

# SCIENTIFIC REPORTS



OPEN

## Targeted mutagenesis in soybean using the CRISPR-Cas9 system

Xianjun Sun<sup>1,2,\*</sup>, Zheng Hu<sup>2,\*</sup>, Rui Chen<sup>3</sup>, Qiyang Jiang<sup>2</sup>, Guohua Song<sup>2</sup>, Hui Zhang<sup>2</sup> & Yajun Xi<sup>2,\*</sup>

Received: 15 December 2014

Accepted: 10 April 2015

Published: 29 May 2015

Genome editing is a valuable technique for gene function analysis and crop improvement. Over the past two years, the CRISPR-Cas9 system has emerged as a powerful tool for precisely targeted gene editing. In this study, we predicted 11 *U6* genes in soybean (*Glycine max* L.). We then constructed two vectors (pCas9-GmU6-sgRNA and pCas9-AtU6-sgRNA) using the soybean *U6-10* and Arabidopsis *U6-26* promoters, respectively, to produce synthetic guide RNAs (sgRNAs) for targeted gene mutagenesis. Three genes, *Glyma06g14180*, *Glyma08g02290* and *Glyma12g37050*, were selected as targets. Mutations of these three genes were detected in soybean protoplasts. The vectors were then transformed into soybean hairy roots by *Agrobacterium rhizogenes* infection, resulting in efficient target gene editing. Mutation efficiencies ranged from 3.2–9.7% using the pCas9-AtU6-sgRNA vector and 14.7–20.2% with the pCas9-GmU6-sgRNA vector. Biallelic mutations in *Glyma06g14180* and *Glyma08g02290* were detected in transgenic hairy roots. Off-target activities associated with *Glyma06g14180* and *Glyma12g37050* were also detected. Off-target activity would improve mutation efficiency for the construction of a saturated gene mutation library in soybean. Targeted mutagenesis using the CRISPR-Cas9 system should advance soybean functional genomic research, especially that of genes involved in the roots and nodules.

Genome editing is an important tool for gene function analysis, gene therapy and crop improvement. In recent years, three genome editing techniques have been developed that rely on zinc-finger nucleases (ZFNs), transcription activator-like nucleases (TALENs) and clustered regularly interspaced short palindromic repeat (CRISPR)/CRISPR-associated protein (Cas) system<sup>1</sup>. All three methods induce double-stranded breaks (DSBs) in the target genome DNA, which are subsequently repaired through non-homologous end joining (NHEJ) and homologous recombination (HR)<sup>2</sup>. ZFNs and TALENs target the genome through protein-DNA interactions, whereas genomic DNA editing by the CRISPR-Cas system is based on short RNA-DNA base pairing<sup>1,3,4</sup>. Compared with ZFNs and TALENs, genome editing with the CRISPR-Cas system is simpler, faster and more efficient. The CRISPR-Cas system has consequently been widely applied to genome editing in eukaryotic cells during the past two years<sup>1</sup>.

The CRISPR-Cas system, an adaptive immunity system against foreign nucleic acid invaders in prokaryotes, exists in most archaea and numerous bacteria<sup>5</sup>. The system has been categorized into three types (I, II, and III) based on the *Cas* genes and CRISPR sequences present<sup>6–8</sup>. Types I and III contain multiple CAS proteins that form a complex for degrading foreign DNA/RNA. Type II directs the cleavage of targeted foreign DNA using a single CAS9 protein, which makes it the system of choice for targeted genome engineering. In the Type-II CRISPR-Cas9 system, CRISPR RNA (crRNA) hybridizes with a small trans-activating CRISPR RNA (tracrRNA) to form mature dual crRNA<sup>9,10</sup>. The mature crRNA combines with Cas9 to form a functional complex. When the complex recognizes a short seed sequence in the vicinity of a typical 5'-NGG-3' protospacer-adjacent motif (PAM) by RNA/DNA base pairing,

