CRISPR-offinder

User Manual

Version 1.2

Written by Xie Shengsong
August 2017

Huazhong Agricultural University

Thank you for using CRISPR-offinder. Please read this manual carefully before using this software and save this manual for future use.

Copyright Statement

Contents

I. Introduction

CRISPR/Cas system undoubtedly holds great potential for genome editing. Target site cleavage by CRISPR technology requires a protospacer adjacent motif (PAM) immediately downstream or upstream of the protospacer element to which the sgRNA binds. However, Cas9 from different types of bacteria or variant recognizes different PAM sequences. To meet the needs of different CRISPR system with specific and efficient sgRNA design, CRISPR-offinder was developed.

Given an input FASTA file of the target sites and queries the reference genome as well as a CRISPR system with a defined spacer length and PAM sequence, this standalone tool will identify putative sites and assign a predicted activity based on support vector machine model which conducted by sgRNA Scorer 2.0 [1]. In addition, sgRNAs with minimal off-target activity were predicted by Cas-OFFinder [2], and score with Off-Target Cutting Frequency Determination (CFD) [3].

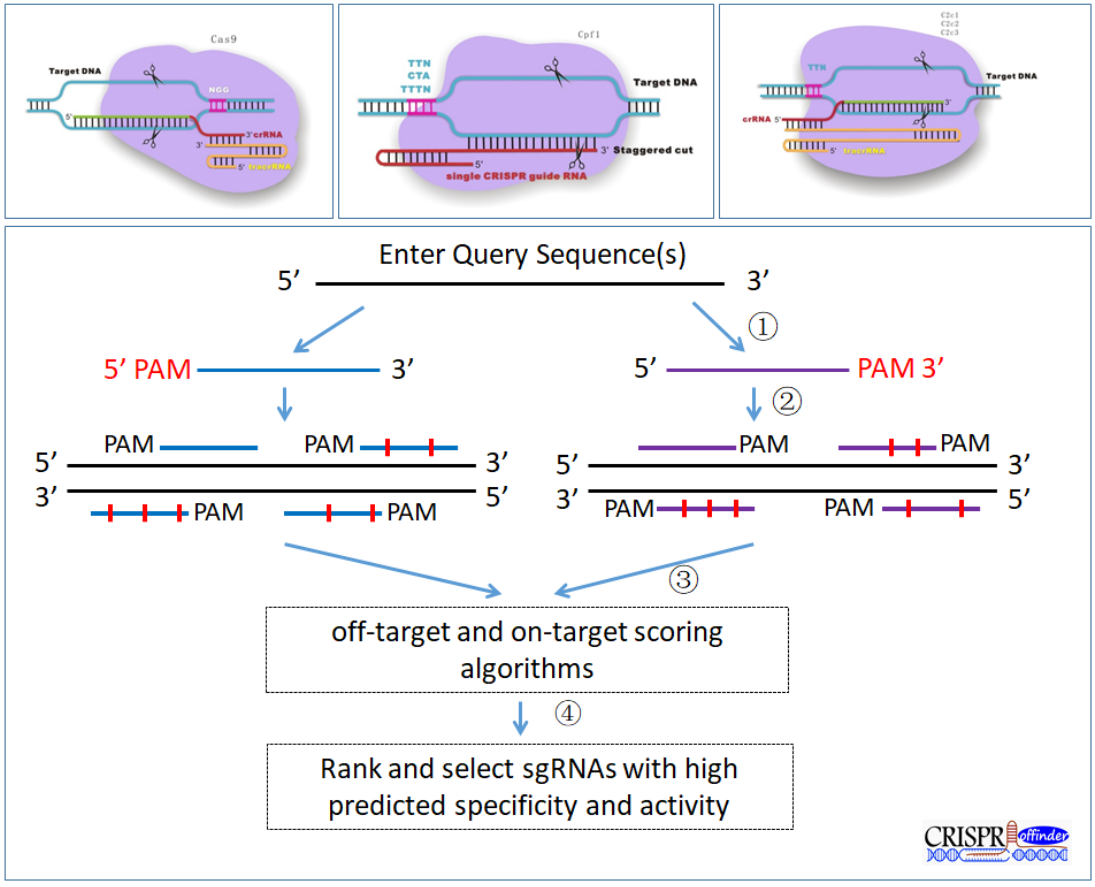

Figure. Workflow of CRISPR-offinder

A. Supported Operating Systems

1. Linux (with proprietary drivers installed)
2. Mac OS X (Snow leopard or higher)

B. Requirements and Installation

Perl;
Python >=2.7 or greater;
Python package: Scipy, Biopython and Sci-kit learn, you can install by commond:

```
sudo pip install BioPython
sudo pip install sklearn
sudo pip install scipy
```

CRISPR-offinder requires an OpenCL-enabled device (CPU, GPU, or etc..) and corresponding runtime driver pre-installed to run properly.

Please download and install the latest driver below:
1.Intel (Download 'OpenCL runtime' in the middle of the page):
http://software.intel.com/en-us/vcsource/tools/opencl-sdk
2.NVidia: http://www.nvidia.com/Download/index.aspx
3.AMD: http://support.amd.com/en-us/download

Before installing CRISPR-offinder, please check whether your device is an OpenCL-supported one and make sure that Cas-OFFinder program work well.

C. Synopsis

$ perl ./ CRISPR-offinder_1.2 .pl <option>
