kRISP-meR: A Reference-free Guide-RNA Design Tool for CRISPR/Cas9

Mahmudur Rahman Hera¹, Amatur Rahman¹, and Atif Rahman^{1,*}

¹Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology, Dhaka, Bangladesh

*Corresponding author. Email: atif@cse.buet.ac.bd

Abstract

Genome editing using the CRISPR/Cas9 system requires designing guide RNAs (sgRNA) that are efficient and specific. Guide RNAs are usually designed using reference genomes which limits their use in organisms with no or incomplete reference genomes. Here, we present kRISP-meR, a reference free method to design sgRNAs for CRISPR/Cas9 system. kRISP-meR takes as input a target region and sequenced reads from the organism to be edited and generates sgRNAs that are likely to minimize off-target effects. Our analysis indicates that kRISP-meR is able to identify majority of the guides identified by a widely used sgRNA designing tool, without any knowledge of the reference, while retaining specificity.

Keywords

CRISPR, SpCas9, gRNA design, reference-free, k-mer

1 Introduction

CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) has revolutionized the field of genome editing. By delivering the Cas9 nuclease complex with synthetic single-guide RNAs (sgRNA) into a cell, the cell's genome can be cut at a desired locations, allowing genes to be knocked out and edited [1]. The ability to program the target by designing sgRNAs makes CRISPR a robust technology and has opened possibilities of endless applications [2]. CRISPR was successfully used to distort sex-ratio in *Anopheles gambiae* [3], reducing number of females in the population and thus controlling the growth. Human T cells were modified using CRISPR for the treatment of X-Linked Hyper-IgM Syndrome, a condition that affects the immune system [4]. CRISPR/Cas9 genome editing system was used to precisely remove the HIV-1 genome spanning from human T-lymphoid cells [5]. Treatment of primary open-angle glaucoma was successfully achieved in mouse eyes using CRISPR [6].

The genomic modifications made by CRISPR are *guided* by the single-guide RNAs deployed in the Cas9 enzyme, i.e. editing takes place where the guide RNAs bind to the genome. It is therefore important to design sgRNAs that bind to the target, and do not have off-target effects. Several models have been proposed to profile the off-target and on-target activities of the sgRNAs in CRISPR [7, 8, 9]. Using these models, a number of computational tools have been developed to design the sgRNAs [10, 11, 12, 13, 14].

However, these tools rely on the reference genome to design the guide RNAs (gRNA). This has a number of limitations. First, constructing a reference genome is often challenging and expensive and hence genomes of most organisms are incomplete. As a result designing guide RNAs for many organisms including agricultural crops is difficult. Second, the strain or individual to be edited may be genetically different from the reference. Recent studies have showed that considerable variation is present across human populations and individuals contain large scale variations such as duplications not present in the reference [15]. The issue of individual specific genomic variations has been addressed in tools that are capable of designing personalized sgRNAs, namely AlleleAnalyzer [16], CrisFlash [17] and CRISPRitz [18] while Sun et al. presented a tool to design sgRNAs for non-reference plant genomes [19]. However, these tools use reference genomes and sequencing reads to detect variations and may lack specificity if the individual contains duplications or regions missing from the reference. Synthetic sex-ratio distortion using CRISPR was achieved by Papathanos et al. [20], using a tool that does not require a reference genome, but it is not applicable as a general purpose sgRNA designing tool.

To address these issues, we present *kRISP-meR*, a reference-free sgRNA designing tool which can operate with only sequenced reads from the organism. *kRISP-meR* takes as input a target region and the reads, identifies candidate sgRNAs and scores each candidate sgRNA using the numbers of times k -mers i.e. sequences of length k appears in the reads. We show that *kRISP-meR* is able to design sgRNAs accurately in absence of the reference genome, and that it can identify nearly all the guides that are identified by a popular sgRNA designing tools, GuideScan [13] while retaining specificity.

2 Methods

2.1 Overview

The steps of *kRISP-meR* are shown in Figure 1. It takes as input a set of sequenced reads from the genome of the individual to be edited, and a target sequence which corresponds to the region in the genome where editing is intended. As the the individual may have genetic variations in the target region, the sequenced reads are aligned to the target sequence and the target sequence is modified to generate a personalized target sequence. The personalized target sequence is then searched for candidate sgRNAs.

