BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 8 2014, pages 1180–1182
doi:10.1093/bioinformatics/btt764

Genome analysis

Advance Access publication January 2, 2014

CasOT: a genome-wide Cas9/gRNA off-target searching tool

An Xiao[1,†], Zhenchao Cheng[1,†], Lei Kong[2], Zuoyan Zhu[1], Shuo Lin[3], Ge Gao[2,*] and Bo Zhang[1,*]

[1]Key Laboratory of Cell Proliferation and Differentiation of the Ministry of Education, [2]State Key Laboratory of Protein and Plant Gene Research, Center for Bioinformatics, College of Life Sciences, Peking University, Beijing 100871, China and [3]Department of Molecular, Cell and Developmental Biology, University of California, Los Angeles, CA 90095, USA

Associate Editor: John Hancock

ABSTRACT

Summary: The CRISPR/Cas or Cas9/guide RNA system is a newly developed, easily engineered and highly effective tool for gene targeting; it has considerable off-target effects in cultured human cells and in several organisms. However, the Cas9/guide RNA target site is too short for existing alignment tools to exhaustively and effectively identify potential off-target sites. CasOT is a local tool designed to find potential off-target sites in any given genome or user-provided sequence, with user-specified types of protospacer adjacent motif, and number of mismatches allowed in the seed and non-seed regions.

Availability: http://eendb.zfgenetics.org/casot/

Contact: zfgenetics@gmail.com or bzhang@pku.edu.cn

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on October 7, 2013; revised on December 6, 2013; accepted on December 24, 2013

1 INTRODUCTION

Cas9/gRNA (guide RNA), or Cas9/single guide RNA, a customized type of clustered regularly interspaced short palindromic repeat/CRISPR-associated (CRISPR/Cas) system, has recently been used for effective gene targeting in many organisms and cultured cells (Pennisi, 2013; Xiao *et al*., 2013). This system is easy to design and engineer; only a single short gRNA, which contains ~20-nt region reverse complementary to one strand of the target DNA (the protospacer) with an -NGG motif in 3′-end (the protospacer adjacent motif, PAM) of the target site, has to be synthesized for a particular target site. The same Cas9 nuclease can be used universally; this notably reduces the cost of both time and labor, promoting the broad and rapidly developing application of this system (Jinek *et al*., 2012).

However, the specificity of Cas9/gRNA needs careful evaluation. For example, Cas9/gRNA seems to have considerable off-target effects in cultured human cells (Cradick *et al*., 2013; Fu *et al*., 2013; Hsu *et al*., 2013; Mali *et al*., 2013a; Pattanayak *et al*., 2013) and in several organisms (Fujii *et al*., 2013; Li *et al*., 2013; Shan *et al*., 2013). *In vitro* experiments suggest that the PAM type could influence the cleavage activity: the -NGG type of PAM shows the highest activity, PAM of -NAG comes second and PAM of -NNGG shows the weakest activity. Mismatches in different regions of the protospacer in the target site also have different effects: a mismatch in the region of the first 12 nt adjacent to the PAM (called the seed region) may cause a notable reduction of the cleavage activity of Cas9/gRNA, whereas mismatches in the other region of the protospacer (the non-seed region) have a much weaker effect (Cong *et al*., 2013; Jiang *et al*., 2013; Liu *et al*., 2013).

An approach to identify potential off-target sites in any given genome or other sequence is necessary to evaluate the off-target effects of Cas9/gRNA *in vivo*. An online tool used for Cas9/gRNA target design also supports the prediction of certain potential off-target sites (Hsu *et al*., 2013); however, neither a detailed scoring algorithm nor an adjustable option is available in this tool, and only a few genomes can be searched for target or off-target sites. Other online tools, TagScan (Iseli *et al*., 2007) or Pattern Match (http://viewer.shigen.info/medakavw/patternmatch/), are available to search for regions in some commonly used genomes or only in the medaka genome, with up to two mismatches. On the other hand, widely used heuristic alignment tools like BLASTN and BLAT are not designed for exhaustive searching for short nucleotide sequences (as short as ~20 nt). In addition, the variable types of PAM with random nucleotides also need to be considered during the search for potential off-target sites.