For help information, type 'perl CRISPR-offinder_1.2.pl'

D. Required parameters

-input: [s] Input file in FASTA format <required>
-pamseq: [s] PAM sequence <required>
-pamori: [s] PAM orientation. Value must either be 5'or 3' (enter 5 or 3) <deafult: 3>
-pamlen: [i] Length of protospacer <default: 20>
-gc_min: [i] The minimum value of GC content <default: 20>
-gc_max: [i] The maximum value of GC content <default: 80>
-mismatches: [i] Number of mismatches[0-9] <default: 5>
-strand: [s] Searching CRISPR target sites using DNA strands based option(s/a/b) <default: b>

-cga: [s] (C: using CPUs, G: using GPUs, A: using accelerators) <default: C>
-gd: [s] genome dir <default: $Bin/genome>
-system: [s] run system (Linux32/Linux64/Mac) <default: Linux64>
-offset_start: [i] The minimum value of sgRNA offset <default: -2>
-offset_end: [i] The maximum value of sgRNA offset <default: 32>
-output: [s] Output dir. Final output files with guide RNA sequences and scores <default: ./>

Notes: 1. [s] stand for string; [i] stand for integer.
2. The input sequences entered should not contain any letters other than A, T, C or G.
3. Whole genome of target organism is needed (in FASTA format). You can find one in one of the below links:
1.UCSC genome sequences library, http://hgdownload.soe.ucsc.edu/downloads.html
2.Ensembl sequence library, http://www.ensembl.org/info/data/ftp/index.html

Extract all FASTA files in a directory. Remember the full path of the FASTA files directory (for option -gd).

E. PAM sequence

RNA-guided Endonucleases(Species/variant)	**PAM sequence(5'→3')**	**Direction(3'/5' side)**
SpCas9 from Streptococcus pyogenes	NGG	3’
SpCas9 from Streptococcus pyogenes	NRG	3’
StCas9 from Streptococcus thermophilus	NNAGAAW	3’
NmCas9 from Neisseria meningitidis	NNNNGATT	3’
SaCas9 from Staphylococcus aureus	NNGRRT	3’
SaCas9 variant (KKH SaCas9)	NNNRRT	3’
SpCas9 D1135E variant	NGG (reduced NAG binding)	3’
SpCas9 VRER variant	NGCG	3’
SpCas9 EQR variant	NGAG	3’
SpCas9 VQR variant	NGAN or NGNG	3’
AsCpf1 from Acidaminococcus, LbCpf1 from Lachnospiraceae	TTTN	5’
FnCpf1 from Francisella novicida strain U112	TTN and/or CTA	5’
C2c1 from four major taxa: Bacilli, Verrucomicrobia, a-proteobacteria, and d-proteobacteria	TTN	5’