Next we determine the number of times each k -mer is present in the set of sequenced reads, where the value of k is determined by the lengths of gRNA and protospacer adjacent motif (PAM) sequences. The counts are then used to calculate scores of the candidate gRNAs that reflect their efficiency and specificity. The steps are described in more details in the following sections.

2.2 Identifying candidate gRNAs

In order to determine the target sequence for the individual, first, the sequencing reads are mapped on the target sequence using Bowtie2 [21]. With the output BAM file, Pilon [22] is used to determine the genomic variations for the individual and the target sequence is polished to generate a personalized target sequence. Next, *kRISP-meR* scans the personalized target sequence for the PAM sequence, which is NGG by default and identifies strings of length 20 bp (excluding the PAM) upstream of the PAM sequence. Together with the PAM, the 23 bp long sequences are identified as the candidate guide RNAs.

2.3 Counting k-mers in sequenced reads

Since *kRISP-meR* is a reference free tool for guide RNA design, it cannot directly determine the number of times a sequence, where a guide RNA may bind to, is present in the genome. Instead, it first counts the number of times

Figure 1. Overview of kRISP-meR. kRISP-meR takes as input a target sequence and sequenced reads from the organism. It starts by aligning the reads to the target sequence to detect sample specific variations. Then the candidate guide RNAs are identified. Next, scores to indicate efficiency and specificity of each candidate are determined using number of times k-mers appear in the sequences reads with the use of probabilistic models.

sequences of length 23 are present in the set of sequenced reads using Jellyfish 2 [23], and uses the counts to estimate the number of times they are present in the genome. The k -spectrum is generated by counting all possible k -mers in the sequencing reads and determining the number of i -mers in the reads $H(i)$ for all i . The first peak in the spectrum is due to large number of k -mers that arise due to sequencing errors and appear very few times. The next peak appears at the k -mer coverage λ .

2.4 Scoring candidate gRNAs

The cutting efficiency for each candidate gRNA is determined using the scheme suggested by Doench et al. [7] As in GuideScan [13], we use Rule Set 2 [7] which takes as input a sequence of length 30 at the target site and the 20-mer from the candidate gRNA and calculates cutting efficiency of the candidate gRNA using a boosted regression tree model.

Next, kRISP-meR assigns a score to each guide RNA indicating its off-target activity. First it calculates the total expected number of cuts, both on and off-target, for the gRNAs. It is known that a gRNA, when deployed in CRISPR/Cas9, in addition to making a cut where it finds a perfect matching sequence in the genome, may make cuts at sites with number of mismatches as well, albeit with lower probabilities. If the set of all k -mers within a specified Hamming distance (default is 2) of the complementary sequence of the gRNA is $\mathbb{K}$, we define the expected number of cuts in the genome as follows:

$$\text{Expected number of cuts, } E[x_G] = \sum_{k \in \mathbb{K}} c_{G,k} \times \psi_k$$

where ψ_k is the cutting probability of the gRNA at the k -mer site determined by a experimentally determined mutation matrix [7], and $c_{g,k}$ is the number of the times the k -mer is present in the genome. Since we do not know this in the absence of a reference, we take the expectation and estimate this from the number of times it appears in the read set

$c_{R,k}$ using the Bayes rule. Therefore,

$$\begin{aligned} E[x_G] &= \sum_{k \in \mathbb{K}} \left[\sum_{c_{G,k}=1}^{\infty} P(c_{G,k}|c_{R,k}) \times c_{G,k} \right] \times \psi_k \\ &= \sum_{k \in \mathbb{K}} \left[\sum_{c_{G,k}=1}^{\infty} \frac{P(c_{R,k}|c_{G,k})P(c_{G,k})}{P(c_{R,k})} \times c_{G,k} \right] \times \psi_k \end{aligned}$$

Here, $P(c_{G,k})$ is the prior probability that there are $c_{G,k}$ copies of k in the genome.

Now, the score assigned to a gRNA, which can be interpreted as the inverse of the specificity, is defined as follows:

$$s \approx \frac{E[x_G]}{E[x_t] \times c_{G,t}} \quad (1)$$

where $E[x_t]$ is the expected number of cuts within the target region, which can be calculated using the cutting probabilities and the numbers of times sequences within a given Hamming distance of the gRNA sequence appears in the target. $c_{G,t}$ is the number of times the target appears in the genome. This is calculated by constructing the histogram of k -mer counts in the sequencing reads and finding the k -mer coverage. Restricting within only the k -mers that appear within the target string generates a similar curve, but the first peak after the error component appears at $\lambda \times c_{G,t}$. Given that this peak appears at λ_t , the *target coverage*, we approximate $c_{G,t}$ as $\frac{\lambda_t}{\lambda}$.

