, respectively, and the penalty set for “NAG” were 3, 7, and 10, respectively (see Supplementary Figure S2). The off-target site was considered significant when Σ (penalty × mismatch) < threshold (4). CRISPRi-FGP accepts user to personally determine the threshold. Users are free to choose lower threshold for mor sgRNAs or higher threshold for lower potential of off-targets. Basically, we recommend a threshold of 18-21 for genome-scale library design. Once the off-target selection is complete, the last step, GC-content threshold filter is conducted on selected sgRNAs using the upper and lower limits defined by user.

Negative control sgRNA library design

Negative control sgRNAs are sgRNAs with no significant target across the genome, it can serve as a negative control in subsequent data analysis expressing the impact of experimental procedures such as transforming plasmids on the cell phenotype in the experiment. CRISPRi-FGP designs those sgRNAs by generating random N20 sequences and use the same off-target and GC content selection as mentioned above. The penalty score is set to 25, GC limits are the same as sgRNA library. Five or more consecutive identical bases is avoided. The number of negative control sgRNAs are defined by users.

Results visualization and download

The final designed sgRNA library is arranged in a result table. In the result table, sgRNAs are classified and counted with their targeting gene. For each gene, sgRNAs are ranked by their distance to ATG codon, this is because sgRNAs targeting closer to starting point of transcription generally have better knock-down activity (17). Another result table contains the sgRNAs with possible off-target sequences and their off-target penalty score. Both tables are saved as tables and temporarily saved in the server. The two tables can be either downloaded by user from the website or sent to email address determined by user.

Individual sgRNA can be inspected in a graphical interactive page based on D3GB Genome Browser (24). The whole genome and all sgRNAs are displayed in a zoomable and interactable page so that uses can easily zoom in their interested regions or use the searching function to locate a specific target gene. Each sgRNA and gene are displayed at its appropriate location in genome. Clicking on any individual sgRNA or gene target site results in more detailed information of sequence, position, etc.

Identification of potential off-target sgRNA in existing tools

The experiment was conducted in *Escherichia coli* strain MCm. For avoiding knock down of endogenous genes, we used CHOPCHOP (11), Synthego and CRISPRi-FGP to design sgRNA for *Corynebacterium glutamicum* strain ATCC 13032, respectively. The off-target penalty score was 18. For each gene of *C. glutamicum*, sgRNAs have high rank in CHOPCHOP and Synthego but are considered to have potential off-target sites in CRISPRi-FGP are chosen as candidate, and a second comparation with *E. coli* genome to ensure no off-target with endogenous genes. Finally, four sgRNAs, R1-R4, and potential

off-target pairs are chosen, all are within the top three recommended of CHOPCHOP and Synthego (listed in Table 1). Each sgRNA also has positive (complementary pairing with the off-target sequence, different for each test group) and negative control (no annealing with the off-target sequence or endogenous genes, the same sgRNA for each test group). Moreover, to ensure all sgRNAs have no influence on host metabolism, an extra blank control is conducted for all sgRNA used in this work.

DNA manipulations and reagents

Plasmid extraction and DNA purification were performed using kits from Vazyme. PCR reactions were carried out using KOD one from TOYBO Life Science. All PCR primers were ordered from Genewiz (Supplementary Table S2). Plasmids were constructed through Gibson Assembly mix of 10 U/μL T5 Exonuclease, 2 U/μL Phusion High-Fidelity DNA Polymerase and 40 U/μL Taq DNA Ligase, all are from New England Biolabs. Antibiotic concentrations for kanamycin and ampicillin were 50 and 100 mg/L, respectively. All strains and plasmids used in this work are listed in Supplementary Table S3.

Cell growth conditions and strain construction

In this work, cells were grown in 1mL LB medium under 37ºC at a 220-rpm shaker in 48 well deep-well plate. All twenty-one experiment groups (Table 2) were grown with kanamycin and ampicillin, and an empty control group (MCm-PdCas9-J23111) were grown with kanamycin. The host *E. coli* strain MCm was constructed from previous work by integrating a chloramphenicol expression cassette cloned from pKM154 (Addgene plasmid #13036) into the *smf* locus of wild-type *E. coli* K12 MG1655 (4). The start stain MCm-PdCas9-J23111 was construcedin previous work (4). A second plasmid was transformed into the host using the heat shock method.

Plasmid construction

The CRISPR-dCas9 system is constructed in two plasmids containing dCas9 and sgRNA, potential off-target site and reporting protein separately. The dCas9 expressing plasmid pdCas9-J23111 was constructed in previous work (4). The plasmid for sgRNA and reporting fluorescence expression is based on pN20test-114mCherry-r0-m1 from previous work (7). In this plasmid, a pJ23119 promoter controls the expression of sgRNA with Cas binding site. The sgRNA targeting sites were inserted into the upstream of -35 region of a pJ23114 controlling the expression of mCherry protein. As a result, the higher the binding activity of sgRNA and target, the lower fluorescence would be observed. Finally, twenty-one different pN20test-114mCherry plasmids were constructed (Supplementary Figure S3). Plasmids fractions were acquired using PCR (primers are listed in Supplementary Table S1, template was the original pN20test-114mCherry-r0-m1 plasmid) and combined using Gibson assembly. All the constructed plasmids were confirmed by Sanger sequencing.

