Bioinformatics, 2016, 1–3
doi: 10.1093/bioinformatics/btw476
Advance Access Publication Date: 10 July 2016
Applications Note

OXFORD

Genome analysis

CRISPR-DO for genome-wide CRISPR design and optimization

Jian Ma[1], Johannes Köster[2,3,4,5], Qian Qin[1], Shengen Hu[1], Wei Li[2,3], Chenhao Chen[2,3], Qingyi Cao[6], Jinzeng Wang[1], Shenglin Mei[1], Qi Liu[1,*], Han Xu[2,3,*] and Xiaole Shirley Liu[1,2,3*]

[1]School of Life Science and Technology, Tongji University, Shanghai 200092, China, [2]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard School of Public Health, Boston, MA 02115, USA, [3]Center for Functional Cancer Epigenetics, Dana-Farber Cancer Institute, Boston, MA 02215, USA, [4]Division of Molecular and Cellular Oncology, Department of Medical Oncology, Dana-Farber Cancer Institute and [5]Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA and [6]State Key Laboratory for Diagnosis and Treatment of Infectious Diseases, Collaborative Innovation Center for Diagnosis and Treatment of Infectious Diseases, the First Affiliated Hospital, College of Medicine, Zhejiang University, Hangzhou 310003, China

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on November 5, 2015; revised on June 14, 2016; accepted on July 4, 2016

Abstract

Motivation: Despite the growing popularity in using CRISPR/Cas9 technology for genome editing and gene knockout, its performance still relies on well-designed single guide RNAs (sgRNA). In this study, we propose a web application for the Design and Optimization (CRISPR-DO) of guide sequences that target both coding and non-coding regions in spCas9 CRISPR system across human, mouse, zebrafish, fly and worm genomes. CRISPR-DO uses a computational sequence model to predict sgRNA efficiency, and employs a specificity scoring function to evaluate the potential of off-target effect. It also provides information on functional conservation of target sequences, as well as the overlaps with exons, putative regulatory sequences and single-nucleotide polymorphisms (SNPs). The web application has a user-friendly genome–browser interface to facilitate the selection of the best target DNA sequences for experimental design.

Availability and Implementation: CRISPR-DO is available at http://cistrome.org/crispr/

Contact: qiliu@tongji.edu.cn or hanxu@jimmy.harvard.edu or xsliu@jimmy.harvard.edu

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

1 Introduction

CRISPR/Cas9 (clustered regularly interspaced short palindromic repeats/CRISPR-associated protein 9) originally came from bacterial host defense and has provided new insight into site-specific genome editing (Hsu *et al.*, 2014). The CRISPR/Cas9 technology requires an sgRNA with a ~20 bp guide sequence to pair with the target DNA, which enables the Cas9 protein loading to the correct location and introduces a DNA double-strand break (DSB) (Cho *et al.*, 2013; Cong *et al.*, 2013; Jinek *et al.*, 2012; Mali *et al.*, 2013).

To fully utilize the CRISPR genome editing technology, one must consider two essential factors: off-target effect and cleavage efficiency of sgRNA. On one hand, it has been reported that the mismatches in off-target sites, especially those in the 10-12 bases proximal to the PAM, allow less off-target binding (Cong *et al.*, 2013; Hsu *et al.*, 2013; Jinek *et al.*, 2012; Mali *et al.*, 2013). About 17–19 bp truncated sgRNAs are more sensitive to mismatches thus more specific (Fu *et al.*, 2014). CRISPR design tools like *CRISPR-P* (Lei *et al.*, 2014), *E-CRISPR* (Heigwer *et al.*, 2014) and *CasOT*

(Xiao *et al.*, 2014) are mainly focused on the prediction of off-target effect. *Cas-OFFinder* (Bae *et al.*, 2014) is another off-target detection tool with multiple CRISPR Systems supported. On the other hand, for many CRISPR/Cas9 applications, especially for CRISPR screens, sgRNA-induced Cas9 cleavage efficiency is also important. The sgRNA efficiency is predominantly determined by the sequence of the guide and its 3′ flanking region (Wang *et al.*, 2014; Xu *et al.*, 2015). Currently, CRISPR design tools such as *CRISPR-ERA* (Liu *et al.*, 2015), *Benchling* (Benchling, RRID:SCR_013955) and *sgRNA designer* from Broad Institute (Doench *et al.*, 2016) have consideration for both on- and off-target sgRNA design. Meanwhile, other genomic features of an sgRNA target, such as its evolutional conservation, regulatory potential and genetic variations, should also be considered in functional analysis using CRISPR/Cas9 systems (Shi *et al.*, 2015).

To address this need, we propose a web application for the Design and Optimization (CRISPR-DO) of sgRNA sequences in human, mouse, zebrafish, fly and worm genomes. CRISPR-DO integrates an sgRNA efficiency prediction model and an off-target scoring function, which allows the users to evaluate the 'goodness' of an sgRNA in both sensitivity and specificity.

In CRISPR-DO, we annotate each target sequence with the PhastCons conservation score as well as the overlaps with exons, DNase I hypersensitive sites (DHSs), and single-nucleotide polymorphisms (SNPs) for better functional characterization when such data is available. We also integrated our target sequence search result with the powerful WashU Epigenome Browser (Zhou *et al.*, 2011) to enable loading of other genomic tracks and facilitate the visualization and selection of target sequences. Details of online CRISPR-DO can be found in Supplementary materials.

