Molecular Plant Advance Access published February 19, 2014

Genome-Wide Prediction of Highly Specific Guide RNA Spacers for the CRISPR–Cas9-Mediated Genome Editing in Model Plants and Major Crops

Dear Editor,

RNA-guided genome editing (RGE) using the *Streptococcus pyogenes* CRISPR–Cas9 system (Jinek et al., 2012; Cong et al., 2013; Mali et al., 2013b) is emerging as a simple and highly efficient tool for genome editing in many organisms. The Cas9 nuclease can be programmed by dual or single guide RNA (gRNA) to cut target DNA at specific sites, thereby introducing precise mutations by error-prone non-homologous-end-joining repairing or by incorporating foreign DNAs via homologous recombination between target site and donor DNA. The gRNA–Cas9 complex recognizes targets based on the complementarity between one strand of targeted DNA (referred as protospacer) and the 5′-end leading sequence of gRNA (referred to as gRNA spacer) that is approximately 20 base pairs (bp) long (Figure 1A). Besides gRNA–DNA pairing, a protospacer-adjacent motif (PAM) following the paired region in the DNA is also required for Cas9 cleavage. Recent studies reveal that Cas9 could cut the PAM-containing DNA sites that imperfectly match gRNA spacer sequences, resulting in genome editing at undesired positions. This off-target editing of engineered gRNA–Cas9 has been extensively examined recently (Hsu et al., 2013; Mali et al., 2013a). Thus, gRNA–Cas9 specificity becomes a major concern for RGE application, and it is very important to evaluate the potential constraint of Cas9 specificity and develop straightforward bioinformatic tools to facilitate the design of highly specific gRNAs to minimize off-target effects.

Nucleotide mismatch between a gRNA spacer sequence and a PAM-containing genomic sequence was shown to significantly reduce the Cas9 affinity at the target site (Hsu et al., 2013; Mali et al., 2013a; Pattanayak et al., 2013). Cas9 generally tolerates no more than three mismatches in the gRNA–DNA paired region and the presence of mismatches adjacent to PAM would greatly reduce Cas9 affinity to the site imperfectly matching the gRNA. Thus, the off-target risk of a designed gRNA could be assessed by similarity searching against whole-genome sequence *in silico*; and, vice versa, genome-wide sequence analysis could be used to predict gRNA spacer with high specificity for RGE in designated specie. For plants, especially crops whose genome sizes range from ~1×10^8 to 2×10^9 bp with different levels of sequence complexity and duplication, genome-wide prediction of specific gRNAs would help evaluate the potential constraint for Cas9 off-target effects and greatly facilitate the application of the RGE technology in plant functional genomics and genetic improvement of agricultural crops. To this end, we analyzed the assembled nuclear genome sequences of eight representative plant species (Supplemental Table 1), including *Arabidopsis thaliana*, *Medicago truncatula*, *Glycine max* (soybean), *Solanum lycopersicum* (tomato), *Brachypodium distachyon*, *Oryza sativa* (rice), *Sorghum bicolor*, and *Zea mays* (maize) to predict specific gRNA spacers which are expected to have little or no off-target risk in RGE.

The choice of gRNA spacer sequences is limited to locations with PAMs in the genome. The gRNA–Cas9 complex recognizes two PAMs, 5′-NGG-3′ and 5′-NAG-3′, but shows much less affinity and less tolerance of mismatches at the NAG–PAM site (Hsu et al., 2013). Thus, we only predicted specific gRNA spacers targeting NGG–PAM sites. Potential gRNA spacer sequences (20 nt long) were extracted from the genomic sequences before NGG–PAM (GG-spacer). We also extracted the 20-nt sequences before NAG–PAM (AG-spacer) but only used them in off-target assessment. The off-target risk of a gRNA spacer is dependent on its similarity to all GG-spacers and AG-spacers. After the pair-wise sequence comparison, we have taken two steps to classify these GG-spacer sequences according to their off-target potential (Figure 1B; see details in Supplemental Method, Supplemental Figure l, and Supplemental Table 2). First, each GG-spacer was sorted to Class0 (no significant sequence similarity with other GG-spacers), Class1 (four or more mismatches, or three mismatches adjacent to PAM in all GG-spacer alignments), or Class2 (fewer than three mismatches, or three mismatches distant to PAM in all GG-spacer alignments). A Class2 candidate is considered to have off-target possibilities because it shares significant sequence identity with other GG-spacers