2.5 Learning probability distributions and estimating priors

To calculate the scores using the equations mentioned above, the prior probability distribution on numbers of copies of k -mers in the genome and conditional probability distributions on k -mer counts in the reads given the copy numbers in the genome need to be defined. When a k -mer appears in the read set due to the presence of one or more copies of the sequence in the genome, Poisson distributions have been observed to model the counts well in genome sequencing data [24]. If a genomic region is present i times, then the counts of the k -mers within that region are assumed to be Poisson distributed with mean λi , where λ is the k -mer coverage of the dataset.

However, k -mers may also appear due to sequencing errors even though it is not present in the genome. This results in large number k -mers that appear one or few times. We observe that this is best approximated by the geometric distribution among the widely used distributions.

Since organisms differ in their repeat structure, we estimate priors on copy numbers in the genome from the data. First we use the expectation-maximization (EM) algorithm [25] to estimate the k -mer coverage in the sequencing dataset. We ignore the entries up to the first trough in the histogram as they are likely due to sequencing errors. The rest of the k -mer counts are modeled as a mixture of N Poisson distributions with means λi for $i = 1$ to N .

In the EM algorithm, the k -mer coverage λ is initialized using the histogram as mentioned in Section 2.3. Then the E-step and the M-step are repeated until convergence where in the E-step, membership probabilities are re-calculated and in the M-step, λ is updated. Once the k -mer coverage λ is estimated, we calculate, for all values of k , the probability that it was generated by each of the N Poisson distributions representing copy counts as well as the geometric distribution modeling the error component. The membership probabilities are then summed to obtain priors on copy numbers in the genome and the prior on the error component.

3 Results

To assess the performance of kRISP-meR, we first analyzed datasets for which both the reference genomes and high throughput sequencing reads are available. We used the Illumina sequencing reads of *Staphylococcus aureus* and Chromosome 14 of *Homo sapiens*, which are publicly available in the GAGE dataset [26].

We took random segments from the reference genomes to be used as the target sequences, and ran kRISP-meR on the targets and the reads. The identified gRNAs are then scored according to Equation 1 as described in Section 2. We also evaluate scores for the gRNAs using Equation 1 with the use of the reference genome. That is, instead of the k -mer counts from the reads, we use the exact number of times the k -mers are present in the reference genome. For all the guides identified, scored and listed by kRISP-meR, the scores are shown in Figure 2 along with the score calculated with the reference.

For the genome of *S. aureus* which has low repeat complexity, we find that majority of the gRNAs have scores close to 1.0 according to both the schemes. The scores calculated without the reference genome have greater spread compared to the ones obtained using the reference. This is expected because of the variation in sequencing depth across the genome caused by stochasticity and biases in the sequencing process as well as the presence of extraneous k -mers

Figure 2. Comparison of scores given by kRISP-meR and scores calculated using reference genome.

arising because of sequencing errors. However, we observe that almost half the gRNAs have scores close to 1.2 or less and almost three-fourths are below 1.5.

For the repeat rich Chromosome 14 of *H. sapiens*, there are many gRNAs with high scores indicating lack of specificity and highlighting the importance of properly scoring and avoiding these as guides. For this dataset, we observe that the scores computed using both approaches have similar median and inter-quantile ranges.

Next we go on to compare the guide RNAs listed by kRISP-meR with the gRNAs identified by a widely used gRNA designing tool, GuideScan [13], which requires reference genomes.

We used GuideScan on the two mentioned above and built genome-wide sgRNA library with a maximum of 2 mismatches. Next, we identified the sgRNAs from the library for the same target region using kRISP-meR. We only consider the guides with scores less than 3.0 for kRISP-meR, as kRISP-meR score all potential sgRNAs but the ones with scores greater than this are likely to have off-target effects. Figure 3a shows the extent of agreement between the sets of sgRNAs generated by kRISP-meR and GuideScan. It reveals that kRISP-meR is able to recognize the majority of the sgRNAs identified by GuideScan while reporting very few additional ones for both the datasets.

