Prediction of off-target effects of the CRISPR/Cas9 system for design of sgRNA

Calvin Guo1*, David Zhen1**†

1*No. 989, Baise Road, Shanghai, China.

Abstract. CRISPR/Cas9 genome editing technology is the frontier of life science research. They have been used to cure human genetic diseases, achieve cell personalized treatment, develop new drugs, and improve the genetic characteristics of crops and other fields. This system relies on the enzyme Cas9 cutting target DNA (on target) under the guidance of sgRNA, but it can also cut non-target sites, which results in off-target effects, thus causing uncontrollable mutations. The risk of off-target effect in CRISPR technology is the main limiting factor that affects the widespread application of CRISPR technology. How to evaluate and reduce the off-target effect is the urgent problem to be solved. In this work, we build up a model that can predict the score of being off-target. Through comparison with the complete genome of the target and precise mathematics that calculate the potential risk of being off-target, we optimize the sgRNA, which is capable of reducing the off-target effect. The result has proven that we can efficiently and quickly identify and screen the best editing target sites with our model. The CRISPR/Cas9 system, not even being perfected yet, has already demonstrated its potential in the field of genome editing. Hopefully through our model, the CRISPR/Cas9 system can quickly apply to more branches in life science and cure those diseases that have been previously incurable.

1 INTRODUCTION

Targeted genome editing has become an increasingly popular field in bioengineering as more methods of the editing have been discovered and can be applied to a wider range of organisms1-3. One method, known as the clustered regularly interspaced short palindromic repeats (CRISPR) associated 9 system (Cas9), has recently been discovered and applied to edit the genome of various organisms4, 5. It is the third generation of enzyme-related gene editing tool, right after zinc-finger nucleases (ZFN) and transcription activator-like effectors nucleases (TALENs)6. The CRISPR/Cas9 system was initially discovered as the “immune system” for bacteria and archaea, which uses this strategy to fight against viruses that inject DNA into them7-9. The system has a prominent aspect as it only requires designing a single-guided RNA (sgRNA), which usually consists of twenty base pairs, that is capable of leading the Cas9 enzyme to bind to the site where the DNA sequence needs to be modified10-14. The sgRNA also helps identify with a specific sequence, called the protospacer adjacent motif (PAM), which is crucial in locating the segment of gene which needs modifications15-18. Thus, the CRISPR/Cas9 system can target any locus in any given genome19-22. This method has become more widely used because of its accuracy, relatively cheap costs, and convenience23-26. Other potential usages of such method include killing viruses and cancer cells27-30. However, this method has not been perfected as the limitation of 20 nucleotides in the sgRNA exposes this method to the possibility of being off-target31-34. Since it only contains 20 nucleotides, the same sequence is vulnerable to repetition within the genome and therefore the Cas9 enzyme has the possibility of being guided to the wrong locus, which in turn leads to mutations35-39.

Therefore, we will explore deep into the CRISPR/Cas9 system, and analyze statistical data from experiments to generate a relatively simple model that can be used to calculate the score of the mismatches in order to reduce the off-target effect.

2 Methods

1. Design a single-guided RNA sequences that helps guiding the Cas9 to the targeted genome (Figure 1).
2. Compare the targeted genome with the database of gene sequences.
3. Identify the repeating sequences of DNA with the targeted segment of gene in the targeted genome.
4. Identify the position of the repeating bases.
5. Input the position into the designated formula.
6. For multiple repeating strands of DNA, we take the strand which has the highest score as the final mismatch score.

*Corresponding author: calvingyw@outlook.com
** Corresponding author: DaveZaye@outlook.com
†The two authors contributed equally to this paper.
7. After gaining all the score values, the sgRNA sequence with the highest score value will be selected.

![Workflow of our model. The input sequence is scanned to identify sgRNA target sites according to the parameters specified. For each candidate target site, the potential off-target sites are determined using Bowtie1. Each potential off-target and its score is computed. With this information each candidate is ranked and finally the results are provided.](image)

3 Results and discussion

From any provided DNA sequence like green fluorescent protein (GFP) all sgRNA target sites will be identified according to parameters like the type of PAM (Figure 2). For off-target predictions, the selection of a PAM type should be made separately (Figure 2). Experimental evidence indicates Cas9 nuclease activity strongly correlating with the mismatch position. Mismatches close to the PAM site will most likely abolish the introduction of a double strand break, while more distant mismatches are tolerated 40, 41. We incorporate these findings as a parameter, and based on the experimental data, we build a mathematical model to calculate the probability of editing success when different bases are mismatched at different positions.