<sup>1</sup>College of Agronomy, Northwest A&F University, Yangling, Shaanxi 712100, China. <sup>2</sup>National Key Facilities for Crop Genetic Resources and Improvement, Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing 100081, China. <sup>3</sup>Tianjin Institute of Agricultural Quality Standard and Testing Technology, Tianjin Academy of Agricultural Sciences, Tianjin 300381, China. \*These authors contributed equally to this work. Correspondence and requests for materials should be addressed to Z.H. (email: huzheng@caas.cn) or Y.X. (email: xiyajun2002@126.com)

Cas9 cleaves the target DNA<sup>11–13</sup>. Mature crRNA containing tracrRNA and crRNA can be replaced in the laboratory with a single synthetic guide RNA (sgRNA)<sup>10</sup>. Consequently, only sgRNA and Cas9 protein are needed to make genome editing simple and efficient. The CRISPR-Cas9 system has been widely applied in genetic studies of prokaryotes and eukaryotes over the past two years<sup>14</sup>.

In plants, the CRISPR-Cas9 system has been successfully used in various species including *Arabidopsis thaliana*, *Nicotiana benthamiana*, rice, tobacco, sorghum, wheat, maize, orange and liverwort<sup>15–22</sup>. In these applications, Cas9 is expressed by the Cauliflower mosaic virus (CaMV) 35s promoter or a gene-specific promoter. A nuclear localization signal (NLS) sequence is fused to the Cas9 gene, which delivers CAS9 to the genomic nuclei. To design the sgRNA, the 20-bp sequence following the PAM in the target DNA is selected as the sgRNA seed. These 20-bp sgRNA seed regions are abundant in plant genomes<sup>16</sup>, with more than 90% of rice genes containing a specific sgRNA seed<sup>23</sup>. Several software programs have been developed to identify target-specific sgRNA seeds<sup>24–26</sup>. RNA polymerase-III promoters such as *U6* and *U3* are typically used to express the sgRNA<sup>27</sup>. These polymerase-III promoters express transcripts with a purine initiated at the first nucleotide. The purine may or may not affect base pairing between the sgRNA and the target DNA. The CaMV 35s promoter is also used to express the sgRNA without an additional nucleotide<sup>21</sup>. Several different sgRNAs can be co-expressed within a single CRISPR-Cas9 system to target multiple DNA sites simultaneously. In plants, DSBs induced by CRISPR-Cas9 can be repaired by HR or NHEJ, the latter primarily responsible for genomic insertion or deletion (indel) mutations. The CRISPR-Cas9 system is a highly efficient tool to obtain targeted mutant transgenic plants with a high frequency of mutation<sup>27</sup>. Biallelic mutations have also been observed at high frequencies in T<sub>0</sub> transgenic plants<sup>16,19</sup>. Mutations can be stably inherited in the next generation<sup>28–30</sup>. The CRISPR-Cas9 system is a powerful tool for the advancement of functional genomics research. Targeted mutagenesis has also been applied to crop improvement. In wheat, *TaMLO* mutants induced by the CRISPR-Cas9 system have shown broad-spectrum resistance to powdery mildew<sup>31</sup>.

Mutagenesis has played an important role in functional genomics over the past two decades. Targeted mutagenesis is an efficient tool for functional genomics research. Although ZFNs have been used in recent years for targeted mutagenesis in soybean (*Glycine max* L.)<sup>32</sup>, construction difficulties, high cost and modest efficacy limit their application<sup>1</sup>. In this study, we used the CRISPR-Cas9 system to efficiently perform targeted mutagenesis in soybean protoplasts and hairy roots. Targeted mutagenesis using the CRISPR-Cas9 system can advance soybean functional genomic research, especially that of genes involved in roots and nodules.