Here, we report CasOT, a downloadable and open-source tool designed to identify potential off-target sites of given Cas9/gRNA targets or target pairs in any user-specified sequence or genome, with several tunable parameters like PAM type and the mismatch number in the seed (up to 4–6 nt) and non-seed regions (Supplementary Material S1).

2 IMPLEMENTATION AND FEATURES

2.1 Single-gRNA searching mode

CasOT is a Perl script and can run locally in Windows, *nix and Mac OS systems. Three searching modes, single-gRNA, paired-gRNA and target-and-off-target, are provided. In the single-gRNA mode, users are required to provide (i) an input FASTA file including individual 21–33-nt target sites (18–30-nt protospacer plus the PAM -NGG in the 3′-end), (ii) genome sequence files to search for off-target sites or any user-specified sequence and (iii) (optional) a genome annotation file (Ensembl GTF format) for determining whether the potential off-target sites are located in exons. Links to public sequences and

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

annotation files for several widely used genomes are available at the CasOT Web site. Users can specify several parameters for the search, including the allowed level of PAM type (four levels are available), and the maximum number of mismatches allowed in the seed region (0–6 are available, but allowing 5 or 6 mismatches needs a large computer random access memory; two by default) and in the non-seed region (no limit by default).

The program reads the input target sequences and divides them into PAMs, seed regions and non-seed regions. A pool is built to contain all possible 12-nt sequences similar to the seed regions and their reverse complementary sequences, with mismatches no more than that allowed by the user-specified option. The genome sequences or other user-provided sequences are scanned nucleotide-by-nucleotide to identify whether any 12-nt fragment appears in the pre-built seed pool. If found, adjacent genomic sequences of the fragment are extracted and additional test steps are executed to check whether the PAM type and mismatch number in the non-seed region match the user-specified options. If all criteria are met, the position and strand of the fragment in the genome or sequence are recorded. Then the detailed information of potential off-target sites, including locations, sequences, PAM type, numbers and detailed description of mismatches, is output as separate files corresponding to each target site. If the annotation file is provided, the gene ID and gene symbol are output appended to the off-target information when this site is located in an exon (Supplementary Material S4A). A statistic file to count the types of potential off-target sites is also output.

2.2 Paired-gRNA searching mode

A new strategy for gene targeting using two gRNAs (paired-gRNA) instead of a single gRNA was recently reported to reduce off-target effects; in this system, the Cas9 protein is mutated from its original nuclease form to a nickase form (Mali *et al*., 2013b; Ran *et al*., 2013). With the guidance of a single gRNA, the Cas9-nickase only cleaves one strand of the double-stranded DNA and leaves a single nick, which is precisely repaired by the cell repair mechanism and does not cause mutations in general. For the purpose of gene targeting, the Cas9-nickase and a pair of gRNAs are used to generate two nearby nicks located in opposite strands within a short distance; this can mediate a double-strand break and subsequent mutagenesis.

In the paired-gRNA searching mode of CasOT, most of the input and supporting files and the options are similar to the single-gRNA mode, except that the two target sequences in each pair in the input FASTA file must share the same name with different suffixes '_#F' and '_#R' (Fig. 1A). An additional option is supported in this mode where users can specify the maximum distance allowed between the two potential off-target sequences (100 nt by default). The search results for each individual gRNA are output similarly to the single-gRNA mode; and pairs of the potential off-target sites located in opposite strands of the same chromosome within the allowed offset are output as another file (Fig. 1B, Supplementary Material S4B).

2.3 Target-and-off-target searching mode

Sometimes it is necessary to design Cas9/gRNA target sites and search for potential off-targets of these sites sequentially.