Mismatch off-target effects characterization

To demonstrate the off-target repression ability of sgRNAs designed by other tools, we constructed three paired sgRNA-N20 groups for each sgRNA: (i) The experiment group, sgRNA from other tools and potential off-target N20 sequence; (ii) The positive control

group, potential off-target N20 sequence with a reversed and complemented sgRNA and (iii)The negative control group, potential off-target N20 sequence with a negative control sgRNA which has no binding activity with the N20 sequence. To further demonstrate the CRISPRi systems used in this study has no influence on mCherry expression, we designed nine control groups of all nine sgRNAs. In these control groups, N20 sequence has no binding activity with sgRNA. All test groups are listed in Table 2. To fully express the mCherry and dCas protein, a relatively long culture procedure was conducted. Strains grew overnight in LB at 37°C, 220-rpm shaker. A second round of dilution and growth was carried out to improve expression stability of the fluorescent proteins and reduce cell agglomeration (25). Each step the culture was diluted with a volume ratio of 5‰ and grew at the same condition for 10 hours. Cultures were next diluted in fresh LB medium to OD_{600}=0.02 and then grew 4 hours to logarithmic phase, also for the mature of mCherry protein. After cultivation, 5 μL of culture medium were diluted in 200 μL phosphate-buffered saline (PBS). Three biological replicates were prepared for flow cytometry analysis.

Sorting was performed on a LSRFortessa flow cytometer (BD Biosciences) in a 48 well plate. Gating based on FSC-Area and SSC-Area was carried out to exclude non-cell particles. The population in this gated area is referred to as P1. The autofluorescence is measured using the blank untransfected MCm-PdCas9-J23111 strain and was subtracted in the following analysis. Note that the blank strain is completely negative for mCherry expression, mCherry fluorescence for each group are calculated with resulting positive area in the region corresponding to the PE-Texas Red-Area.

Obtaining expression strength from cytometry data

For each cytometry assay, the log10-transformed fluorescence intensity distribution was matched with a two-component Gaussian mixture model with parameters $(\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2)$ through the expectation–maximization (EM) algorithm, where λ and $1-\lambda$ refer to the mixing coefficients of two Gaussian components; μ_1 and σ_1 are the mean and standard deviation of the first Gaussian component, respectively; μ_2 and σ_2 are the mean and standard deviation of the 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 $V_2 = ((1-\lambda_i)\sigma_{2i} \log(10))^2$.

Average repression rate (rep) of each group (R1-R4) is calculated using Equation 3:

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

Where i = 1-4.

RESULTS

Implementation of CRISPRi-FGP web tool

We have developed CRISPRi-FGP, an easy-to-use genome-scale sgRNA library designing web tool for CRISPRi system in prokaryotes. It is available at http://www.thu-big.net/sgRNA_design/. To the best of our knowledge, CRISPRi-FGP is the first tool developed for designing genome-scale sgRNA library of CRISPRi in prokaryotes. It employs a powerful system for finding off-target sites, and displays the output in an interactive page.

Input and configuration

CRISPRi-FGP has a built-in database containing reference files for over 1,300 microorganisms for users to choose from. It also accepts uploaded reference files in case one organism is not included in database or some modification is needed. To start the design, users are required to initialize the program by setting five parameters, we describe each parameter as below: (i) design target, this parameter controls the design target of sgRNAs, coding sequence (CDS) or RNA coding genes (RNA); (ii) off-target threshold, the off-target penalty threshold used in off-target selection (see materials and methods); (iii) GC content max and min, the maximum and minimum GC content of sgRNA (in percentage), which is reported to have influence on binding efficiency (26,27); (iv) spacer length, the length of designed sgRNA which has influence on efficacy and off-target effects (28). It has been reported that off-target effects induced by the 17nt but not the 20nt sgRNAs (29); (v) strand, either Template or Nontemplate, this parameter tells CRISPRi-FGP which strand on target gene should be used as the binding target of sgRNAs. Literature indicated that sgRNAs binding to the nontemplate DNA strand demonstrated effective gene silencing when targeting to coding sequence (23). sgRNA binding to template strain are for other usage like gene editing or dynamic imaging (30) where binding strain have no influence.