doi:10.1093/mp/ssu009

Received 19 January 2014; accepted 21 January 2014

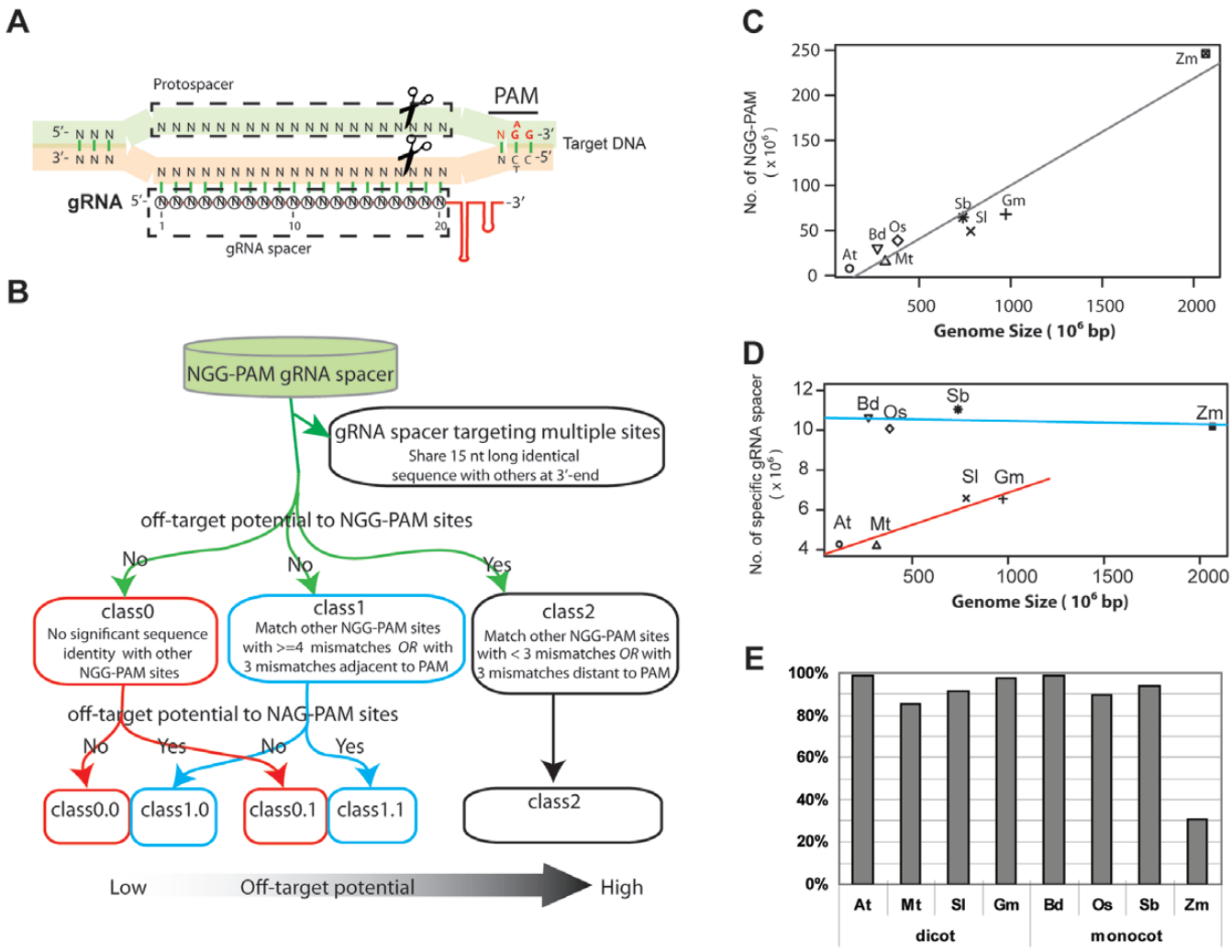

Figure 1. Genome-Wide Prediction of Specific gRNA Spacers and Assessment of Off-Target Constraints for CRISPR–Cas9 in Eight Plant Species.

(A) Diagrammatic illustration of targeted DNA cleavage by gRNA–Cas9. A gRNA consists of a 5′-end spacer sequence paired to target DNA protospacer and the conserved scaffold (red lines). PAM, protospacer-adjacent motif.

(B) A simplified scheme for genome-wide prediction of specific gRNA spacers (see Supplemental Methods and Supplemental Figure 1 for details). Class0.0 and Class1.0 gRNA spacers are considered specific for RGE.

(C) Positive correlation between genome size and NGG–PAM number in eight plant species.

(D) Positive correlation between genome size and the number of specific gRNA spacers was found in eudicots but not in monocots of the grass family. The linear regressed trend line is shown in red for eudicots and blue for monocots.

