CRISPRdigger manual

Ruiquan Ge[1,2,*], Guoqin Mai[1,*], Pu Wang[1,2,*],Manli Zhou[1,2], Youxi Luo[3], Jikui Liu[1], Fengfeng Zhou[1, #].

[1]Shenzhen Institutes of Advanced Technology, and Key Lab for Health Informatics, Chinese Academy of Sciences,

[2]Shenzhen College of Advanced Technology, University of Chinese Academy of Sciences Shenzhen, Guangdong, P.R. China, 518055.

[3]School of Science, Hubei University of Technology, Wuhan, 430068

Contact: fengfengzhou@gmail.com, ruiquange@gmail.com

Dec. 6th, 2015

1. Introduction

Clustered regularly interspaced short palindromic repeats play a key role in the prokaryote immune systems. CRISPR can apply to gene editing, phage resistance, and biological evolution. we developed a *de novo* CRISPR detection program named CRISPRdigger, which could identify more truncated DR and has higher accuracy and more contents in the test genomes compared with the present other tools-CRISPRFinder[1], CRT[2] and PILER-CR[3]. CRISPRs are composed of a series of highly conserved direct repeats varying in length from 23 to 47 bp, separated by unique spacer sequences in the range of 0.6-2.5 times length of DR. Fig. 1 gives a schematic structure of a CRISPR.

Fig. 1. The schematic structure of a CRISPR. DR represents the direct repeats. SPx (X=1,2,3…) represents the different spacer.

No single software is the complete answer for all CRISPRs' discovery. So in the result part, we also give all the tools' results for users to analysis.

2. Software preparation

2.1 CRISPRdigger

The list of installed software

Perl version 5.8.8 or later : http://www.perl.org/ , Bioperl : http://bioperl.org/DIST/ ,

BLAST : ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/ ,

RepeatMasker[5]: http://www.repeatmasker.org/RMDownload.html ,

RepeatScout-v1.05 [4]: http://bix.ucsd.edu/repeatscout/ ,

ClustalW2.1[6]: http://www.clustal.org/clustal2/ .

The list of modified software

RepeatScout parameter L value was set by the length of the genome file [length(genome)]. In script, the parameter MaxDR was the largest length of the DR [length(MaxDR)]. L value=length(MaxDR)-int(1+log(length(genome))/log(4)).

In scripts, all the parameters had been optimized and suitable for most prokaryotic genomes. You just use the new file (**filter-stage-1.prl**) replace the old one into its directory.

Table 1 The parameters or scripts lists

Programme	Script	Parameter	Parameter value
RepeatScout-v1.05	**build_lmer_table**	tandem	60
		min	2
	RepeatScout	L	
		goodlength	20
		minthresh	2
		tandemdist	60
	filter-stage-1.prl	$MIN_LENGTH	15
		$MAX_LENGTH	60
		$NSEG_THRESHOLD	0.9
RepeatMasker		cutoff	120
BLAST(rmblast-2.2.27)	**blastall**	p	blastn
		W	7
		M	8

The auto-installation of the softwares

Some softwares need to be installed successfully before execute the CRISPRdigger. The AutoconfigCRISPRdigger script can achieve it and auto-configure the envionment variables at the same time. The command line like this: *perl AutoconfigCRISPRdigger.pl*

The RepeatMasker need to be manually configured. Because its installation is interactive. After executing the last script, the tips was given about how to install the RepeatMasker.

After all software were installed successfully, the environment variables should be set in order that all tools could run fluently, or, all the software' executable path should be modified in the scripts which included the command lines like "system(…)". At the same time, all software or executable scripts should be executable for users.

3. Scripts explanation

Table 3 scripts explanation

Script name	Explanation	Command line example
CRISPRdigger.pl	It was only applied in one genome or a DNA sequence.	*perl CRISPRdigger.pl –i inputfile*
batchCRISPR.pl	It was applied in a batch of genomes. The script should be put in the same directory with the genomes.	*perl batchCRISPR.pl*

After the CRISPRdigger script was installed successfully, you can execute other scripts. But, you must modify the bioperl library path and some other softwares' path in the right way in the scripts.

4. Tools' application in one genome

4.1 Input options in CRISPRdigger

In all the input parameters, the parameter *i* is required, and other parameters are optional. The default value of parameters were given in bracket.

The command line example: *perl CRISPRdigger.pl –i inputfile*

The input parameters list as below:

-i *inputfile*

Input one genome or a DNA sequence file in fasta format.

-h

It will display the help information.

-l [0.5]

Spacer's length should be longer than *l* times of DR's.

-m [3.0]

Spacer's length should be shorter than *m* times of DR's.

-g [10]

Spacer's length should be longer than *g* bps.

-p [120]

Spacer's length should be shorter than *p* bps.

-t [0.80]

In RepeatScout, filter-stage-1.prl script output one file named like *. stg1.fasta. In this file, all the same sequence were found. This parameter will combine the high similar sequence as one

repeat template. Its recommended range was [0.7,0.9].

-s [0.5]

In order to confirm the CRISPR, the spacers should be different. This parameter will set the highest similarity as a true CRISPR. If the similarity is more than this value, the CRISPR maybe set questionable or false CRISPR. Its recommended range was [0.4,0.8].

-R [0.5]

This parameter was used to extend the lost DRs between the DRs found by RepeatMasker. If the similarity between the extend sequence and the consistent DR is over the R value, the extend sequence was added into the repeat template arrays. Its recommended range was [0.4,0.8].

-B [0.35]

This parameter was used to extend truncated DRs at the both ends of the CRISPRs. If the similarity between the extend sequence and the consistent DR is over the R value, the extend sequence was added into the CRISPR. Its recommended range was [0.2,0.7].

-d [47]

DR's length should be shorter than *d*.

4.2 Output result files in CRISPRdigger

Table 4 The explanation of the output files in one genome

Tool	Output file	Explanation
CRISPRdigger	*.gff3	the CRISPR positional information including all DR location and questionable CRISPR
	*.dr	the CRISPR detailed information including all DR and spacers sequences
	*.sp	collect all the spacers sequences in the genome arraying by fasta format

5. Tools' application in a group of genomes

The results files in all tools look like *.oldresult and *.result. The *.oldresult files include the CRISPR number , every CRISPR's location and spacers' number. The *.result files include every DRs' location except the *.oldresult files' information. The command line should be in the same directory with the fasta files and has no parameter.

The command line example: *perl batchCRISPR.pl*

Table 5 The explanation of the output files in group of genomes

Output file	Explanation
*. oldresult	the CRISPR numbers; every CRISPR's location and spacers' number
*. result	the CRISPR numbers; every CRISPR's location, spacers' number and DRs' location

6. References

1. **Grissa, I., G. Vergnaud, and C. Pourcel,** ***CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats.*** **Nucleic Acids Res, 2007. 35(Web Server issue): p. W52-7.**
2. **Bland, C., et al.,** ***CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats.*** **BMC Bioinformatics, 2007. 8: p. 209.**
3. **Edgar, R.C.,** ***PILER-CR: fast and accurate identification of CRISPR repeats.*** **BMC Bioinformatics, 2007. 8: p. 18.**
4. **Price, A.L., N.C. Jones, and P.A. Pevzner,** ***De novo identification of repeat families in large genomes.*** **Bioinformatics, 2005. 21 Suppl 1: p. i351-8.**
5. **RepeatMasker,** ***www.repeatmasker.org*****.**
6. **Larkin, M.A., et al.,** ***Clustal W and Clustal X version 2.0.*** **Bioinformatics, 2007. 23(21): p. 2947-8.**