Code	Base	Code	Base
A	Adenine	K	G or T
C	Cytosine	M	A or C
G	Guanine	B	C or G or T
T	Thymine	D	A or G or T
R	A or G	H	A or C or T
Y	C or T	V	A or C or G
S	G or C	N	any base
W	A or T		

Notes that CRISPR-offinder allows mixed bases to account for the degeneracy in PAM sequences, Custom, Enter your PAM - enter user defined-PAM

F. PAM orientation and location

The orientation of the CRISPR PAM can be set on the 5’ or 3’.
sense strand - PAM located on the sense strand
anti-sense strand - PAM located on the anti-sense strand
both strands - PAM located on the both DNA strands
Note that if user wants to design “Paired-gRNAs”, user can set the value of (minimum) min and (maximum) max of “gRNA offset” for designing Paired-gRNAs.

G. Mismatches number

The maximum number of mismatches that allowed in the "sgRNA" region when perform whole genome alignment, 'N' in PAM sequence are not counted as mismatched bases.

H. On-target score

The on-target score of sgRNA sequence based on sgRNA Scorer 2.0 [1]. The higher the score, the better the predicted activity. Guide scores should typically be used in a relative manner as opposed to absolute. For example, if one guide is scored 3.9 and another guide is scored 1.9, the first guide would be considered better than the second guide. The score is purely based on on-target activity and does not incorporate off-target activity.

I. Off-target score

The off-target score module was designed based on the improved model of off-target according to their interference potential. We employed cutting frequency determination (CFD) score models [3] to predict the

off-target activities of sgRNAs for CRISPR/Cas9 system in CRISPR-offinder.

For a given sgRNA, CRISPR-offinder enumerates all its neighbors up to Q mismatches, calculates the CFD total score for each neighbor, and then multiplies that score by the number of times the neighbor occurs in the genome. It then aggregates the CFD values into a single composite score using the formula used by Perez et al [4]:

$$\text{Specificity Score} = \frac{1}{\sum_{i=1}^{n} CFD_i * q_i}$$

In the off-target score model, sgRNA mismatche's possition, number and PAM's mismatches are taken into account. The total CFD score range from 0 to 1, sgRNA with total lower off-target score has much off-target potential, which ususally should be avoided.

In general, sgRNA with high off-target score (CFD total) sites and with high on-target score are the ideal ones user needs.

G. Description of result file

"CRISPR_offinder_report.xls", single sgRNAs

sgRID	Start	End	CRISPR_target_sequence(5'-3')	Length(nt)	GC%	0M	1M	2M	3M	4M	5M	6M	7M	8M	9M	Total	Scorer2.0	CFD
gI_PRV_S_62	1116	1138	GAAGCTGGCCACCATCGCAGAGG	23	65.22%	1	0	0	0	0	0	0	0	0	0	0	3.9182	1
gI_PRV_S_47	801	823	CCTGGACGCGAACGGCACGATGG	23	69.57%	1	0	0	0	0	0	0	0	0	0	0	2.8599	1
gI_PRV_S_21	355	377	GTGCACCACGAGGCCTTCCGCGG	23	73.91%	1	0	0	0	0	0	0	0	0	0	0	2.6154	1
gI_PRV_S_44	768	790	GGCGACGACGGCGATGACCCCGG	23	78.26%	1	0	0	1	0	0	0	0	0	0	1	0.7769	0.9623
gI_PRV_S_5	90	112	CCTCCTCGCCGCCCTGACCCTGG	23	78.26%	1	0	0	1	0	0	0	0	0	0	1	1.3081	0.9
gI_PRV_S_38	741	763	CGACGAAGAGGAGGAGGACGAGG	23	65.22%	1	0	2	17	0	0	0	0	0	0	19	1.9114	0.2346
gI_PRV_S_40	745	767	GAAGAGGAGGAGGACGAGGAGGG	23	65.22%	2	0	9	9	0	0	0	0	0	0	19	2.5545	0.1808
gI_PRV_S_41	746	768	AAGAGGAGGAGGACGAGGAGGGG	23	65.22%	1	1	0	14	0	0	0	0	0	0	15	3.1917	0.1798
gI_PRV_S_39	744	766	CGAAGAGGAGGAGGACGAGGAGG	23	65.22%	1	0	8	15	0	0	0	0	0	0	23	3.5817	0.1264
gI_PRV_S_42	747	769	AGAGGAGGAGGACGAGGAGGGGG	23	69.57%	1	1	4	21	0	0	0	0	0	0	26	3.1259	0.1074