Process and output

After configuration, a request is sent to the server to start the pipeline. The calculation process runs totally online, at this step users can leave the page and wait for the final result to be sent to the email determined. A progress bar will be displayed on the page to show the status of calculation. For a typical genome-scale library design, we expected the total time to be 5 to 20 minutes. Larger genomes would require longer process time for its mass amount of calculation. The results are displayed in two tables and an interactive page. A table contains the designed sgRNA library. For each gene the table list its position, attribution and the number, position and sequence of all sgRNAs targeting to it. The other table contains the sgRNA candidates not chosen for library for off-target. This table shows the off-target number of each sgRNA with the sequence and penalty score of sgRNA to off-target site.

The sgRNA library map in genome can be viewed in a separated interactable page. Each gene and sgRNA are mapped in their natural position in genome. This graphical output is generated using D3GB (24) and enables the user to zoom and scroll across the gene. This view suppresses the visibility confirmation of sgRNA coverage. A downloadable fasta file of genome or targeted region is also available for further analysis. An example of interactable page and result files is shown in Figure 1.

Feature comparation with other tools

Two main factors that set CRISPRi-FGP apart from other sgRNA design software that are widely utilized are its capability to generate sgRNA on genome-scale for prokaryotes and the updated off-target selection principle specially modified for prokaryotes.

Although there are tools available that provide similar sgRNA design functions for CRISPR/Cas systems (12,31,32), there are no tool for CRISPRi systems. Our previous work showed that two types of CRISPR systems would require different sgRNA designing principles (4). Additionally, most tools can only design sgRNA library for a single gene and only limited to their built-in database, which is a drawback in function gene study in prokaryotes. Most tools, such as CHOPCHOP or CRISPOR (12,32), depended on the built-in genome database of organisms to design sgRNAs. For this reason, the supporting organisms are limited and the genome sequence are unable to be modified by users. CRISPRi-FGP is specially modified for prokaryotes, with a built-in library of over 1,300 microorganisms and the ability of processing reference files from users, CRISPRi-FGP is able to design sgRNA library for any microorganism. Accepting customized reference files also allows users to modify the files themselves by adding or deleting sequences from the genome file.

CRISPRi-FGP also has revised model for sgRNA off-target scoring comparing with existing tools. It has two sgRNA scoring model, the seed region model and the penalty score model. The seed region model is from experiments where we found that in seed region, double mutants exhibited apparently lower binding affinities than the sum of effects of two relevant single mutants (7). A second off-target scoring method is a penalty score method and based on experimental parameters (4). These two models have strong experimental prove of accuracy.

CRISPRi-FGP is superior to existing tools in avoiding off-target

Though multiple tools for sgRNA design exists, only two existing design tools: CHOPCHOP and Synthego, both are widely used, have the both the ability of designing sgRNAs in prokaryotes and an off-target ranking system. Thus, results from the two tools were compared with CRISPRi-FGP for off-target activity. We discovered that some sgRNAs are high in rank in those two tools but are identified as off-target in CRISPRi-FGP (Table 1). We further constructed a reporter system based on mCherry to test the binding activity of those sgRNAs with the N20sequences identified by CRISPRi-FGP. The results of mismatch off-target effects characterization suggest that all tested sgRNAs from other tools showed off-target activity with sequences predicted by CRISPRi-FGP (Figure 3B). The flow cytometer results showed fold change in the fluorescence of different experiment group (Figure 3D-3G). The average repression rate was 69.47%, ranging from 51.12% in R1 and 90.70% in R4. Fluorescence control groups of all nine sgRNAs showed close mCherry fluorescence (Figure 3C). Careful analysis of the flow cytometer results reviewed no static difference between each group, demonstrated sgRNA along has no influence to mCherry fluorescence and the repression rate was contributed by off-target effects. These results indicated that CRISPRi-FGP is able to recognize off-target positions unable to be found by existing tools.

DISCUSSION

In this study we developed CRISPRi-FGP, an easy-to-use web tool for designing genome-scale sgRNA library for CRISPRi in prokaryotes. The tool is deployed on a web server thus requires no local hardware or software. It maximizes user flexibility while maintaining a simple interface. CRISPRi-FGP can run with default organism database, and also accepts customized reference files. CRISPRi-FGP is superior to existing tools in predicting off-target sites, it also provides the output in both zoomable interactive page and easy-to-read tables. It proves to be an excellent tool in all kinds of CRISPR systems and especially in some applications which require stricter off-target selection like genotype-phenotype association based on CRISPRi-seq.

CRISPRi-FGP uses standard annotation files for sgRNA library design, which can be acquired from Prokaryotic Genome Annotation Pipeline (PGAP) (33). Moreover, the seed region principle is proved by *in vitro* experiments, so it also works in eukaryotic organisms. This universality of CRISPRi-FGP makes it possible for designing sgRNA library for any kind of organism and thus opens up the prospect of genotype-phenotype association. We have applied CRISPRi-FGP in the genotype-phenotype association of an industry *Corynebacterium glutamicum* strain, which lacks information of its genes. The genome sequence was annotated using PGAP. In total, 3,115 coding genes and 82 non-coding genes were labeled for genome-scale sgRNA library design. The final sgRNA library showed adequate coverage of all genes.

