Bioinformatics, 2018, 1–3
doi: 10.1093/bioinformatics/bty970
Advance Access Publication Date: 30 November 2018
Applications Note

Sequence analysis

CRISPR-Local: a local single-guide RNA (sgRNA) design tool for non-reference plant genomes

Jiamin Sun[1,†], Hao Liu[1,2,†], Jianxiao Liu[1,2], Shikun Cheng[2], Yong Peng[1], Qinghua Zhang[1], Jianbing Yan[1], Hai-Jun Liu[1,*] and Ling-Ling Chen[1,2,*]

[1]National Key Laboratory of Crop Genetic Improvement and [2]Hubei Key Laboratory of Agricultural Bioinformatics, College of Informatics, Huazhong Agricultural University, Wuhan 430070, People's Republic of China

*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.

Associate Editor: Inanc Birol

Received on July 10, 2018; revised on November 4, 2018; editorial decision on November 23, 2018; accepted on November 29, 2018

Abstract

Summary: CRISPR-Local is a high-throughput local tool for designing single-guide RNAs (sgRNAs) in plants and other organisms that factors in genetic variation and is optimized to generate genome-wide sgRNAs. CRISPR-Local outperforms other sgRNA design tools in the following respects: (i) designing sgRNAs suitable for non-reference varieties; (ii) screening for sgRNAs that are capable of simultaneously targeting multiple genes; (iii) saving computational resources by avoiding repeated calculations from multiple submissions and (iv) running offline, with both command-line and graphical user interface modes and the ability to export multiple formats for further batch analysis or visualization. We have applied CRISPR-Local to 71 public plant genomes, using both CRISPR/Cas9 and CRISPR/cpf1 systems.

Availability and implementation: CRISPR-Local can be freely downloaded from http://crispr.hzau.edu.cn/CRISPR-Local/.

Contact: heroalone@webmail.hzau.edu.cn or llchen@mail.hzau.edu.cn

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

1 Introduction

The CRISPR-derived editing system has been widely used for genome editing and has recently reached a high-throughput level with the construction of a genome-wide mutant library and the development of large-scale genetic screening (Lu *et al.*, 2017; Meng *et al.*, 2017; Shalem *et al.*, 2014; Wang *et al.*, 2014). For genome editing experiments, it is critical to design a reliable single-guide RNA (sgRNA). However, most existing tools do not consider variations in genetic background. Taking this variation into account is especially important for plant genome editing studies, since the transformed lines are rarely the sequenced reference genomes. Therefore, mismatches between transformed and reference genomes are very common. For example, the maize (*Zea mays*) genome has a minimum of one single nucleotide polymorphism (SNP) per 44 bp, and ~30% of the entire genome sequence in a maize population may not be captured by the B73 reference genome (Gore *et al.*, 2009). These factors could potentially make sgRNAs designed from the reference genome unusable. Here, we introduce a local genotype-specific genome-wide sgRNA design tool, CRISPR-Local, that integrates the reference genome with sequencing data from specific transgenic receptor lines.

2 Materials and methods

The major concept of the CRISPR-Local tool can be divided into two parts: the first is the 'one-for-all' strategy to build a whole-genome sgRNA database with high efficiency, from which all possible sgRNAs for a given reference or user-defined genome are generated and stored on a local computer. The other is to retrieve (or *de novo* design) candidate sgRNAs by integrating data from whole-genome sequencing, mRNA sequencing or known variants for specific transgenic receptor lines.

The detailed implementation features of CRISPR-Local (Supplementary Fig. S1) are as follows:

i. Reference sgRNAs database (RD)-build model: to build the whole-genome sgRNA database from a reference (or user-defined) genome and corresponding annotation files. This function is the same as used in other sgRNA-design tools but at a genome-wide level: (a) Screening all possible on-target sgRNAs for CRISPR/Cas9 and CRISPR/cpf1 systems and scoring them with the Rule set 2 algorithm (Doench *et al.*, 2016) for CRISPR/Cas9 and the method developed by Kim *et al.* (2017) for CRISPR/cpf1; (b) Using SeqMap program to predict the effects of each off-target site with the highest cutting frequency determination (CFD) score for each sgRNA (Jiang and Wong, 2008). All target and off-target data determined across the entire genome are exported into RD.