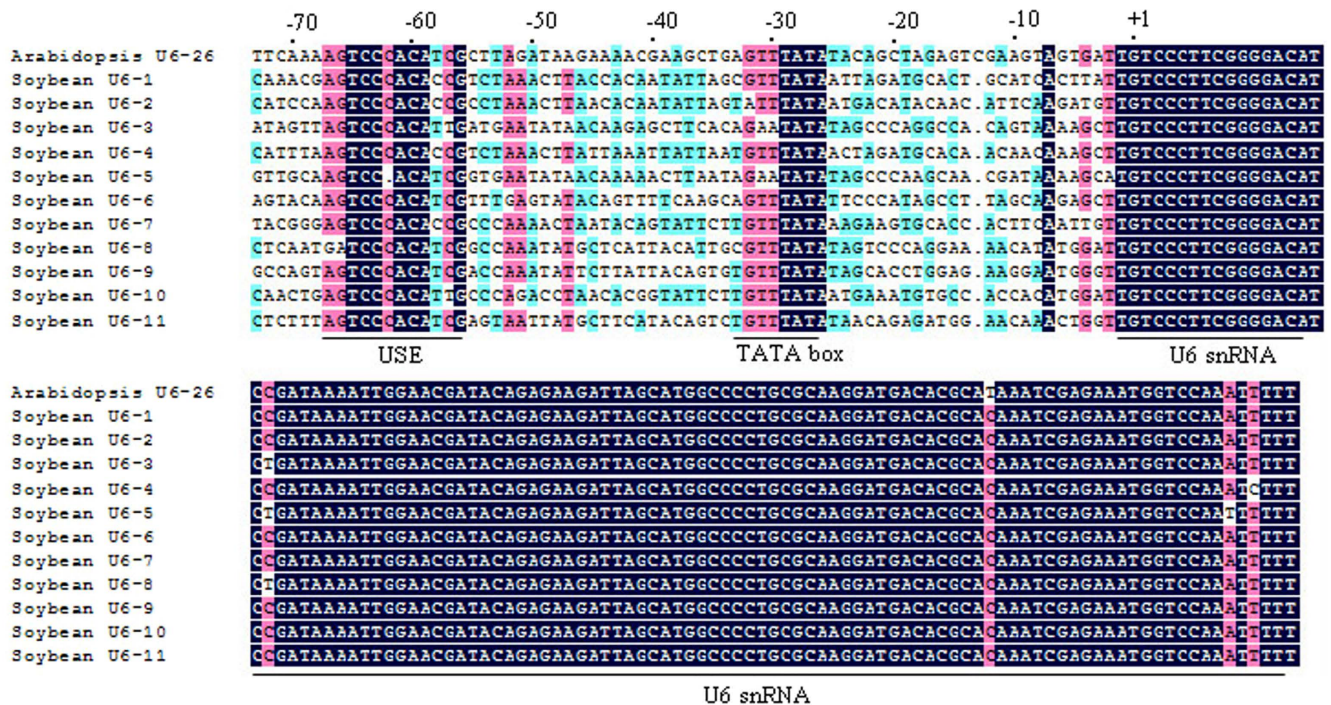
## Results

**Prediction of *U6* promoters in soybean.** The *U6* promoter is typically used to drive the expression of sgRNA in various plants<sup>27</sup>. The *Arabidopsis U6-26* promoter has been used to generate sgRNA in *Arabidopsis* and *N. benthamiana*; the rice *U6* promoter has been used in rice and sorghum. By comparison with the *Arabidopsis U6* small nuclear RNA (snRNA) sequence, we predicted 11 *U6* genes in the soybean genome. These 11 *U6* genes were distributed on seven chromosomes, with 2 *U6* genes (*U6-7* and *U6-8*) clustered on a 6.1-kb fragment on chromosome 16 (Supplementary Table S1). Plant *U6* promoters contain the following two conserved elements: an upstream sequence element (USE; consensus sequence RTCCACATCG) and a TATA-like box<sup>33</sup>. These two elements, separated by a suitable distance, are necessary for *U6* gene transcription<sup>33</sup>. Promoter sequences of the 11 *U6* genes were extracted from the soybean genome. Multiple sequence alignments revealed that the *U6-5* promoter had a C-nucleotide deletion in the USE, whereas the other 10 *U6* promoters contained both conserved elements (Fig. 1). The USE and TATA-like-box conserved sequences in soybean are RTCCACA(T/C)(T/C)G and GTTTATA, respectively. The presence of these conserved elements suggests that the 10 soybean *U6* promoters may have the transcriptional activity to generate sgRNAs in soybean.