However, some drawbacks exist in the sgRNA design principle. While the seed region principle was proved by *in vitro* experiments and been widely accepted, the off-target penalty region and scoring principles are completely based on experience, lacking both agreement or rational explanation. Furthermore, the contribution of different kinds of Watson-Crick and mismatched nucleobases pairs are considered to be the same. However, literatures showed that the stability of different nucleobases pairs (34,35), there are also some mismatches have close stability as Watson-Crick pairs like rU/dG, rG/dT and rG/dG (36). This different in stability has severe impact on binding activity but is neglected in existing principles. Nucleic acids thermodynamics provides a potential lead to rational calculation of RNA-DNA combination (37–40). Literature has reported a calculation model predicting off-target activity based on Markov chain using parameters from deep learning (7), in such progress each step is controlled by thermodynamics of single base pair breaking in DNA-DNA and formation in RNA-DNA. This provides us a physical model of sgRNA binding. There are already models for predicting the thermodynamic parameters of RNA-DNA combination (31), and some parameters of different Watson-Crick and mismatched pairs had already been fitted from experiments (34,35). However, the parameters of mismatches are not sufficient, leading to the inability of predicting the properties of hybrid duplexes containing some mismatches. A comprehensive measurement of nearest-neighbor parameters is urgently required therefore a rational model of sgRNA off-target activity can be further established.

ACKNOWLEDGEMENT

We would like to thank members from the C. Zhang and H. Xing lab oratories for critical discussions of this work.

FUNDING

National Key R&D Program of China [2018YFA0901500 to X.X.]; National Natural Science Foundation of China [21938004 to X.X.]; National Natural Science Foundation of China [U2032210 to C.Z.].

conflict of interest statement. None declared.

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 Oct 20;6(10):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 Jan 12;35(1):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 Nov 20;4(11):1244–53.
4. 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-net. Nat Commun [Internet]. 2018;9(1). Available from: http://dx.doi.org/10.1038/s41467-018-04899-x
5. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science (1979). 2012 Aug 17;337(6096):816–21.
6. 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 Sep 21;31(9):827–32.
7. Feng H, Guo J, Wang T, Zhang C, Xing X hui. Guide-target mismatch effects on dCas9–sgRNA binding activity in living bacterial cells. Nucleic Acids Res [Internet]. 2021 Feb 22;49(3):1263–77. Available from: https://academic.oup.com/nar/article/49/3/1263/6121456
8. Shalem O, Sanjana NE, Zhang F. High-throughput functional genomics using CRISPR–Cas9. Nat Rev Genet. 2015 May 9;16(5):299–311.
9. Heigwer F, Kerr G, Boutros M. E-CRISP: fast CRISPR target site identification. Nat Methods. 2014 Feb;11(2):122–3.
10. Sander JD, Joung JK. CRISPR-Cas systems for editing, regulating and targeting genomes. Nat Biotechnol. 2014;32(4):347–55.
11. Montague TG, Cruz JM, Gagnon JA, Church GM, Valen E. CHOPCHOP: a CRISPR/Cas9 and TALEN web tool for genome editing. Nucleic Acids Res [Internet].

2014 Jul 1;42(W1):W401–7. Available from: http://academic.oup.com/nar/article/42/W1/W401/2437392/CHOPCHOP-a-CRISPRCas9-and-TALEN-web-tool-for