![The identification of sgRNA target sites and the off-target prediction.](image)

Before implementing the model, we input the target sequence of DNA into the database to identify similar sequence that exist in the genome. After this, we start designing the sgRNA so that we can compare the off-target score between the distinguished combinations of sgRNA. To calculate the off-target effect, we define the position of each base of the sgRNA. The base most far away from the PAM sequence is defined as 20, and the position decreases along with the distance from the PAM sequence, measured in units of base. By defining the positions, we utilize the fact that when the similar sequence exists near the PAM, the possibility of being off-target increases significantly.

![The relation between position and active rate of different bases.](image)

Through the experimental data (Supporting tables 1-2) about the percent activeness relating to the position of the nucleotide bases, a graph can be simulated in supporting table 3-6 42. Though the data only provide 20 position, it is enough information to formulate a graph with it (Figure 3). After the data is provided, then the information is inputted into excel to generate a line...
The y-axis shows the percent active in the sgRNA for each mismatch bases, and the x-axis represents the position. There are four graphs being derived through excel according to different bases. It is worth noticing that the general trend of the percent activeness of the sgRNA is decreasing as the mismatch position is getting closer to the PAM. Therefore, it verifies our hypothesis that mismatch bases affect the off-target rate the highest when the base is closer to PAM, and as the mismatch base is further away from the PAM sequence, it will be less effective.

To find a graph to fit our hypothesis, the formula $Y_A = -0.0052X + 0.4617$, $Y_T = -0.0075X + 0.7642$, $Y_G = -0.0319X + 0.8741$, $Y_C = -0.0144X + 0.7698$ are calculated (Figure 3. b)-e), Supporting tables 3-6). The $Y$ is activeness the candidate site, and the $P$ is the position of the mismatch. The $Y$ provides a value of activeness to each mismatch bases at the different position. If the candidate site have more than one base mismatch, $P_i = Y_A * Y_T * Y_G * Y_C$. $P_i$ is the activeness of this candidate site. If the candidate site have more than one base strand mismatch, $P_i = \text{MIN}(P_i)$. Finally, output the $\text{MAX}(P_i)$, then, the candidate is the best one and have the largest activeness.

4 Conclusion

By building this model, we can predict the off-target effects of the CRISPR/Cas9 system in the design of sgRNA. With knowing the off-target effects to each corresponding sgRNA, it can tell us whether the that sgRNA is applicable or not. The model can eliminate the sgRNA that have high off-target rates, and allow us to select the ones that have lower off-target rate. The model can improve the efficiency in the process of designing the sgRNA, and it can also reduce the probability of failure. Meanwhile, the model is simple as everything needed is to calculate the mismatch score of each base with the modified function and multiply the scores of mismatched bases. The final result is a direct indication of the applicability of sgRNA.

Through this model, as introduced, we wish to widen to applicability of the CRISPR/Cas9 system as it is a
an advanced gene-modifying technique that possess huge potentials. For instance, the first genetically modified human babies are edited by CRISPR, and the babies are immune towards the HIV disease, which is currently incurable by any means. Despite the ethical issues, the CRISPR technique has actually displayed itself as an advanced gene-modifying method that will be widely used in the future in order to cure more diseases without interfering with ethical issues. Hopefully our model can contribute to the progress of CRISPR becoming a world-wide technique with high efficiency and accuracy by means of eliminating the possibility of off-target effect through deliberate selection of sgRNA.

References

1. Miller, J. C.; Tan, S.; Qiao, G.; Barlow, K. A.; Wang, J.; Xia, D. F.; Meng, X.; Paschon, D. E.; Leung, E.; Hinkley, S. J.; Dulay, G. P.; Hua, K. L.; Ankoudinova, I.; Cost, G. J.; Urnov, F. D.; Zhang, H. S.; Holmes, M. C.; Zhang, L.; Gregory, P. D.; Rebar, E. J., A TALE nuclease architecture for efficient genome editing. Nat Biotechnol 2011, 29 (2), 143-8.