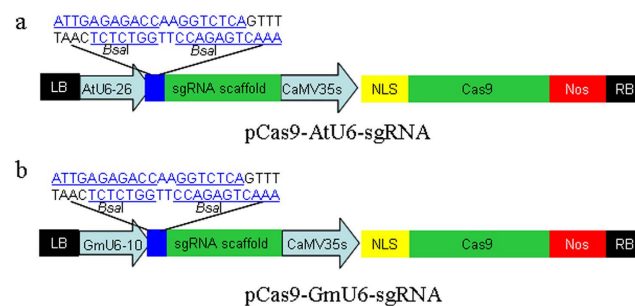
**Evaluation of the CRISPR-Cas9 system for gene editing in soybean.** We constructed two binary vectors to express sgRNAs and Cas9 for gene editing (Fig. 2). In both vectors, the CaMV 35s promoter was used to drive the expression of Cas9. Two RNA polymerase-III (Pol III) promoters, *AtU6-26* and *GmU6-10*, were selected to generate sgRNAs in the two vectors (pCas9-*AtU6*-sgRNA and pCas9-*GmU6*-sgRNA, respectively). Two *BsaI* sequences, easily replaceable by 20-bp sgRNA seed sequences, were introduced between the *U6* promoter and the sgRNA scaffold in the vectors.

To detect the activity of these two vectors in soybean, we selected three genes (*Glyma06g14180*, *Glyma08g02290* and *Glyma12g37050*) as targets for gene editing in soybean. For each gene, we designed a different sgRNA seed with a restriction site in the vicinity of the PAM (Supplementary Table S2). A total of six binary vectors were therefore generated to evaluate targeted mutagenesis in this study.

**Targeted mutagenesis in soybean protoplasts.** We first verified the activity of the CRISPR-Cas9 system in soybean protoplasts. The vectors were transformed into soybean protoplasts using the polyethylene glycol (PEG)-mediated transformation method. After 48 h of incubation in darkness at room temperature, the transformed protoplasts were collected for genomic DNA extraction. A restriction enzyme PCR (RE-PCR) assay was used to detect mutations in the targeted genes. The genomic DNAs for various targeted mutant genes were completely digested with restriction enzymes. The mutant genes were not digested as they lost the enzyme sites and could be amplified using the gene-specific primers. The PCR results confirmed that all six vectors were able to induce targeted gene mutations (Fig. 3a). Sequence