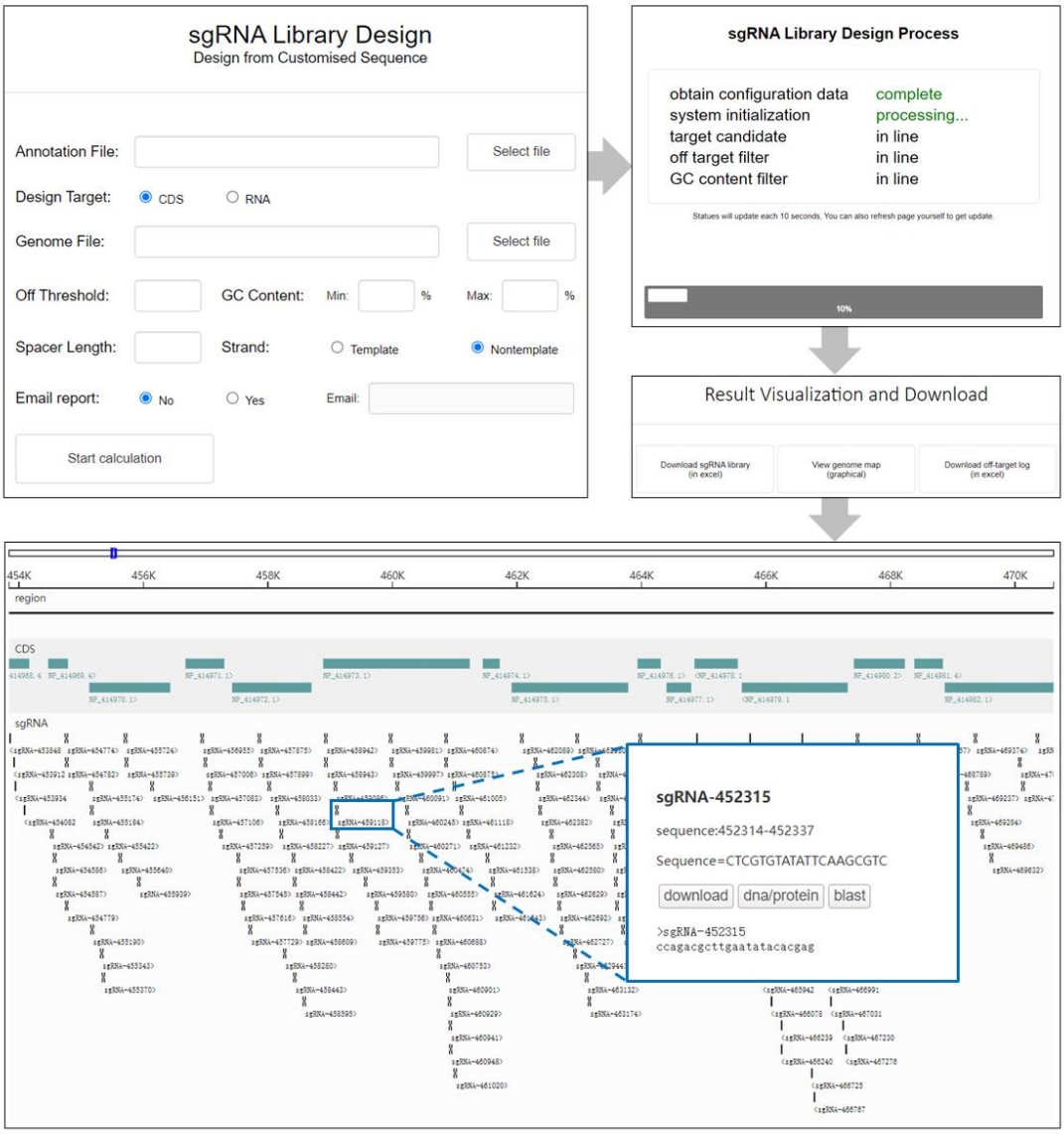
12. Concordet JP, Haeussler M. CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. Nucleic Acids Res [Internet]. 2018 Jul 2 [cited 2022 Oct 16];46(W1):W242–5. Available from: https://academic.oup.com/nar/article/46/W1/W242/4995687
13. Struhl K. Fundamentally Different Logic of Gene Regulation in Eukaryotes and Prokaryotes. Cell. 1999 Jul;98(1):1–4.
14. Kuzminov A. The Precarious Prokaryotic Chromosome. J Bacteriol. 2014 May 15;196(10):1793–806.
15. Sternberg SH, LaFrance B, Kaplan M, Doudna JA. Conformational control of DNA target cleavage by CRISPR–Cas9. Nature. 2015 Nov 28;527(7576):110–3.
16. 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. Proceedings of the National Academy of Sciences. 2017 May 23;114(21):5461–6.
17. 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 [Internet]. 2018 Aug 21;46(14):7052–69. Available from: doi.org/10.1093/nar/gky572
18. Jiang H, Wong WH. SeqMap: mapping massive amount of oligonucleotides to the genome. Bioinformatics. 2008 Oct 15;24(20):2395–6.
19. 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 [Internet]. 2014 Oct;159(3):647–61. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0092867414011787
20. 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 [Internet]. 1993 [cited 2022 Oct 15];175(8):2445–7. Available from: https://journals.asm.org/doi/10.1128/jb.175.8.2445-2447.1993
21. Schrider DR, Hahn MW. Gene copy-number polymorphism in nature. Proceedings of the Royal Society B: Biological Sciences [Internet]. 2010 Nov 7 [cited 2022 Oct 15];277(1698):3213–21. Available from: https://royalsocietypublishing.org/doi/10.1098/rspb.2010.1180
22. Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL. NCBI BLAST: a better web interface. Nucleic Acids Res [Internet]. 2008 Jul 1 [cited 2022 Oct 15];36(suppl_2):W5–9. Available from: https://academic.oup.com/nar/article/36/suppl_2/W5/2505810
23. 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 [Internet]. 2013;152(5):1173–83. Available from: http://dx.doi.org/10.1016/j.cell.2013.02.022
24. Barrios D, Prieto C. D3GB: An Interactive Genome Browser for R, Python, and

WordPress. J Comput Biol [Internet]. 2017 May 1 [cited 2022 Oct 24];24(5):447–9. Available from: https://pubmed.ncbi.nlm.nih.gov/28437135/