2. Straimer, J.; Lee, M. C.; Lee, A. H.; Zeitler, B.; Williams, A. E.; Pearl, J. R.; Zhang, L.; Rebar, E. J.; Gregory, P. D.; Llinas, M.; Urnov, F. D.; Fidock, D. A., Site-specific genome editing in Plasmodium falciparum using engineered zinc-finger nucleases. Nat Methods 2012, 9 (10), 993-8.

3. Urnov, F. D.; Rebar, E. J.; Holmes, M. C.; Zhang, H. S.; Gregory, P. D., Genome editing with engineered zinc finger nucleases. Nat Rev Genet 2010, 11 (9), 636-46.

4. Zhang, H.; Zhang, J.; Wei, P.; Zhang, B.; Gou, F.; Feng, Z.; Mao, Y.; Yang, L.; Zhang, H.; Xu, N.; Zhu, J. K., The CRISPR/Cas9 system produces specific and homozygous targeted gene editing in rice in one generation. Plant Biotechnol J 2014, 12 (6), 797-807.

5. Chen, C.; Fenk, L. A.; de Bono, M., Efficient genome editing in Caenorhabditis elegans by CRISPR-targeted homologous recombination. Nucleic Acids Res 2013, 41 (20), e193.

6. Bae, T.; Kim, H.; Kim, J. H.; Kim, Y. J.; Lee, S. H.; Ham, B. J.; Hur, J. K., Specificity Assessment of CRISPR Genome Editing of Oncogenic EGFR Point Mutation with Single-Base Differnces. Molecules 2019, 25 (1).

7. Banakar, R.; Eggenberger, A. L.; Lee, K.; Wright, D. A.; Murugan, K.; Zarecor, S.; Lawrence-Dill, C. J.; Sashital, D. G.; Wang, K., High-frequency random DNA insertions upon co-delivery of CRISPR-Cas9 ribonucleoprotein and selectable marker plasmid in rice. Sci Rep 2019, 9 (1), 19902.

8. Barrangou, R., Foresight is 2020: Ten Bold Predictions for the New CRISPR Year. CRISPR J 2019, 2 (6), 341-342.

9. Becu-Villalobos, D., CRISPR-CAS9 in medicine, the saga continues]. Medicina (B Aires) 2019, 79 (6), 522-523.

10. Crowley, V. M.; Catching, A.; Taylor, H. N.; Borges, A. L.; Metcalf, J.; Bondy-Denomy, J.; Jackson, R. N., A Type IV-A CRISPR-Cas System in Pseudomonas aeruginosa Mediates RNA-Guided Plasmid Interference In Vivo. CRISPR J 2019, 2 (6), 434-440.

11. Cui, X.; Balcerzak, M.; Schernthaner, J.; Babic, V.; Datla, R.; Brauer, E. K.; Labbe, N.; Subramaniam, R.; Ouellet, T., Correction to: An optimised CRISPR/Cas9 protocol to create targeted mutations in homoeologous genes and an efficient genotyping protocol to identify edited events in wheat. Plant Methods 2019, 15, 163.

12. de Maat, V.; Stege, P. B.; Dedden, M.; Hamer, M.; van Pijkeren, J. P.; Willems, R. J. L.; van Schaik, W., CRISPR-Cas9-mediated genome editing in vancomycin-resistant Enterococcus faecium. FEMS Microbiol Lett 2019, 366 (22).

13. Decker, C. E.; Young, T.; Pasnioskwi, E.; Chiu, J.; Song, H.; Wei, Y.; Thurston, G.; Daly, C., Genome-scale CRISPR activation screen uncovers tumor-intrinsic modulators of CD3 bispecific antibody efficacy. Sci Rep 2019, 9 (1), 20068.