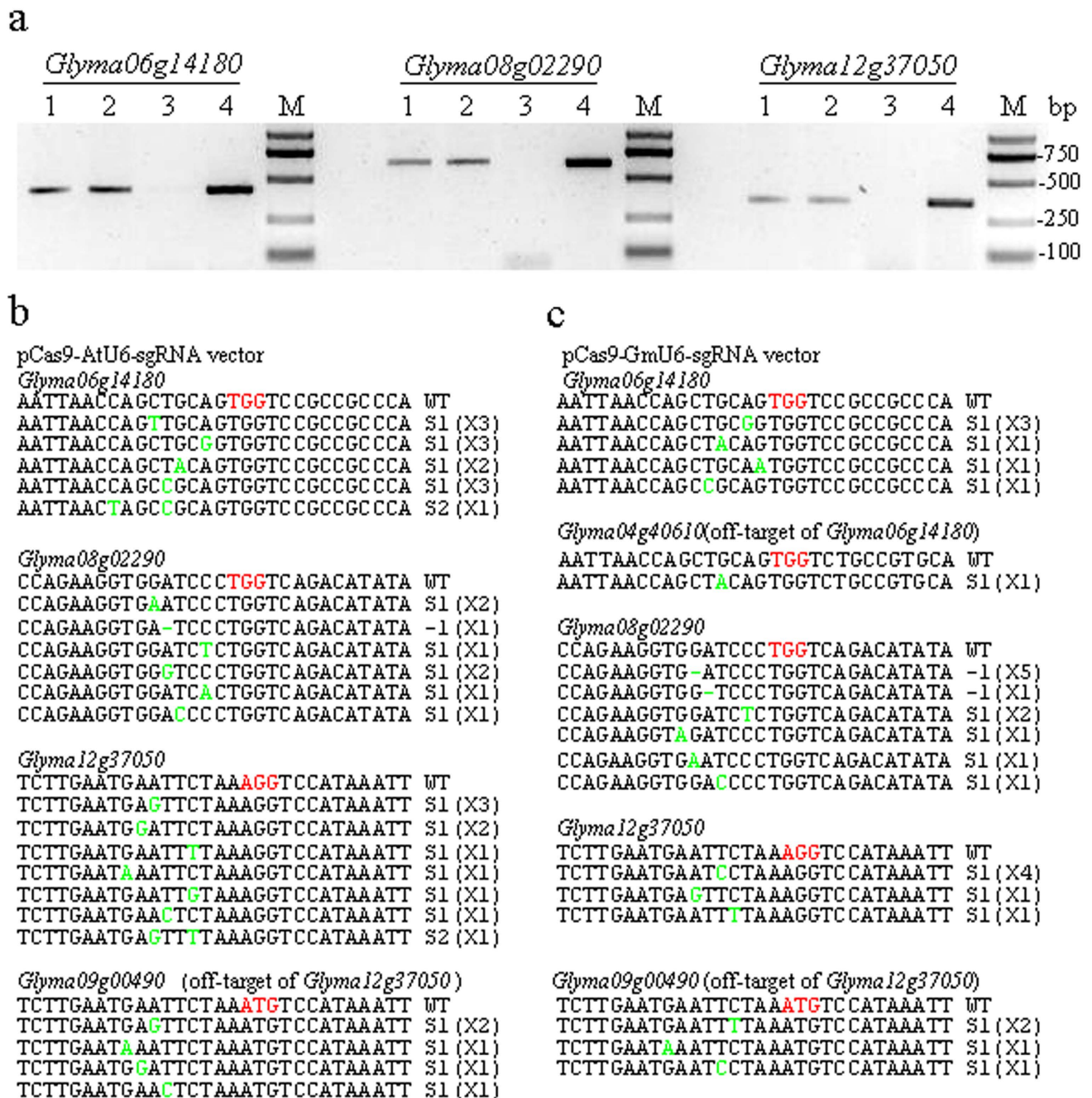
**Figure 1.** Alignment of 11 soybean *U6* and Arabidopsis *U6-26* genes. Soybean *U6* and Arabidopsis *U6-26* genes are highly conserved. Upstream sequence element (USE), TATA-box and *U6* small nuclear RNA (snRNA) sequence regions are underlined.



**Figure 2.** Construction of binary vectors for genome editing in soybean. Cas9 fused with a single nuclear localization signal (NLS) is expressed with a Cauliflower mosaic virus 35s (CaMV 35s) promoter. Synthetic guide RNA (sgRNA) is derived using *U6* promoters. (a) *Arabidopsis thaliana U6-26* promoter (b) *Glycine max U6-10* promoter. Sequences containing two *BsaI* sites are located between the *U6* promoter and the sgRNA scaffold. These sequences can be easily replaced with a gene-specific sgRNA seed. LB: left border; RB: right border.