25. Lin MY, Lindsay HM, Weitz DA, Ball RC, Klein R, Meakin P. Universality in colloid aggregation. Nature 1989 339:6223 [Internet]. 1989 [cited 2022 Oct 24];339(6223):360–2. Available from: https://www.nature.com/articles/339360a0
26. Gagnon JA, Valen E, Thyme SB, Huang P, Ahkmetova L, Pauli A, et al. Efficient mutagenesis by Cas9 protein-mediated oligonucleotide insertion and large-scale assessment of single-guide RNAs. PLoS One [Internet]. 2014 May 29 [cited 2022 Oct 25];9(5). Available from: https://pubmed.ncbi.nlm.nih.gov/24873830/
27. Wang T, Wei JJ, Sabatini DM, Lander ES. Genetic screens in human cells using the CRISPR-Cas9 system. Science (1979) [Internet]. 2014 Jan 3 [cited 2022 Oct 25];343(6166):80–4. Available from: https://www.science.org/doi/10.1126/science.1246981
28. Matson AW, Hosny N, Swanson ZA, Hering BJ, Burlak C. Optimizing sgRNA length to improve target specificity and efficiency for the GGTA1 gene using the CRISPR/Cas9 gene editing system. PLoS One [Internet]. 2019 Dec 1 [cited 2022 Oct 25];14(12):e0226107. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0226107
29. Zhang JP, Li XL, Neises A, Chen W, Hu LP, Ji GZ, et al. Different Effects of sgRNA Length on CRISPR-mediated Gene Knockout Efficiency. Scientific Reports 2016 6:1 [Internet]. 2016 Jun 24 [cited 2022 Oct 25];6(1):1–10. Available from: https://www.nature.com/articles/srep28566
30. 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 Dec 19;155(7):1479–91.
31. Müller Paul H, Istanto DD, Heldenbrand J, Hudson ME. CROPSR: an automated platform for complex genome-wide CRISPR gRNA design and validation. BMC Bioinformatics. 2022 Dec 1;23(1).
32. Montague TG, Cruz JM, Gagnon JA, Church GM, Valen E. CHOPCHOP: a CRISPR/Cas9 and TALEN web tool for genome editing. Nucleic Acids Res [Internet]. 2014 Jul 1;42(Web Server issue):W401-7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24861617
33. 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 [Internet]. 2021 Jan 8 [cited 2022 Nov 22];49(D1):D1020–8. Available from: https://pubmed.ncbi.nlm.nih.gov/33270901/
34. Sugimoto N, Nakano S ichi, Katoh M, Matsumura A, Nakamuta H, Ohmichi T, et al. Thermodynamic Parameters To Predict Stability of RNA/DNA Hybrid Duplexes. Biochemistry [Internet]. 1995 Sep 5;34(35):11211–6. Available from: https://pubs.acs.org/doi/abs/10.1021/bi00035a029
35. Watkins NE, Kennelly WJ, Tsay MJ, Tuin A, Swenson L, Lee HR, et al. Thermodynamic contributions of single internal rA·dA, rC·dC, rG·dG and rU·dT mismatches in RNA/DNA duplexes. Nucleic Acids Res [Internet]. 2011 Mar;39(5):1894–902. Available from: https://academic.oup.com/nar/article-

lookup/doi/10.1093/nar/gkq905
36. Sugimoto N, Nakano M, Nakano S ichi. Thermodynamics−Structure Relationship of Single Mismatches in RNA/DNA Duplexes. Biochemistry [Internet]. 2000 Sep 1 [cited 2022 Sep 27];39(37):11270–81. Available from: https://pubs.acs.org/doi/10.1021/bi000819p
37. Borer PN, Dengler B, Tinoco I, Uhlenbeck OC. Stability of ribonucleic acid double-stranded helices. J Mol Biol [Internet]. 1974 Jul;86(4):843–53. Available from: https://linkinghub.elsevier.com/retrieve/pii/002228367490357X
38. Freier SM, Kierzek R, Jaeger JA, Sugimoto N, Caruthers MH, Neilson T, et al. Improved free-energy parameters for predictions of RNA duplex stability. Proceedings of the National Academy of Sciences [Internet]. 1986 Dec;83(24):9373–7. Available from: https://pnas.org/doi/full/10.1073/pnas.83.24.9373
39. Breslauer KJ, Frank R, Blocker H, Marky LA. Predicting DNA duplex stability from the base sequence. Proc Natl Acad Sci U S A. 1986;83(11):3746–50.
40. Turner DH, Sugimoto N. RNA structure prediction. Annu Rev Biophys Biophys Chem. 1988;17:167–92.