(E) Percentage of annotated transcript units that could be targeted by specific gRNAs. Eudicots: At, *Arabidopsis thaliana*; Mt, *Medicago truncatula*; Sl, *Solanum lycopersicum*; Gm, *Glycine max*. Monocots: Bd, *Brachypodium distachyon*; Os, *Oryza sativa*; Sb, *Sorghum bicolor*; Zm, *Zea mays*.

and contains fewer mismatches. Second, GG-spacers from Class0 and Class1 were further classified to subclasses after comparing with all AG-spacers. Class0.0 and Class1.0 spacers are expected to be highly specific whereas Class0.1 and Class1.1 may cause off-target effects on other NAG–PAM sites. A GG-spacer may have off-target effects on other NAG-sites if it matches other AG-spacers with fewer than three mutations. These criteria were selected based on the recent reports regarding the gRNA specificity and off-target analyses in animals (Hsu et al., 2013; Mali et al., 2013a; Pattanayak et al., 2013) and observations in plants (Li et al., 2013; Nekrasov et al., 2013; Shan et al., 2013; Xie and Yang, 2013). As a result, Class0.0 and Class1.0 gRNA spacers are expected to provide high specificity in the CRISPR–Cas9-mediated genome editing, with class0.0 gRNA spacers being the most specific.

Among these eight plant species, 5–12 NGG–PAMs were identified every 100 bp in chromosomes (Supplemental Table 2), and the total number of NGG–PAMs is positively correlated to genome size (correlation coefficient R = 0.97, Figure 1C). The total number of specific gRNA spacers (Class0.0 and 1.0) ranges from 4 to 11 million, and more specific gRNAs were predicted in monocots (*Brachypodium*, rice, *Sorghum*,

and maize) than in eudicots (*Arabidopsis*, *Medicago*, tomato, and soybean) despite their genome size. The number of specific gRNA spacers is positively correlated to genome size ($R = 0.95$) in four eudicot species (Figure 1D). In four monocot species, however, the number of specific gRNA spacers is not proportional to the genome size ($R = –0.30$, Figure 1D), nor to the total transcript number ($R = –0.67$) or the NGG–PAM number ($R = –0.37$). Comparable numbers of specific gRNA spacers ($10 – 11 \times 10^6$) were found in four monocot species despite the significant difference (two to eight-fold) in their genome sizes (Figure 1D and Supplemental Table 2). Although the 20-nt-long gRNA spacer sequences have more chance to be aligned with other PAM sites with fewer mismatches in bigger genomes, the number of specific gRNA spacers also depends on the genome sequence content.

We calculated the proportion of annotated genes that could be targeted by specific gRNAs designed from Class0.0 and Class1.0 spacer sequences. Based on the current genome annotation for seven of the eight plant species, specific gRNAs could be designed to target 85.4%–98.9% of annotated transcript units (TU), and 83.4%–98.6% of TUs could be targeted in exons (Figure 1E and Supplemental Table 3). The exception, maize, has the largest genome and the largest number of annotated TUs among these eight species, but only 30% of maize TUs are targetable by the specific gRNA (Supplemental Table 3). For the other seven plant species, 67.9%–96.0% of TUs have at least 10 NGG–PAM sites that could be targeted by specific gRNAs containing Class0.0 or Class1.0 spacers (Supplemental Figure 2). Thus, we predict that the off-target effect of CRISPR–Cas9 could be minimized and will not constrain genome editing in *Arabidopsis*, *Medicago*, tomato, soybean, rice, *Sorghum*, and *Brachypodium*. However, further bioinformatic analysis and careful prediction of specific gRNA spacers are required to avoid off-target effects for the CRISPR–Cas9-mediated genome editing in maize.

We have examined the feasibility of specifically targeting the nucleotide-binding site leucine-rich repeat (NBS–LRR) genes, which comprise one of the largest plant gene families and evolve rapidly to mediate host resistance against pathogen infection. The number of predicted NBS–LRR genes varies from 112 to 502 in these eight species (Supplemental Table 4). Specific gRNAs could be designed to target almost all NBS–LRR genes in *Arabidopsis*, soybean, rice, tomato, *Brachypodium*, and *Sorghum*. However, specific gRNAs are not available to target 41 (8.7%) and 40 (33.9%) of the NBS–LRR genes in *Medicago* and maize, respectively (Supplemental Table 4). We reasoned that those NBS–LRR genes share a high level of sequence identity to other genomic sites because of their gene duplication and diversification history.