First column - sgRID - unique identifier for the sgRNA sequence

#_A_1, A stand for PAM located on the anti-sense strand

#_S_1, S stand for PAM located on the sense strand

Second column - start position of the sgRNA target site in given query sequence

Third column - end position of the sgRNA target site in given query sequence

Forth column - nucleotide sequence of the sgRNA sequence (including PAM)

Fifth column - length of target sequence

Sixth column - GC contents of protospacer

Seventh column - the number of the perfect matched site, if 0M (mismatch) =1, represent unique on-target site in genome; if 0M = 0, represent no perfect matched site in genome; if 0M >1, please check the target gene whether is a multi-copied gene, it's may target to the same sequence, otherwise, it's may contain perfect matched off-target sites.

Eighth column - the number of the off-target sites with 1 mismatched bases (1M)
Ninth column - the number of the off-target sites with 2 mismatched bases (2M)
Tenth column - the number of the off-target sites with 3 mismatched bases (3M)
Eleventh column - the number of the off-target sites with 4 mismatched bases (4M)
Twelfth column - the number of the off-target sites with 5 mismatched bases (5M)
Thirteenth column - the number of the off-target sites with 6 mismatched bases (6M)
Fourteenth column - the number of the off-target sites with 7 mismatched bases (7M)
Fifteenth column - the number of the off-target sites with 8 mismatched bases (8M)
Sixteenth column - the number of the off-target sites with 9 mismatched bases (9M)
Seventeenth column - the total number of the mismatched sites ('N' in PAM sequence are not counted as mismatched bases)
Eighteenth column - Scorer2.0 - score of sgRNA sequence based on sgRNA Scorer 2.0. The higher the score, the better the predicted activity. Guide scores should typically be used in a relative manner as opposed to absolute.
Last column – CFD (total) - The score range from 0 to 1, sgRNA with total lower off-target score has much off-target potential, which ususally should be avoided.

folder of "off-target"

On-target sequence	Chromosome	Start	On- or off-target sequences	Strand	mismatches	CFD score
AGAGGAGGAGGACGAGGAGGNNN	PRV_genome	122998	AGAGGAGGAGGACGAGGAGGGGG	+	0	1
AGAGGAGGAGGACGAGGAGGNNN	PRV_genome	12110	AGAGGAGGAGGACGAGGAtGGGG	+	1	0.666666667
AGAGGAGGAGGACGAGGAGGNNN	PRV_genome	12014	cGAGGAGGAGGAgGAGGAGGAGG	+	2	0.116883117
AGAGGAGGAGGACGAGGAGGNNN	PRV_genome	12017	gGAGGAGGAGGAgGAGGAGGAGG	+	2	0.136363636
AGAGGAGGAGGACGAGGAGGNNN	PRV_genome	106356	gGAGGAcGAGGACGAGGAGGAGG	+	2	0.6875

First column - given query sequence
Second column - FASTA sequence title (if you downloaded it from UCSC or Ensembl, it is usually a chromosome name)
Third column - position of the potential off-target site
Forth column - actual sequence located at the position (mismatched bases noted in lowercase letters)
Fifth column - indicates forward strand(+) or reverse strand(-) of the found sequence
Sixth column - the number of the mismatched bases ('N' in PAM sequence are not counted as mismatched bases)

Last column - CFD score - The score range from 0 to 1, site with higher off-target score has much off-target potential, which ususally should be avoided.