14. Dempster, J. M.; Pacini, C.; Pantel, S.; Behan, F. M.; Green, T.; Krill-Burger, J.; Beaver, C. M.; Younger, S. T.; Zhivich, V.; Najgebauer, H.; Allen, F.; Goncalves, E.; Shepherd, R.; Doench, J. G.; Yusa, K.; Vazquez, F.; Parts, L.; Boehm, J. S.; Golub, T. R.; Hahn, W. C.; Root, D. E.; Garnett, M. J.; Tsherniak, A.; Iorio, F., Agreement between two large pan-cancer CRISPR-Cas9 gene dependency data sets. Nat Commun 2019, 10 (1), 5817.

15. Jackow, J.; Guo, Z.; Hansen, C.; Abaci, H. E.; Doucet, Y. S.; Shin, J. U.; Hayashi, R.; De Lorenzo, D.; Kabata, Y.; Shinkuma, S.; Salas-Alanis, J. C.; Christiano, A. M., CRISPR/Cas9-based targeted genome editing for correction of recessive dystrophic epidermolysis bullosa using iPSC cells. Proc Natl Acad Sci U S A 2019.

16. Jeon, J.; Park, J. S.; Min, B.; Chung, S. K.; Kim, M. K.; Kang, Y. K., Retroelement Insertion in a CRISPR/Cas9 Editing Site in the Early Embryo Intensifies Genetic Mosaicism. Front Cell Dev Biol 2019, 7, 273.

17. Jeong, K.; Munoz-Bodnar, A.; Arias Rojas, N.; Poulin, L.; Rodriguez, R. L.; Gagnevin, L.; Verniere, C.; Pruvost, O.; Koebnik, R., CRISPR elements provide a new framework for the genealogy of the citrus canker pathogen Xanthomonas citri pv. citri. BMC Genomics 2019, 20 (1), 917.

18. Johansen, I. E.; Liu, Y.; Jorgensen, B.; Bennett, E. P.; Andreasson, E.; Nielsen, K. L.; Blennow, A.; Petersen, B. L., High efficacy full allelic CRISPR/Cas9 gene editing in tetraploid potato. Sci Rep 2019, 9 (1), 17715.

19. Viswanatha, R.; Brathwaite, R.; Hu, Y.; Li, Z.; Rodiger, J.; Merckaert, P.; Chung, V.; Mohr, S. E.; Perrimon, N., Pooled CRISPR Screens in
Drosophila Cells. Curr Protoc Mol Biol 2019, 129 (1), e111.

20. Vyas, V. K.; Bernstein, D. A., An Introduction to CRISPR-Mediated Genome Editing in Fungi. J Microbiol Biol Educ 2019, 20 (3).

21. Wang, M.; Chen, K.; Wu, Q.; Peng, R.; Zhang, R.; Li, J., RCasFISH: CRISPR/dCas9-Mediated In Situ Imaging of mRNA Transcripts in Fixed Cells and Tissues. Anal Chem 2019.

22. Wang, P. A.; Xiao, H.; Zhong, J. J., CRISPR-Cas9 assisted functional gene editing in the mushroom Ganoderma lucidum. Appl Microbiol Biotechnol 2019.

23. Wu, J.; Tang, Y.; Zhang, C. L., Targeting N-Terminal Huntingtin with a Dual-sgRNA Strategy by CRISPR/Cas9. Biomed Res Int 2019, 2019, 1039623.

24. Wu, J.; Yin, H., Engineering guide RNA to reduce the off-target effects of CRISPR. J Genet Genomics 2019.

25. Xie, L.; Huang, J.; Li, X.; Dai, L.; Lin, X.; Zhang, J.; Luo, J.; Zhang, W., Generation of a homozygous HDAC6 knockout human embryonic stem cell line by CRISPR/Cas9 editing. Stem Cell Res 2019, 41, 101610.

26. Xing, H.; Meng, L. H., CRISPR-cas9: a powerful tool towards precision medicine in cancer treatment. Acta Pharmacol Sin 2019.

27. Xu, C. F.; Chen, G. J.; Luo, Y. L.; Zhang, Y.; Zhao, G.; Lu, Z. D.; Czarna, A.; Gu, Z.; Wang, J., Rational designs of in vivo CRISPR-Cas delivery systems. Adv Drug Deliv Rev 2019.

