J. of xxxxxx, YYYY, 0–0
doi: 10.xxxx/xxxxxx/xxxxx
Advance Access Publication Date: DD Month YYYY
Application Note

Genome analysis

Cas_GUIDE: An on-line database for selecting practical sgRNA for your gene of interest

Xiao Zeng[1,*], Yufang Wang[1] and Hong Zhou[2]

[1]PBSG, LLC, 905 W 7th St., Frederick, MD 217o1, USA

[2]Department of Mathematical Science, School of Health and Natural Sciences, University of Saint Joseph, 1678 Asylum Avenue, West Hartford, CT 06117, USA

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: Cas_GUIDE is a web portal that provides practical sgRNA choices for gene editing with the CRISPR/Cas9 systems. Using a modularized algorithm, which incorporates the most up-to-date rules for Cas9 on-target efficiency with an innovative off-target search module, Cas_GUIDE presents users with up to 3 top sgRNA pairs for their gene of interest. To minimize off-target effect, Cas_GUIDE recommends Cas9 nickase by default. For those genes without paired designs, individual sgRNAs are offered. Cas_GUIDE also provide researchers with PCR primer pairs flanking the on-target as well as relevant off-target sites, should there be any.
Availability: Human sgRNA library is freely available at http://pbsgweb.com/Cas_GUIDE.
Contact: xiao.zeng@planbscientific.com
Supplementary information: Supplementary information is available at upon request.

1 Introduction

Once demonstrated as being functional in eukaryotes, especially in mammalian systems, the CRISPR/Cas9 has become the fastest adopted tool in the molecular biology tool box since the invention of PCR (Ledford, 2015). Very few additional rules are required to construct a working sgRNA (Doench *et al.*, 2014), besides the complimentary requirement of twenty or so nucleotides (20 nt) immediately upstream of a PAM sequence, or the protospacer adjacent motif. This lack of stringency is also apparently the reason for the so-called off-target effect. In some cases, sgRNAs with as many as five mismatches can still augment a Cas9-mediated DNA editing (Fu *et al.*, 2013; Hsu *et al.*, 2013; Mali *et al.*, 2013). These reports suggest that a trade-off between on-target efficiency and off-target effect must be considered when selecting sgRNA for any specific experiment. To address this issue experimentally, Ran *et al.* created a Cas9 nickase mutant that requires a paired sgRNAs to introduce double-strand breaks. This modification has improved target specificity by 50- to 1,500-fold (Ran *et al.*, 2013).

In current study, we present Cas_GUIDE, which incorporates a novel and comprehensive off-target search algorithm that walks through the whole genome sequences for mismatches up to 4nt with any given sgRNA candidate. DNA or RNA bulges in those mismatch formations are included. In its final format, Cas_GUIDE generates up to three sgRNA pairs for each protein-coding gene with minimum potential off-target effect in humans. The libraries for mouse, rat and other experimental model organisms will be added once their designs are finished.

2 Methods

Protein coding genes are obviously the first targets that most biomedical researchers would try on with the newly discovered CRISPR/Cas9 gene editing tool. To test our Cas_GUIDE algorithm, we employed the human protein-coding refseq sequences as input.

For on-target efficiency score, we adopted the updated rational design by Doench *et al.* (Doench *et al.*, 2016) with the following modifications: **a**) we preferred sgRNAs with a narrower GC content between 30-50%, which have a lower tendency to cause off-target effect (Lin *et al.*, 2014); **b**) we only selected designs of sgRNA inside the 5'-half of CDS (coding sequences) and no cross exon boundary designs were allowed; **c**) we avoided sgRNA containing a run of (continuous) four or more thymidines (Ts), as well as adenosines (As); **d**) we filtered out a common trinucleotide tandem repeat, $(cac)_3$ /$(gtg)_3$.

We qualified genomic sequence as an "off-target site" for any given sgRNA based on the criteria described before (Lin *et al.*, 2014; Doench *et al.*, 2016; Supplementary).