“report_protospacer_pairs.xls”, paired-gRNAs

sgRID_S	target_seq_S	Start_S	End_S	GC%_S	<->	sgRID_A	target_seq_A	Start_A	End_A	GC%_A	sgRNA_offset(bp)
seq1_A_4	AAGCGGTTCCGCAGGACCCAGGG	856	878	69.57%	<->	seq1_S_54	CTTCTTCGGCCACCGCTTCCAGG	876	898	65.22%	-2
seq1_A_5	AGCGGTTCCGCAGGACCCAGGGG	855	877	73.91%	<->	seq1_S_54	CTTCTTCGGCCACCGCTTCCAGG	876	898	65.22%	-1
seq1_A_6	GCAGGACCCAGGGGTAGAAATGG	846	868	60.87%	<->	seq1_S_54	CTTCTTCGGCCACCGCTTCCAGG	876	898	65.22%	8
seq1_A_7	ACCCAGGGGTAGAAATGGAGAGG	841	863	56.52%	<->	seq1_S_54	CTTCTTCGGCCACCGCTTCCAGG	876	898	65.22%	13
seq1_A_7	ACCCAGGGGTAGAAATGGAGAGG	841	863	56.52%	<->	seq1_S_53	GTCCTGCGGAACCGCTTCTTCGG	862	884	65.22%	-1
seq1_A_8	CCCAGGGGTAGAAATGGAGAGGG	840	862	60.87%	<->	seq1_S_54	CTTCTTCGGCCACCGCTTCCAGG	876	898	65.22%	14
seq1_A_8	CCCAGGGGTAGAAATGGAGAGGG	840	862	60.87%	<->	seq1_S_53	GTCCTGCGGAACCGCTTCTTCGG	862	884	65.22%	0
seq1_A_9	GGTAGAAATGGAGAGGGTCCCGG	834	856	60.87%	<->	seq1_S_54	CTTCTTCGGCCACCGCTTCCAGG	876	898	65.22%	20
seq1_A_9	GGTAGAAATGGAGAGGGTCCCGG	834	856	60.87%	<->	seq1_S_53	GTCCTGCGGAACCGCTTCTTCGG	862	884	65.22%	6
seq1_A_10	AATGGAGAGGGTCCCGGTGCTGG	828	850	65.22%	<->	seq1_S_54	CTTCTTCGGCCACCGCTTCCAGG	876	898	65.22%	26

First column – sgRID_S - unique identifier for the sgRNA sequence
Second column - Target sequence, which located on the anti-sense strand
Third column - start position of the sgRNA target site in given query sequence
Forth column - end position of the sgRNA target site in given query sequence
Fifth column - GC contents of protospacer
Sixth column - sgRID_A
Seventh column - Target sequence, which located on the sense strand
Eighth column - start position of the sgRNA target site in given query sequence
Ninth column - end position of the sgRNA target site in given query sequence
Tenth column - GC contents of protospacer
Last column - sgRNA offset for paired-gRNA

II. Frequently Asked Questions

1. What is CRISPR-offinder?

CRISPR-offinder is a CRISPR sgRNA design and off-target searching tool for user-defined protospacer adjacent motif (PAM). This tool ranks and picks candidate sgRNA sequences for the targets provided, while attempting to maximize on-target activity and minimizing off-target activity. It uses the "sgRNA Scorer 2.0" scoring model from Raj Chari et al., ACS Synthetic Biology 2017 to assess sgRNA on-target activity, and the CFD (Cutting Frequency Determination) score to evaluate off-target sites.