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	Off-target sequence	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 CRISPRi-FGP query. The homepage guides users to input or database design page, which allows users to upload reference files and set parameters. The main design process starts when a 'start calculation' order is received, the website is automatically redirected to the progress page which displays real-time progress of calculation. Finally, after calculation, CRISPRi-FGP redirects again to the result page. Users are free to download the result tables and view the sgRNA library in the interactive page. **(A)** The workflow and interactable page that provides a zoomable genome map of genes and sgRNA library, each gene and sgRNA is located in its natural

position. **(B)** The final result table containing the sgRNA library, sgRNAs are classified by its binding gene, sequence, position and guide number are provided. **(C)** The off-target log table, lists all not chosen sgRNAs with their predicted off-target sequence and penalty score.

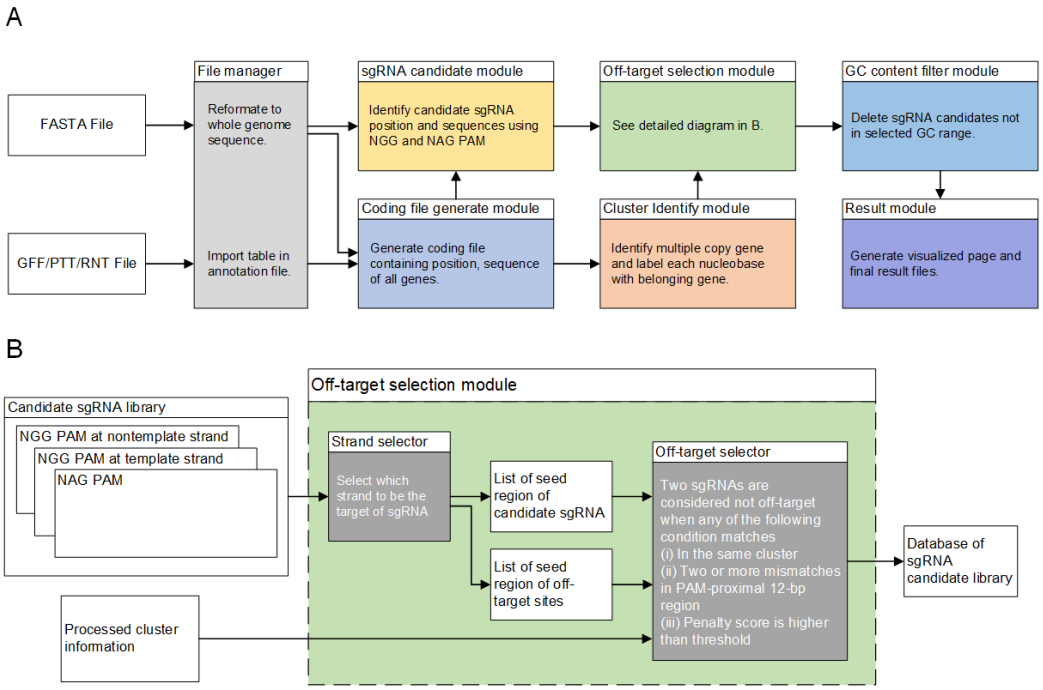


Figure 2. Functional block diagram of CRISPRi-FGP modules. **(A)** The workflow based on functional modules block CRISPRi-FGP. Two input data files (FASTA, GFF/PTT/RNT) are imported by system or user and processed by multiple modular programs. A coding file is generated based on the two files. The following modules did the sgRNA searching, off-target filter (shown in detail in B), GC content selection and result collecting. **(B)** The off-target selection module takes data from the sgRNA candidate module (which generates three lists of sgRNA candidates of NGG PAM in nontemplate strand, NGG PAM in template strand and NAG PAM), and the Cluster identify module. Each sgRNA candidate is compared with other sgRNAs to rule out off-target. Final sgRNA library is then added to the database to be associated with GC content selection and final results collating.