Figure 3. (a) Venn diagrams showing numbers of candidate guide RNAs found by only kRISP-meR, only GuideScan and both tools for randomly chosen regions in *S. aureus* genome and *H. sapiens* chromosome 14. (b) Numbers of guide RNAs predicted by kRISP-meR and GuideScan with off-target effects with no mismatches.

Finally, we assess potential off-target effects of the sgRNAs generated by kRISP-meR with scores below 3.0. For all the guides identified by GuideScan, and by kRISP-meR, we determine whether they match any site outside of the target region considering no mismatches i.e. have perfect matches. The results are shown in Figure 3b. We observe that for the *S. aureus* dataset, kRISP-meR generates one sgRNA with off-target effect whereas GuideScan has none, and for the human chromosome 14 dataset, kRISP-meR and GuideScan generate 1 and 3 sgRNAs with off-target effects.

4 Conclusion

In this paper, we presented a reference-free guide RNA designing tool for CRISPR/Cas9 system, namely **kRISP-meR**, which only needs sequenced reads from the organism. We experimented kRISP-meR with reads from two organisms, and compared the sgRNAs identified by kRISP-meR with those identified by GuideScan. Extensive analysis will follow but our results suggest, kRISP-meR identifies majority of the guides identified by GuideScan without any knowledge of the reference while keeping number of sgRNAs with off-target effects low. kRISP-meR can be used to design sgRNAs for organisms with no or incomplete reference genomes. In addition, even when a reference genome is available, it may be used as an auxiliary tool to ensure the candidate sgRNAs do not have off-target effects that may be missed by reference based tools due to presence of regions in the genome of the individual being edited which is missing in the reference.

References

- [1] Jennifer A Doudna and Emmanuelle Charpentier. The new frontier of genome engineering with CRISPR-Cas9. *Science*, 346(6213):1077–1087, 2014.
- [2] Patrick D Hsu, Eric S Lander, and Feng Zhang. Development and applications of CRISPR-Cas9 for genome engineering. *Cell*, 157(6):1262–1278, 2014.
- [3] Philippos Aris Papathanos and Nikolai Windbichler. Redkmer: An assembly-free pipeline for the identification of abundant and specific x-chromosome target sequences for x-shredding by crispr endonucleases. *The CRISPR journal*, 1(1):88–98, 2018.
- [4] Carrie M Margulies, Frank Buquicchio, Giulia Schiroli, Valentina Vavassori, Jennifer Gori, Kiran Gogi, Fred Harbinski, Luigi Naldini, Pietro Genovese, Charles Albright, et al. Efficient targeted integration in human t cells with crispr-cas9 for the treatment of x-linked hyper-igm syndrome. In *Molecular Therapy*, volume 26, pages 237–237, 2018.
- [5] Rafal Kaminski, Yilan Chen, Tracy Fischer, Ellen Tedaldi, Alessandro Napoli, Yonggang Zhang, Jonathan Karn, Wenhui Hu, and Kamel Khalili. Elimination of HIV-1 genomes from human T-lymphoid cells by CRISPR/Cas9 gene editing. *Scientific Reports*, 6, no. 22555:1–15, 2016.
- [6] Ankur Jain, Gulab Zode, Ramesh B Kasetti, Fei A Ran, Winston Yan, Tasneem P Sharma, Kevin Bugge, Charles C Searby, John H Fingert, Feng Zhang, et al. CRISPR-Cas9–based treatment of myocilin-associated glaucoma. *Proceedings of the National Academy of Sciences*, 114(42):11199–11204, 2017.
- [7] John G Doench, Nicolo Fusi, Meagan Sullender, Mudra Hegde, Emma W Vaimberg, Katherine F Donovan, Ian Smith, Zuzana Tothova, Craig Wilen, Robert Orchard, et al. Optimized sgRNA design to maximize activity and minimize off-target effects of crispr-cas9. *Nature biotechnology*, 34(2):184, 2016.
- [8] Benjamin P Kleinstiver, Vikram Pattanayak, Michelle S Prew, Shengdar Q Tsai, Nhu T Nguyen, Zongli Zheng, and J Keith Joung. High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects. *Nature*, 529(7587):490–506, 2016.
- [9] Hannah R Kempton and Lei S Qi. When genome editing goes off-target. *Science*, 364(6437):234–236, 2019.
- [10] Florian Heigwer, Grainne Kerr, and Michael Boutros. E-crisp: fast crispr target site identification. *Nature methods*, 11(2):122–123, 2014.
- [11] Miguel A Moreno-Mateos, Charles E Vejnar, Jean-Denis Beaudoin, Juan P Fernandez, Emily K Mis, Mustafa K Khokha, and Antonio J Giraldez. Crisprscan: designing highly efficient sgRNAs for crispr/cas9 targeting in vivo. *Nature methods*, 12(10):982, 2015.
- [12] Maximilian Haeussler, Kai Schöning, Hélène Eckert, Alexis Eschstruth, Joffrey Mianné, Jean-Baptiste Renaud, Sylvie Schneider-Maunoury, Alena Shkumatava, Lydia Teboul, Jim Kent, et al. Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool crispor. *Genome biology*, 17(1):148, 2016.
- [13] Alexendar R Perez, Yuri Pritykin, Joana A Vidigal, Sagar Chhangawala, Lee Zamparo, Christina S Leslie, and Andrea Ventura. Guidescan software for improved single and paired crispr guide RNA design. *Nature Biotechnology*, 35(4):347–349, 2017.
- [14] Tao Li, Shaokai Wang, Feng Luo, Fang-Xiang Wu, and Jianxin Wang. Multiguidescan: a multi-processing tool for designing crispr guide RNA libraries. *Bioinformatics*, 2019.
- [15] Mitchell R Vollger, Philip C Dishuck, Melanie Sorensen, AnneMarie E Welch, Vy Dang, Max L Dougherty, Tina A Graves-Lindsay, Richard K Wilson, Mark JP Chaisson, and Evan E Eichler. Long-read sequence and assembly of segmental duplications. *Nature methods*, 16(1):88, 2019.
- [16] Kathleen C Keough, Svetlana Lyalina, Michael P Olvera, Sean Whalen, Bruce R Conklin, and Katherine S Pollard. Alleleanalyzer: a tool for personalized and allele-specific sgRNA design. *Genome biology*, 20(1):1–9, 2019.
- [17] Adrien LS Jacquin, Duncan T Odom, and Margus Lukk. Crisflash: open-source software to generate crispr guide RNAs against genomes annotated with individual variation. *Bioinformatics*, 2019.