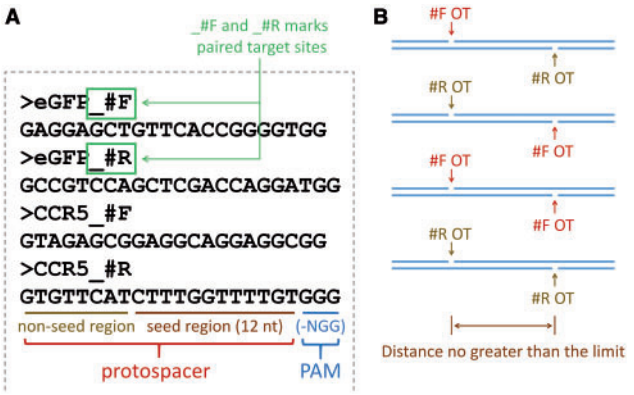

Fig. 1. Sample input file and possible off-target types in the CasOT paired-gRNA searching mode. (**A**) A sample input file (framed by the dashed lines) in which the target sites are paired in groups (#F and #R as a pair). (**B**) Four possible cleavage types caused by two close nicks in the potential off-target sites of a pair of gRNAs (#F and #R), which are all output as potential paired-gRNA off-target sites. OT, potential off-target.

The target-and-off-target mode of CasOT accepts a sequence <1 kb and searches for candidate target sites first. Two parameters for this step can be specified by users: the allowed length range of the protospacers of the sites (18–30 nt; 19–20 nt by default) and the allowed type of sites [only -NGG PAM is required, or a G in the first position (5′-end) for the T7 promoter is required as well]. Then the identified candidate targets are used to search for potential off-target sites in the given genome or sequence, similar to the single-gRNA searching mode.

3 CONCLUSION

We present CasOT, the first exhaustive tool to accurately identify potential off-target sites in any user-specified genome or other sequence with flexible options in an acceptable period of time (Supplementary Material S3). Further description of its usage is available in Supplementary Material S2 and the webpage of this tool.

ACKNOWLEDGEMENTS

The authors thank Dr I. Bruce for language editing of the manuscript, N. Zheng and Y. Zhang for providing Mac OS X platform and all laboratory members for testing the feasibility and applicability of CasOT.

Funding: This work was supported by National Natural Science Foundation of China (NSFC) [31110103904], the 973 programs [2011CBA01000, 2012CB945101, 2011CBA01102], and NSFC [31171242].

Conflict of interest: none declared.

REFERENCES

Cong,L. *et al.* (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science*, **339**, 819–823.

Cradick,T.J. *et al.* (2013) CRISPR/Cas9 systems targeting beta-globin and CCR5 genes have substantial off-target activity. *Nucleic Acids Res.*, **41**, 9584–9592.

Fu,Y. *et al.* (2013) High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat. Biotechnol.*, **31**, 822–826.
Fujji,W. *et al.* (2013) Efficient generation of large-scale genome-modified mice using gRNA and CAS9 endonuclease. *Nucleic Acids Res.*, **41**, e187.
Hsu,P.D. *et al.* (2013) DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.*, **31**, 827–832.
Iseli,C. *et al.* (2007) Indexing strategies for rapid searches of short words in genome sequences. *PLoS One*, **2**, e579.
Jiang,W. *et al.* (2013) RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nat. Biotechnol.*, **31**, 233–239.
Jinek,M. *et al.* (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, **337**, 816–821.
Li,W. *et al.* (2013) Simultaneous generation and germline transmission of multiple gene mutations in rat using CRISPR-Cas systems. *Nat. Biotechnol.*, **31**, 684–686.
Liu,D. *et al.* (2013) Efficient gene targeting in zebrafish mediated by a zebrafish-codon-optimized Cas9 and evaluation of off-targeting effect. *J. Genet. Genomics*, [Epub ahead of print, doi: 10.1016/j.jgg.2013.11.004].
Mali,P. *et al.* (2013a) RNA-guided human genome engineering via Cas9. *Science*, **339**, 823–826.
Mali,P. *et al.* (2013b) CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat. Biotechnol.*, **31**, 833–838.
Pattanayak,V. *et al.* (2013) High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nat. Biotechnol.*, **31**, 839–843.
Pennisi,E. (2013) The CRISPR craze. *Science*, **341**, 833–836.
Ran,F.A. *et al.* (2013) Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell*, **154**, 1380–1389.
Shan,Q. *et al.* (2013) Targeted genome modification of crop plants using a CRISPR-Cas system. *Nat. Biotechnol.*, **31**, 686–688.
Xiao,A. *et al.* (2013) EENdb: a database and knowledge base of ZFNs and TALENs for endonuclease engineering. *Nucleic Acids Res.*, **41**, D415–D422.