2. What is a PAM sequence and where is it located?

CRISPR-Cas9/Cpf1/C2c1 mechanisms recognize DNA targets that are complementary to a short CRISPR sgRNA sequence. The part of the sgRNA sequence that is complementary to the target sequence is known as a protospacer. In order for Cas9/Cpf1/C2c1 to function it also requires a specific protospacer adjacent motif (PAM) that varies depending on the bacterial species of the Cas9/Cpf1/C2c1 gene.
Recognition of the PAM by the Cas9/Cpf1/C2c1 nuclease is thought to destabilize the adjacent sequence, allowing interrogation of the sequence by the sgRNA, and resulting in RNA-DNA pairing when a matching sequence is present. Cas9 nucleases with alternative PAMs have also been characterized and successfully used for genome editing. It is important to note that the PAM is not present in the sgRNA sequence but needs to be immediately downstream or upstream of the target site in the genomic DNA.

3. What is the CRISPR/Cas9 system?

Cas9/sgRNA system, one type of the CRISPR/Cas systems, is a kind of engineered endonuclease (Fig1). It consists of two components: Cas9, a protein with DNA nuclease activity can be used universally in this system; and sgRNA, an ~100-nt single guide-RNA, of which the first ~20 nt in the 5'-end is responsible for recognizing the target site DNA in a DNA-RNA complementary manner. Cas9/sgRNA recognizes and cleaves the target DNA and causes a DSB (double-strand break), which provides the opportunity of gene mutagenesis and other types of genome manipulation.

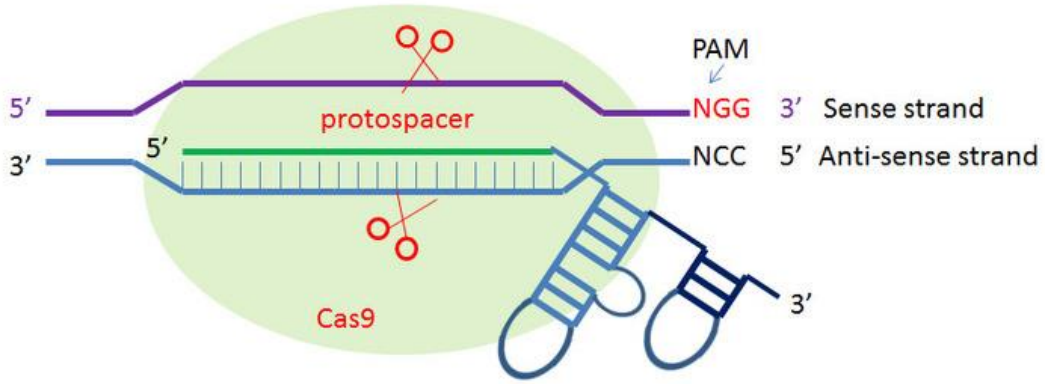

Fig1. The CRISPR/Cas9 system for targeted genome editing

4. What is the CRISPR/Cpf1 system?

CRISPR/Cpf1 is a DNA-editing technology (Fig2). It works analogously to CRISPR/Cas9 which has revolutionized biological research. Like its predecessor, it is derived from a mechanism that bacteria use to prevent genetic damage from viruses. CRISPR/Cpf1 may be better than CRISPR/Cas9 in that Cpf1 is a smaller and simpler endonuclease (a type of enzyme) than Cas9. That simplifies delivery to the cells whose genes need modifying. Two candidate enzymes from Acidaminococcus and Lachnospiraceae display efficient genome-editing activity in human cells.