The genome-wide prediction of specific gRNA spacers suggests that the off-target effect is unlikely to constrain RGE in most model plants and major crops, except maize. Besides maize, wheat and barley, which are important cereal crops with larger genome than maize, may also present a similar challenge for the CRISPR–Cas9-mediated RGE specificity. Considering the functional redundancy of some homologous genes with high sequence identity, specific gRNAs could be designed using spacer sequences other than Class0.0 or 1.0 to target duplicated genes without causing off-target effects to other transcripts. It was reported that Cas9 specificity was increased with a lower gRNA–Cas9 concentration (Hsu et al., 2013; Mali et al., 2013a; Pattanayak et al., 2013). Therefore, more gRNA spacer sequences, like some Class2 spacers, could be considered for specific RGE in practice. Alternative approaches such as the use of paired gRNAs and nickase mutation of Cas9 for reducing off-target risk (Mali et al., 2013a) or use of Cas9 orthologs recognizing different PAM may also help to increase specifically targetable sites, especially for maize. We have established the CRISPR-PLANT Database (www.genome.arizona.edu/crispr; Supplemental Figure 3) to enable the plant research community to access genome-wide predictions of specific gRNAs, and facilitate the application of CRISPR–Cas9-mediated genome editing in model plants and major agricultural crops.

SUPPLEMENTARY DATA

Supplementary Data are available at *Molecular Plant Online.*

FUNDING

This research was supported by the NSF Plant Genome Research Program (DBI-0922747) and the Pennsylvania State University.

ACKNOWLODGMENTS

We appreciate the Research Computing and Cyberinfrastructure at the Pennsylvania State University for providing the high-performance computing systems. We would like to thank Prof. Claude dePamphilis at the Pennsylvania State University for reviewing the manuscript and providing helpful comments. No conflict of interest declared.

Kabin Xie[a], *Jianwei Zhang*[b,c], and *Yinong Yang*[a,d,1]

a Department of Plant Pathology and Environmental Microbiology, The Huck Institutes of the Life Sciences, Pennsylvania State University, University Park, PA 16802, USA
b National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan 430070, China
c Arizona Genomics Institute, University of Arizona, Tucson, AZ 85721, USA
d Institute of Genetics, College of Life Sciences, Zhejiang University, Hangzhou 310058, China
[1] To whom correspondence should be addressed. E-mail yuy3@psu.edu, fax +1(814) 863–7217, tel. +1(814) 867-0324

REFERENCES

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., and Zhang, F. (2013). Multiplex genome engineering using CRISPR/Cas systems. Science. **339**, 819–823.

Hsu, P.D., Scott, D.A., Weinstein, J.A., Ran, F.A., Konermann, S., Agarwala, V., Li, Y., Fine, E.J., Wu, X., and Shalem, O. (2013). DNA targeting specificity of RNA-guided Cas9 nucleases. Nat. Biotechnol. **31**, 827–832.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., and Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science. **337**, 816–821.

Li, J.F., Norville, J.E., Aach, J., McCormack, M., Zhang, D., Bush, J., Church, G.M., and Sheen, J. (2013). Multiplex and homologous recombination-mediated genome editing in *Arabidopsis* and *Nicotiana benthamiana* using guide RNA and Cas9. Nat. Biotechnol. **31**, 688–691.

Mali, P., Aach, J., Stranges, P.B., Esvelt, K.M., Moosburner, M., Kosuri, S., Yang, L., and Church, G.M. (2013a). CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. Nat. Biotechnol. **31**, 833–838.

Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E., and Church, G.M. (2013b). RNA-guided human genome engineering via Cas9. Science. **339**, 823–826.

Nekrasov, V., Staskawicz, B., Weigel, D., Jones, J.D., and Kamoun, S. (2013). Targeted mutagenesis in the model plant *Nicotiana benthamiana* using Cas9 RNA-guided endonuclease. Nat. Biotechnol. **31**, 691–693.

Pattanayak, V., Lin, S., Guilinger, J.P., Ma, E., Doudna, J.A., and Liu, D.R. (2013). High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. Nat. Biotechnol. **31**, 839–843.

Shan, Q., Wang, Y., Li, J., Zhang, Y., Chen, K., Liang, Z., Zhang, K., Liu, J., Xi, J.J., Qiu, J.L., and Gao, C. (2013). Targeted genome modification of crop plants using a CRISPR–Cas system. Nat. Biotechnol. **31**, 686–688.

Xie, K., and Yang, Y. (2013). RNA-guided genome editing in plants using a CRISPR–Cas system. Mol. Plant. **6**, 1975–1983.