The Smith–Waterman algorithm is the most thorough local sequence alignment method to identify off-target that was successfully implemented in siRNA designs (Zhou *et al.* 2006). However, using it for a genome-wide sgRNA off-target search is too computationally expensive to be practical. To shorten the time of the search, we first built a dynamic, indexed two-dimensional array representing all the seed sequence variations. Then the off-target search is simply computing the mismatches and DNA/RNA bulges for each element in this array (Supplementary).

3 Conclusions

To understand how disruptive the off-target effect can be for researchers using human cells, Zhou *et al.* built a mathematical model to assess the actual number of off-target sites (Zhou *et al.*, 2016). In this model, using a similar but slightly less stringent off-target definition than ours they estimated that any given sgRNA would have **107** off-target sites in the human genome on average. They also estimated that if a paired sgRNAs are required for any given target, the number of potential off-targets drops to **0.000546** per human genome.

In this study, running through 19,312 mRNA NM_ records in the human genome (one NM record per gene), our algorithm generated 2,744,617 sgRNAs, which gave **142** sgRNAs for each mRNA transcript on average. For these sgRNAs, 2,097,301,944 off-target candidates were found, which means on average **764** off-targets for each sgRNA. The categories of the off-target candidates are summarized in the table below:

Mismatches	0	1	2	3	4
	1.3%	4.0%	8.3%	16.9%	63.5%
Bulge + mismatches	DNA bulges		RNA bulges		
	1-base	2 bases	1-base	2 bases	
	4.00%	NA	1.87%	0.03%	

When we further analyzed the off-target locations, we found that 42.01% of the off-targets reside in intergenic regions, while 57.99% of them are located within annotated genes. 45.73% of all can be mapped to RNA or protein products with either NM or XM records. But only 2.80% are located inside CDS regions. Based on this, one can estimate that on average each sgRNA would have about twenty-one off-targets (764 x 2.8% = **21**) that may directly affect another protein-coding gene. The process of generating an sgRNA pair is quite straightforward, which requires two sgRNAs lined up "back-to-back", whose 5' ends are separated between -4 to +20 bps (Ran *et al.*, 2013). Of the total 19,312 protein-coding genes, 502,061 sgRNA pairs were generated and 395,390 (80%) of them have no off-target match in human genome.

Built upon the aforementioned database and with simplicity in mind, the Cas_GUIDE serves as an easy-to-use, practical sgRNA web portal for the benefit of all researchers. Once the user enters a gene symbol in a search box, three sgRNA pairs will typically presented that have passed the rigorous off-target check mentioned above. For instances where sgRNA pairs are not available, up to six individual sgRNAs will be displayed, which have the least effect on other protein-coding genes predicted by our algorithm.

Cas_GUIDE also provides a list of PCR primer pairs that can be used to validate the on-target gene editing events. For those that may cause off-target effects, additional PCR primer pairs were provided to monitor gene-editing elsewhere in the genome. It is worth noting that all primer pairs in our database have been checked for off-target homologies in the same manner as used in the sgRNA design itself.

Funding

Conflict of Interest: none declared.

Acknowledgements

We would like to thank Mr. Ben Brenneman for critical reading of the manuscript and valuable suggestions on the navigation functions of our web portal.

References

Doench,J.G. *et al.* (2014) Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat. Biotechnol.* **32**, 1262-1267.

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.* (2013) High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat. Biotechnol.*, **31**, 822-826.

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

Ledford,H. (2015) CRISPR, the disruptor. *Nature* **522**, 20–24.

Lin,Y. *et al.* (2014) CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences. *Nucleic Acids Res.* **42**, 7473-7485.

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

Zhou,H. *et al.* (2006) A three-phase algorithm for computer aided siRNA design. *Informatica* **30**, 357-364.

Zhou,H. *et al.* (2016) Mathematical and computational analysis of CRISPR Cas9 sgRNA off-target homologies. *Proc. IEEE Int. Conf. Bioinforma. Biomed*. 449-454