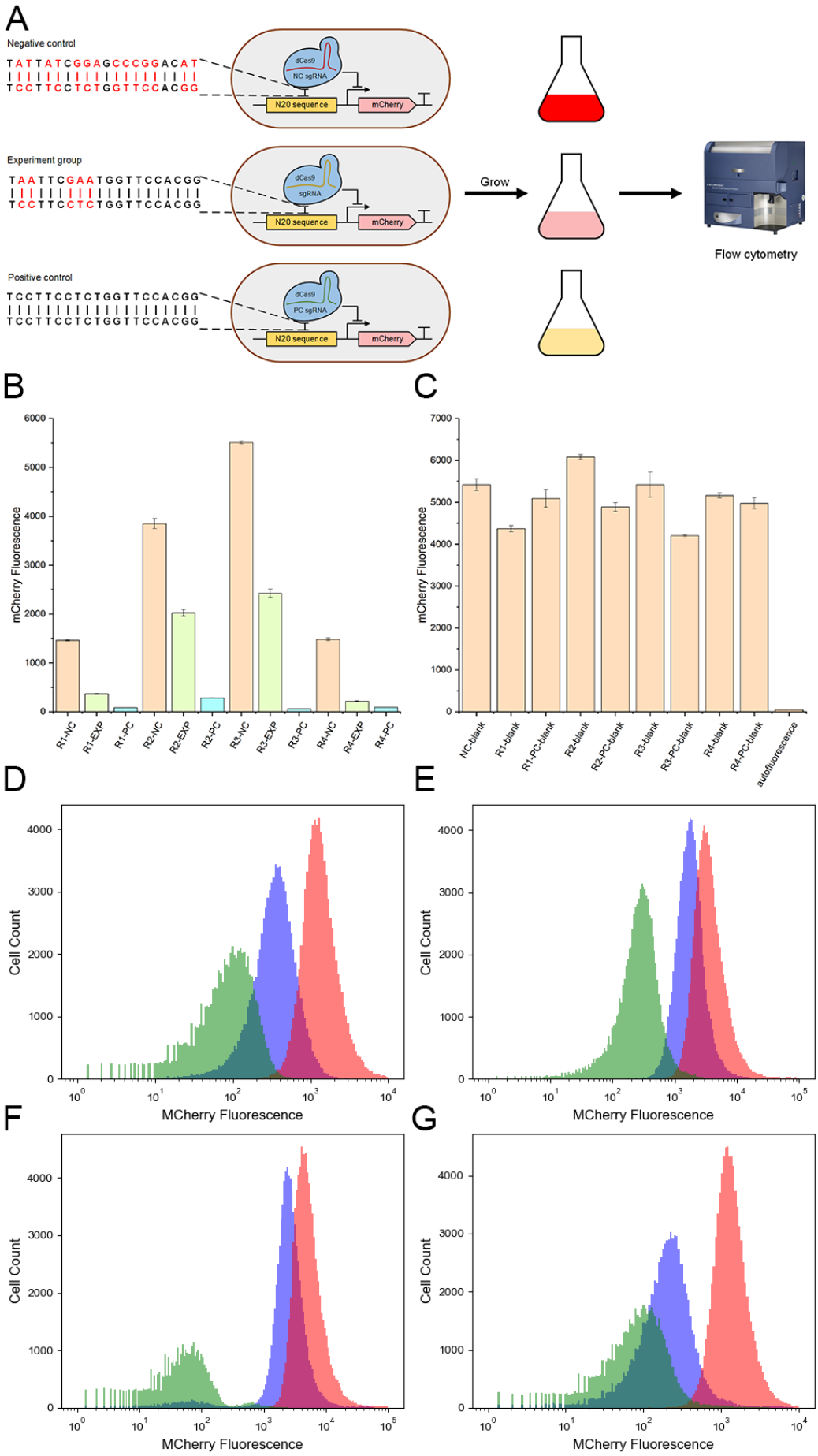

A
Negative control
TATTATCGGAGCCCGGACAT
TCCTTCCTCTGGTTCCACGG
dCas9
NC sgRNA
N20 sequence
mCherry
Experiment group
TAATTCGAATGGTTCCACGG
TCCTTCCTCTGGTTCCACGG
dCas9
sgRNA
N20 sequence
mCherry
Positive control
TCCTTCCTCTGGTTCCACGG
TCCTTCCTCTGGTTCCACGG
dCas9
PC sgRNA
N20 sequence
mCherry
Grow
Flow cytometry
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
autofluorescence
D
Cell Count
MCherry Fluorescence
E
Cell Count
MCherry Fluorescence
F
Cell Count
MCherry Fluorescence
G
Cell Count
MCherry Fluorescence

Figure 3. Flow cytometer results of off-target effects of mismatched sgRNAs. **(A)** Schematic diagram of profiling of dCas9's binding activities mediated by mismatched sgRNAs. The binding activity of each individual mismatched sgRNA are compare with a positive control sgRNA (complete reverse and complement with N20) and a negative control sgRNA (no binding activity with N20). Control groups of the three sgRNAs were used to rule out the influence of different sgRNAs on the fluorescence. **(B)** Mean fluorescence value of test groups (gray), negative control groups (black) and positive control groups (white). The fluorescence can stand for the binding activity of sgRNA-N20 sequences (n = 3). **(C)** Mean fluorescence value of all sgRNAs with negative control N20 sequence and strain without mCherry expression, fluorescence of each group is within 30% difference, showing acceptable parallelism. **(D-G)** Comparation of histogram of corresponding test groups from flow cytometer of R1 (D), R2 (E), R3 (F) and R4 (G), three colors are test groups (blue), negative control groups (red) and positive control groups (green). The binding activity of mismatched sgRNA has fold change comparing to negative control groups.

Table 1. sgRNA designed by other tool and off-target sites predicted by CRISPRi-FGP

No	Design tool	Rank*	Target gene	sgRNA sequence	Off-target gene	Off-target sequence from CRISPRi-FGP
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

*Rank of sgRNA in other tools, ranking method differs when tool change

Table 2. sgRNA 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	
R4-EXP	GAGCTCAGATCGCTCAACAT	TCTCTTGAATCGCTCAACAT
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	