2 Methods

2.1 CRISPR target sequence scan in the whole genome

The workflow to generate the target sequence database is shown in Supplementary Figure S1. We first obtained the full human (GRCh37/hg19 and GRCh38/hg38), mouse (NCBI37/mm9 and GRCm38/mm10), zebrafish (danRer7), fly (dm6) and worm (ce10) genome sequences from UCSC genome database. We removed alternate loci, unlocalized and unplaced (random) sequences. Next we performed a genome-wide sgRNA target sequence scan for PAM sequences on both the forward and the reverse strand in each genome. Here, we identified only 5'-NGG-3' as PAM sequences, since PAM-like NAG sequences have much lower Cas9 loading efficiency (Hsu *et al.*, 2013). We used 19 bp or 20 bp target with its PAM and 7 bp 3' flanking sequence (total 29 bp and 30 bp separately) to build our primary sgRNA target sequence library for further evaluation. The total number of target sequences in each genome is shown in Supplementary Tables S1, S2 and S3.

2.2 Score sgRNA efficiency

The nucleotide composition of the 3' end of the target sequence influences Cas9 loading (Wang *et al.*, 2014). Recently, we have developed a model to predict the efficiency of Cas9 cleavage based on the DNA sequence of an sgRNA target and its flanking regions (Xu *et al.*, 2015). We showed that this model effectively predicts the efficiency of guide RNA in high-throughput CRISPR/Cas9 knockout screens. The coefficients of each nucleotide in the model can be represented as a sequence logo (Supplementary Fig. S2). We applied this model to all target sequences to compute genome-wide efficiency scores. The overall efficiency score distributions are shown in Supplementary Figure S3.

2.3 Measure sgRNA off-target effect

For each target in our primary database, we first used BWA (Li and Durbin, 2009) to map it back to the genome, allowing maximum three mismatches and no gaps. When examining sgRNAs with mismatched mapping, we removed those not followed by NGG/NAG on both strands. For the remaining mismatched mappings, we calculated a specificity score based on Zhang Lab's formula (Hsu *et al.*, 2013) (more details in Supplementary materials). The distributions of the specificity scores in hg38 and mm10 are shown in Supplementary Figure S4.

2.4 CRISPR target sequence annotation

We annotated each target sequence to characterize its evolutionary conservation and to exam its overlap ($\geq$1 bp) with referenced exons, regulatory elements or SNPs. The average conservation score is calculated using UCSC PhastCons (Supplementary Table S4). The exon annotation is from UCSC refGene tables. The SNP annotation is from the NCBI dbSNP database. Peaks from each ENCODE DNase-seq data were merged to form the union of DNase I hypersensitivity regions, representing a comprehensive repertoire of putative regulatory elements in the genome (Consortium, 2012). These annotation features give experimentalists more reference in selecting sgRNAs with the best balance of specificity, efficiency and function.

Acknowledgements

We thank Anya Zhang for polishing the writing of this manuscript. We also want to thank Xin Zhou and Ting Wang for their help setting up the WashU EpiGenome Browser.

Funding

The project was partially supported by the National Natural Science Foundation of China [31329003], NIH R01 HG008728, and the Claudia Adams Barr Award in Innovative Basic Cancer Research from Dana-Farber Cancer Institute.

Conflict of Interest: none declared.

References

Bae,S. *et al.* (2014) Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. *Bioinformatics*, **30**, 1473–1475.

Cho,S.W. *et al.* (2013) Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. *Nat. Biotechnol.*, **31**, 230–232.

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

Consortium,E.P. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Doench,J.G. *et al.* (2016) Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.*, **34**, 184–191.

Fu,Y. *et al.* (2014) Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nat. Biotechnol.*, **32**, 279–284.

Heigwer,F. *et al.* (2014) E-CRISP: fast CRISPR target site identification. *Nat. Methods*, **11**, 122–123.

Hsu,P.D. *et al.* (2013) DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.*, **31**, 827–832.

Hsu,P.D. *et al.* (2014) Development and applications of CRISPR-Cas9 for genome engineering. *Cell*, **157**, 1262–1278.

Jinek,M. *et al*. (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, **337**, 816–821.

Lei,Y. *et al*. (2014) CRISPR-P: a web tool for synthetic single-guide RNA design of CRISPR-system in plants. *Mol. Plant*, **7**, 1494–1496.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Liu,H. *et al*. (2015) CRISPR-ERA: a comprehensive design tool for CRISPR-mediated gene editing, repression and activation. *Bioinformatics*, **31**, 3676–3678.

Mali,P. *et al*. (2013) RNA-guided human genome engineering via Cas9. *Science*, **339**, 823–826.

Shi,J. *et al*. (2015) Discovery of cancer drug targets by CRISPR-Cas9 screening of protein domains. *Nat. Biotechnol*., **33**, 661–667.

Wang,T. *et al*. (2014) Genetic screens in human cells using the CRISPR-Cas9 system. *Science*, **343**, 80–84.

Xiao,A. *et al*. (2014) CasOT: a genome-wide Cas9/gRNA off-target searching tool. *Bioinformatics*, [Epub ahead of print].

Xu,H. *et al*. (2015) Sequence determinants of improved CRISPR sgRNA design. *Genome Res*., **25**, 1147–1157.

Zhou,X. *et al*. (2011) The human epigenome Browser at Washington University. *Nat. Methods*, **8**, 989–990.