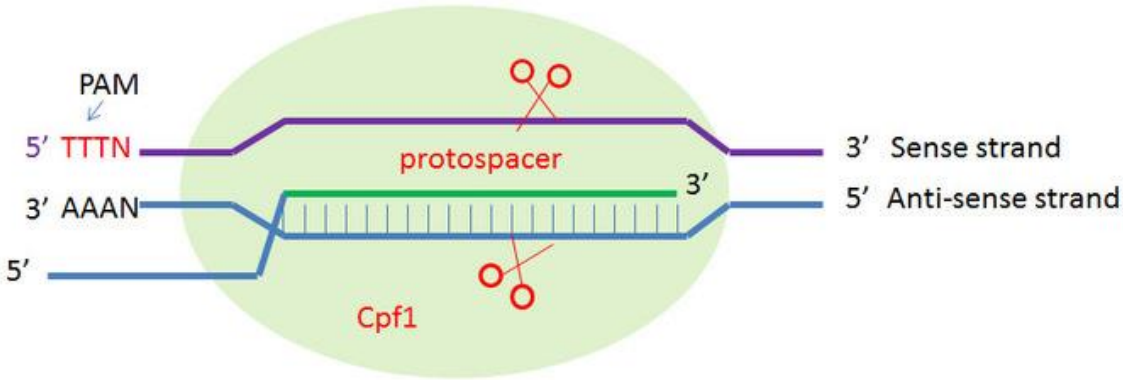

Fig2. The CRISPR/Cpf1 system for targeted genome editing

5. What is the CRISPR/C2c1 system?

CRISPR/C2c1 is a DNA-editing technology (Fig3). It works analogously to CRISPR/Cpf1 which has revolutionized biological research. C2c1 system can mediate DNA interference in a 5'-PAM-dependent fashion analogous to Cpf1. However, unlike Cpf1, which is a single-RNA-guided nuclease, C2c1 depends on both crRNA and tracrRNA for DNA cleavage.

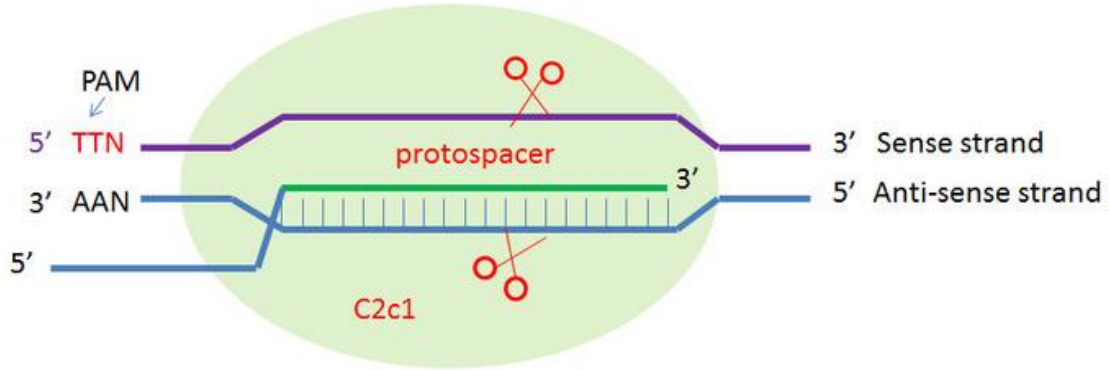

Fig3. The CRISPR/C2c1 system for targeted genome editing

6. What is the Paired-gRNAs strategy of CRISPR TECHNOLOGY?

The paired-gRNAs strategy is reported to show higher specificity than the above single-gRNA strategy, it consists of three components: Cas9-nickase, a mutant form of Cas9 protein, which has no nuclease activity but nickase activity, only cleaving one of the DNA strands with the assist of one sgRNA; and a pair of gRNAs, target two sites with offset no more than tens of nt in the opposite strands of DNA (Fig4). The two close nicks induced by the gRNA pair can cause a DSB. However, single nick in a potential off-target induced by only one gRNA would be difficult to cause DSB.