   RD is the basis for further analysis. It contains information for each sgRNA for every gene, including sequence, physical position, the relative position against transcription start site, on-target score and potential off-target sites with the highest CFD score. During this step, non-exonic regions can also be handled if the user supplies a genomic annotation and a set of options can be established for a specific purpose. For example, by default, sgRNAs are confined to exons; however, users can set a length value (such as 3 bp) that can extend 3 bp 5' or 3' of the exon boundary to obtain more sgRNA candidates, while the core editing region remains in the exon.

ii. User's sgRNA database (UD)-build model: to build a non-reference sgRNA database that is suitable for a specific genetic background. Transgenic receptor lines in plants are usually not the reference genome, which means that sgRNAs designed according to the reference genome may not work well. This optional step is designed to address this concern by allowing users to provide sequencing data for transgenic receptors. Next-generation sequencing reads are first mapped to the reference genome. Each aligned read can be assigned to an annotated gene based on the alignment, specific sequences can be obtained for non-reference lines and the aligned reads can then be used to screen for suitable sgRNAs. Those sgRNAs from completely conserved regions are restored using default parameters. CRISPR-Local can also apply a *de novo* process to design non-reference-specific sgRNAs based on updated sequences by integrating sequence variations. The UD is thus updated by filtering and scoring all candidate sgRNAs, and the final scheme provided in UD is the same as in RD.

iii. Database (DB)-search model: to output results based on the user's gene list. The script in this step can be applied to quickly search the RD/UD to obtain sorted results from all genes, or any gene list provided, under several filtering parameters. The results consist of three parts: The sgRNAs present exclusively in RD (RD only, RO) or UD (UD only, UO) or both (BO). Generally, BO is preferred. Caution should be exercised when using UO, but RO is not recommended when the supplied sequencing data are sufficient. The results can be visualized in the exported text files.

iv. Paralogs (PL)-search model: to output sgRNAs targeting multiple genes. This step is designed for editing multiple PL, but could also be applied to any gene sets of interest, such as for knocking out multiple key genes in the same pathway or related to the same phenotype. This step follows the function of the above DB-search model that assigns targets to RO, UO or BO. Moreover, PL-search will label the sgRNAs as common (matches all candidates) or exclusive (individually matched) according to the list of submitted genes.

3 Results

We have provided a detailed user manual for CRISPR-Local (http://crispr.hzau.edu.cn/CRISPR-Local/help.php), including the requirements for additional software and libraries, the explanation for each parameter and a step-by-step demonstration. CRISPR-Local has both command-line (programmed in perl and python) and graphical user interface (GUI) modes (implemented in Java) provided for diverse users, and can export multiple formats (such as txt and html) for further batch analysis or visualization (Supplementary Fig. S2). Using the maize genome (AGPv4, 2.1GB; Jiao *et al.*, 2017) as an example to demonstrate the general performance of CRISPR-Local, the database-building process was completed in 65 h using 15 CPUs for the CRISPR/Cas9 system and the searching and outputting process for 10 000 randomly selected genes was finished in less than 30 s. Randomly selecting 10 genes from the maize genome (AGPv4), the sgRNAs obtained from CRISPR-Local were all confirmed by existing tools, including CRISPR-P 2.0 (Liu *et al.*, 2017), CHOPCHOP v2 (Labun *et al.*, 2016) and Breaking-Cas (Oliveros *et al.*, 2016) (Supplementary Fig. S3). The obtained sgRNAs were quite similar, and the ranking between them showed significantly ($P < 3E$–102) positive correlations, even though the designing tools have different scoring schemes.

To demonstrate the significance and application of CRISPR-Local, samples from seven tissues (seedling leaf, mature leaf, young stem apex, young tassel, young ear, developing embryo and developing endosperm) of an inbred maize line called ZZC01 were combined for RNA-sequencing with the accession number CRA000750 for raw data from BIG Data Center (BIG Data Center Members, 2017). ZZC01 is a tropical line with significant difference from the B73 reference genome (Supplementary Fig. S4) and has been widely used in maize transformation experiments with high efficiency. CRISPR-Local was performed under default parameters, resulting in approximately 7.37 and 6.98 million sgRNAs against 46 304 genes for B73 and ZZC01, respectively. Among them, 4.70 million (63.8%) sgRNAs derived from the B73 reference genome corresponding to 41 337 (89.3%) genes were shared with ZZC01, while 2.67 million (36.2%) sgRNAs from 45 045 (97.3%) genes were not matched to ZZC01 (Fig. 1A). Furthermore, 185 183 SNPs or small InDels were called between B73 and ZZC01 across 589 378 B73-specific sgRNAs. The presence of substantial variation indicates that these sgRNAs will not be suitable, since a large portion of variants are located close to the protospacer-adjacent motif (PAM), the region that must be matched for effective editing. Even for the 4.70 million shared sgRNAs, 31 615 (matched to 11 657 genes) are uniquely identified in B73 but have multiple fully matched loci in ZZC01 (non-paralog, with the same sgRNA but different flanking sequences), indicating that these sgRNAs are potential off-targets in ZZC01. Together, these data indicate that over one-third of sgRNAs, covering ~10% of genes, would likely fail in editing experiments due to the genetic variation between two maize lines and that even shared sgRNAs would create off-target effects in another line.