28. Xu, S.; Zhan, M.; Jiang, C.; He, M.; Yang, L.; Shen, H.; Huang, S.; Huang, X.; Lin, R.; Shi, Y.; Liu, Q.; Chen, W.; Mohan, M.; Wang, J., Genome-wide CRISPR screen identifies ELP5 as a determinant of gemcitabine sensitivity in gallbladder cancer. Nat Commun 2019, 10 (1), 5492.

29. Yang, L. Z.; Wang, Y.; Li, S. Q.; Yao, R. W.; Luan, P. F.; Wu, H.; Carmichael, G. G.; Chen, L. L., Dynamic Imaging of RNA in Living Cells by CRISPR-Cas13 Systems. Mol Cell 2019, 76 (6), 981-997 e7.

30. Ye, S.; Chen, G.; Kohnen, M. V.; Wang, W.; Cai, C.; Ding, W.; Wu, C.; Gu, L.; Zheng, Y.; Ma, X.; Lin, C.; Zhu, Q., Robust CRISPR/Cas9 mediated genome editing and its application in manipulating plant height in the first generation of hexaploid Ma bamboo (Dendrocalamus latiflorus Munro). Plant Biotechnol J 2019.

31. Zheng, A.; Chevalier, N.; Calderoni, M.; Dubuis, G.; Dormond, O.; Ziros, P. G.; Sykiotis, G. P.; Widmann, C., CRISPR/Cas9 genome-wide screening identifies KEAP1 as a sorafenib, lenvatinib, and regorafenib sensitivity gene in hepatocellular carcinoma. Oncotarget 2019, 10 (66), 7058-7070.

32. Zhuang, X.; Veltrei, D. P.; Long, E. O., Genome-Wide CRISPR Screen Reveals Cancer Cell Resistance to NK Cells Induced by NK-Derived IFN-gamma. Front Immunol 2019, 10, 2879.

33. Applications of Next Generation CRISPR-derived Technologies. Hum Gene Ther 2020, 31 (1-2), 1.

34. Ahmad, S.; Wei, X.; Sheng, Z.; Hu, P.; Tang, S., CRISPR/Cas9 for development of disease resistance in plants: recent progress, limitations and future prospects. Brief Funct Genomics 2020.

35. Bai, H.; Liu, L.; An, K.; Lu, X.; Harrison, M.; Zhao, Y.; Yan, R.; Lu, Z.; Li, S.; Lin, S.; Liang, F.; Qin, W., CRISPR/Cas9-mediated precise genome modification by a long ssDNA template in zebrafish. BMC Genomics 2020, 21 (1), 67.

36. Bailis, W., CRISPR/Cas9 Gene Targeting in Primary Bone Marrow-Derived Macrophages. Methods Mol Biol 2020, 2097, 223-230.

37. Bao, A.; Tran, L. P.; Cao, D., CRISPR/Cas9-Based Gene Editing in Soybean. Methods Mol Biol 2020, 2107, 349-364.

38. Benyousef, A.; Marchitto, L.; Touzot, F., CRISPR genome-engineered CYBB (ko) THP-1 cell lines highlight the crucial role of NADPH-induced reactive oxygen species for regulating inflammasome activation. J Allergy Clin Immunol 2020.

39. Blir, E.; Vatangolu Lutz, E. E.; Ozgonul, M. L., Ethical and scientific issues of gene-edited twin by clustered regularly interspaced short palindromic repeats (CRISPR) Cas9 technology. J Turk Ger Gynecol Assoc 2020.

40. Sternberg, S. H.; Redding, S.; Jinek, M.; Greene, E. C.; Doudna, J. A., DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. Nature 2014, 507 (7490), 62-7.

41. Cong, L.; Ran, F. A.; Cox, D.; Lin, S.; Barretto, R.; Habib, N.; Hsu, P. D.; Wu, X.; Jiang, W.; Marraffini, L. A.; Zhang, F., Multiplex genome engineering using CRISPR/Cas systems. Science 2013, 339 (6121), 819-23.

42. Doench, J. G.; Fusi, N.; Sullender, M.; Hegde, M.; Vaimberg, E. W.; Donovan, K. F.; Smith, I.; Tothova, Z.; Wilen, C.; Orchard, R.; Virgin, H. W.; Listgarten, J.; Root, D. E., Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat Biotechnol 2016, 34 (2), 184-191.