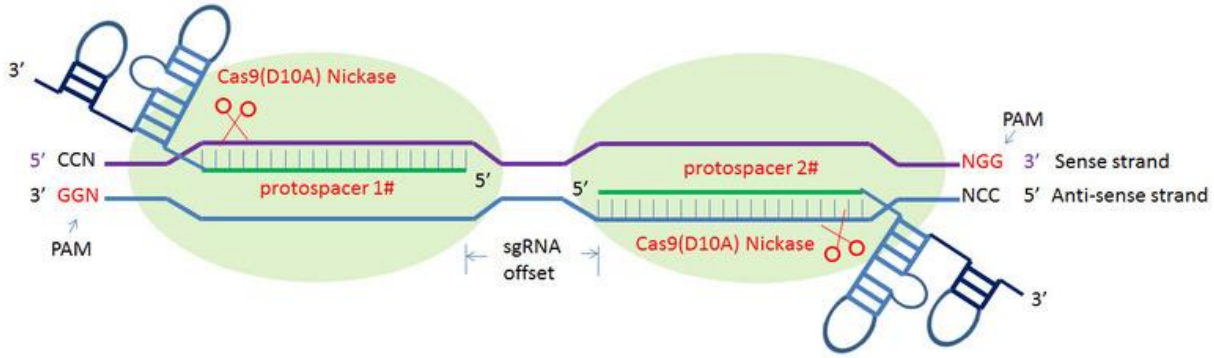

Fig4. The Paired-gRNAs strategy of CRISPR TECHNOLOGY

7. What is the potential off-target site?

Potential off-target site is an unwanted target site in the genome. This is an important consideration when application of CRISPR technology. The specificity of the CRISPR system is determined in large part by how specific the sgRNA targeting sequence is for the genomic target compared to the rest of the genome. Ideally, a sgRNA will have perfect match to the target DNA with no homology elsewhere in the genome. In fact, most of sgRNA targeting sequence will have additional sites throughout the genome where partial homology exists.

8. Is the PAM sequence part of the sgRNA sequence construct?

The PAM sequence is located on the non-complementary strand. In other words, it is on the strand of DNA that contains the same DNA sequence as the target sgRNA. The PAM sequence should not be included in the design of the sgRNA.

III. References

1. Chari R, Yeo NC, Chavez A, Church GM. sgRNA Scorer 2.0: A Species-Independent Model to Predict CRISPR/Cas9 Activity. ACS Synth Biol. 2017. 6(5):902-904.
2. Bae S, Park J, Kim JS. Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. Bioinformatics. 2014. 30(10):1473-5.
3. Doench JG, Fusi N, Sullender M, Hegde M, Vaimberg EW, Donovan KF, Smith I, Tothova Z, Wilen C, Orchard R, Virgin HW, Listgarten J, Root DE. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat Biotechnol. 2016. 34(2):184-191.
4. Perez AR, Pritykin Y, Vidigal JA, Chhangawala S, Zamparo L, Leslie CS, Ventura A. GuideScan software for improved single and paired CRISPR guide RNA design. Nat Biotechnol. 2017. 35(4):347-349.

IV. Technical Support

Feel free to reach out to the CRISPR-offinder team at ssxie@mail.hzau.edu.cn

V. Licensing and Warranty information

If you're interested in getting access to this system under a different license, please contact us.

If you use this program in your research, please cite:
Changzhi Zhao[†], Xiaoguo Zheng[†], Wubin Qu[†], Guanglei Li, Xinyun Li, Yi-Liang Miao, Xiaosong Han, Xiangdong Liu, Zhenhua Li, Yunlong Ma, Qianzhi Shao, Haiwei Li, Fei Sun[*], Shengsong Xie[*] and Shuhong Zhao[*]. CRISPR-offinder: a CRISPR guide RNA design and off-target searching tool for user-defined protospacer adjacent motif. International Journal of Biological Sciences, 2017, under review.