4 Conclusion

In summary, CRISPR-Local is a high-throughput and user-friendly local tool for designing sgRNAs in plants and other species, which is optimized for providing high-throughput sgRNAs and considers the genetic variations among individuals of different genetic

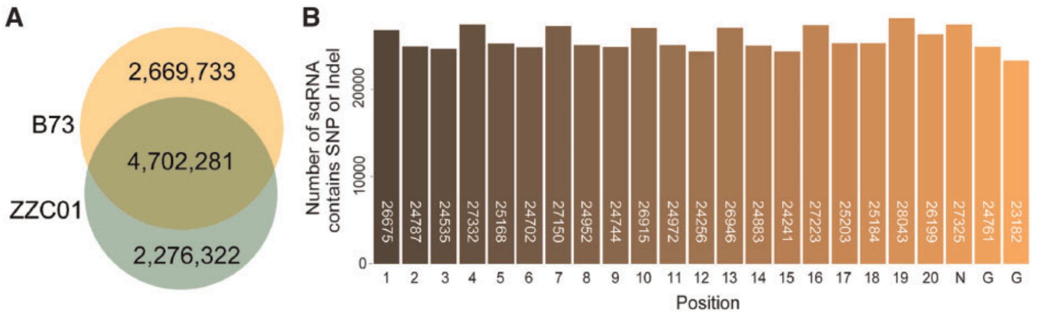

Fig. 1. Discrepancy of sgRNAs identified in the maize reference genome (B73) and transgenic line (ZZC01). (**A**) Overlap of sgRNAs between B73 and ZZC01. (**B**) The distribution of variants between B73 and ZZC01 along B73-specific sgRNAs

backgrounds. CRISPR-Local outperforms previous sgRNA design tools in the following respects: (i) designing appropriate sgRNAs for specific non-reference varieties; (ii) screening sgRNAs that simultaneously target multiple genes (usually PL); (iii) saving computational resources by avoiding repeated calculations from multiple submissions and (iv) running offline, with both command-line and GUI-modes for diverse users and the ability to export multiple formats for further batch analysis or visualization. Once the database is built, the search process for CRISPR-Local is simple and fast. Therefore, CRISPR-Local is a 'one-for-all' strategy that relieves server computing pressure and saves time for users.

Although CRISPR-Local is designed for batch studies, it is also suitable for daily jobs involving only a few loci. We have applied CRISPR-Local to 71 public plant genomes. Most of these genomes are updated from CRISPR-P 2.0 (Liu *et al.*, 2017), and have been analyzed for both CRISPR/Cas9 (with NGG PAM) and CRISPR/cpf1 [with TTV (V represents A, G or C) and TTTV PAM], which can be directly downloaded from http://crispr.hzau.edu.cn/cgi-bin/CRISPR-Local/download. In the updating version, CRISPR-Local will handle more PAMs and adapt to more subsequent analysis (e.g. to determine edited variants in a batch manner) in facing to the era of high-throughput genome editing studies.

Funding

This work was supported by the National Key Research and Development Program of China [2016YFD0101003 to J.Y.], the Fundamental Research Funds for the Central Universities (2662017PY043), the National Natural Science Foundation of China [31571351 to L.-L.C.] and the Post-doctoral Talent Innovation Program [BX201700092 to H.-J.L].

Conflict of Interest: none declared.

References

BIG Data Center Members. (2017) The BIG Data Center: from deposition to integration to translation. *Nucleic Acids Res*., **45**, D18–D24.

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.

Gore,M.A. *et al*. (2009) A first-generation haplotype map of maize. *Science*, **326**, 1115–1117.

Jiang,H. and Wong,W.H. (2008) SeqMap: mapping massive amount of oligonucleotides to the genome. *Bioinformatics*, **24**, 2395–2396.

Jiao,Y. *et al*. (2017) Improved maize reference genome with single-molecule technologies. *Nature*, **546**, 524–527.

Kim,H.K. *et al*. (2017) In vivo high-throughput profiling of CRISPR-Cpf1 activity. *Nat. Methods*, **14**, 153–159.

Labun,K. *et al*. (2016) CHOPCHOP v2: a web tool for the next generation of CRISPR genome engineering. *Nucleic Acids Res*., **44**, W272–W276.

Liu,H. *et al*. (2017) CRISPR-P 2.0: an improved CRISPR-Cas9 tool for genome editing in plants. *Mol. Plant*, **10**, 530–532.

Lu,Y. *et al*. (2017) Genome-wide targeted mutagenesis in rice using the CRISPR/Cas9 system. *Mol. Plant*, **10**, 1242–1245.

Meng,X. *et al*. (2017) Construction of a genome-wide mutant library in rice using CRISPR/Cas9. *Mol. Plant*, **10**, 1238–1241.

Oliveros,J.C. *et al*. (2016) Breaking-Cas—interactive design of guide RNAs for CRISPR-Cas experiments for ENSEMBL genomes. *Nucleic Acids Res*., **44**, W267–W271.

Shalem,O. *et al*. (2014) Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science*, **343**, 84–87.

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