-
- [18] Samuele Cancellieri, Matthew C Canver, Nicola Bombieri, Rosalba Giugno, and Luca Pinello. Crispritz: rapid, high-throughput, and variant-aware in silico off-target site identification for crispr genome editing. *Bioinformatics*, 2019.
- [19] Jiamin Sun, Hao Liu, Jianxiao Liu, Shikun Cheng, Yong Peng, Qinghua Zhang, Jianbing Yan, Hai-Jun Liu, and Ling-Ling Chen. Crispr-local: a local single-guide rna (sgRNA) design tool for non-reference plant genomes. *Bioinformatics*, 2018.
- [20] Andrew Hammond, Roberto Galizi, Kyros Kyrou, Alekos Simoni, Carla Siniscalchi, Dimitris Katsanos, Matthew Gribble, Dean Baker, Eric Marois, Steven Russell, et al. A crispr-cas9 gene drive system targeting female reproduction in the malaria mosquito vector *Anopheles gambiae*. *Nature biotechnology*, 34(1):78, 2016.
- [21] Ben Langmead and Steven L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9:357–359, March 2012. ISSN 1548-7091. doi: 10.1038/nmeth.1923. URL <http://dx.doi.org/10.1038/nmeth.1923>.
- [22] Bruce J Walker, Thomas Abeel, Terrance Shea, Margaret Priest, Amr Abouelliel, Sharadha Sakthikumar, Christina A Cuomo, Qiandong Zeng, Jennifer Wortman, Sarah K Young, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS one*, 9(11):e112963, 2014.
- [23] Guillaume Marçais and Carl Kingsford. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27(6):764–770, 2011. doi: 10.1093/bioinformatics/btr011.
- [24] Atif Rahman, Ingileif Hallgrímsdóttir, Michael B Eisen, and Lior Pachter. Association mapping from sequencing reads using k-mers. *bioRxiv*, page 141267, 2017.
- [25] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.
- [26] Steven L Salzberg, Adam M Phillippy, Aleksey Zimin, Daniela Puiu, Tanja Magoc, Sergey Koren, Todd J Treangen, Michael C Schatz, Arthur L Delcher, Michael Roberts, et al. Gage: A critical evaluation of genome assemblies and assembly algorithms. *Genome research*, 22(3):557–567, 2012.