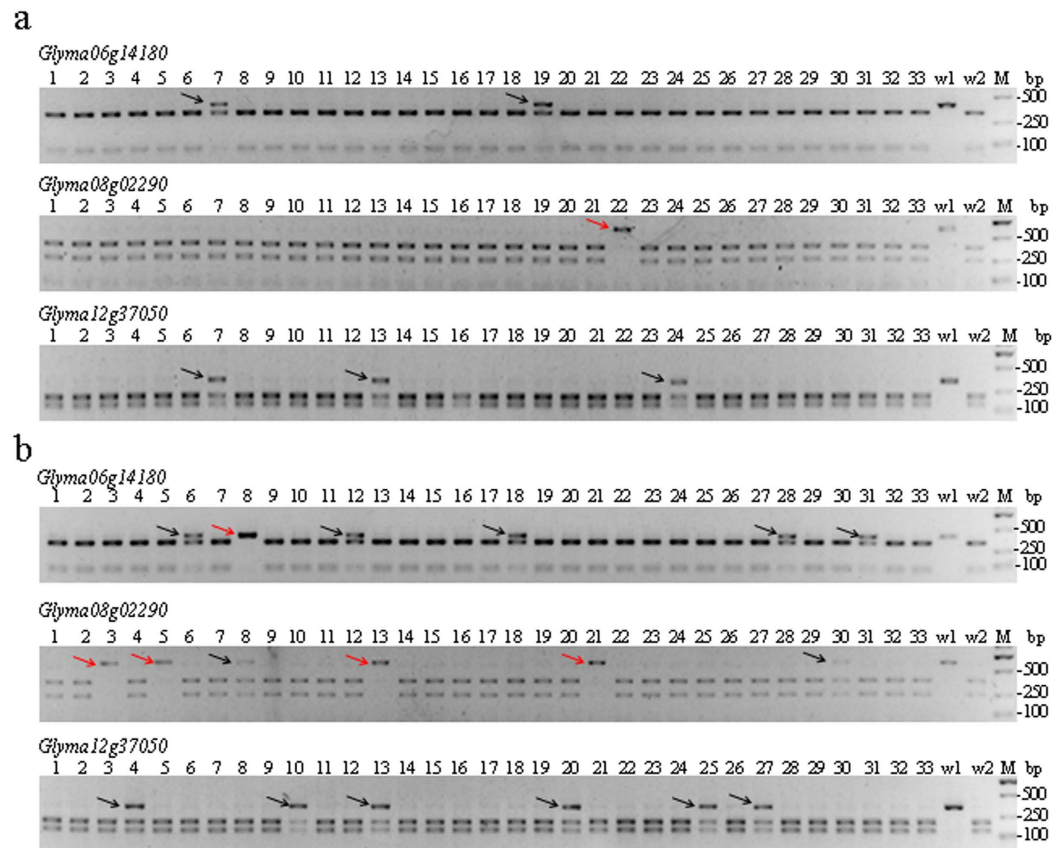
analysis revealed that nucleotide substitutions had occurred in *Glyma06g14180* and *Glyma12g37050* (Figs. 3b,c), suggesting that the DSBs in these two genes were repaired by the HR pathway in the soybean protoplasts. One nucleotide deletion and one substitution were found in *Glyma08g02290* (Figs. 3b,c). The DSBs of *Glyma08g02290* were repaired through both the HR and NHEJ pathways in soybean protoplasts.

**Targeted mutagenesis in soybean hairy roots.** *Agrobacterium rhizogenes* (*A. rhizogenes*)-mediated transformation is a rapid, efficient, simple and inexpensive method for the studying soybean root biology<sup>34</sup>. To detect the targeted gene mutations in soybean roots, we introduced the six binary vectors into *A. rhizogenes* strain K599 and then infected soybean seedling hypocotyls to induce hairy roots. Genomic DNA was collected and extracted for further detection of the target gene mutations from the hairy roots for each of the six vectors. Soybean is a diploid plant and genes have two copies in the homologous chromosomes. The target gene induced by the CRISPR-Cas9 system has three types in the hairy roots.



**Figure 3.** Targeted mutagenesis in soybean protoplasts. (a) Detection of mutations using restriction enzyme-PCR (RE-PCR). Lanes 1 and 2: PCR products of digested genomic DNA from protoplasts treated with pCas9-AtU6-sgRNA and pCas9-GmU6-sgRNA, respectively; Lanes 3 and 4: PCR products of digested and undigested genomic DNA, respectively, from wild-type controls. (b) and (c) Sequence-based detection of mutations induced by pCas9-AtU6-sgRNA and pCas9-GmU6-sgRNA vectors, respectively. Wild-type sequences of the target genes and off-target genes are shown with the protospacer-adjacent motif sequence highlighted in red. The change in the number of nucleotides is shown to the right of each sequence. D: deletion; S: substitution. Nucleotide substitutions are shown in green. The number of clones for each mutant is given in brackets.

Type I is no mutation of the target gene. Type II is a monoallelic mutation where one gene is mutated and the other allelic gene is no mutated. Type III is a biallelic mutation where both of the two allelic genes are mutated (Supplementary Figure S1). The gene is amplified using gene specific primers and then digested completely with the restriction enzyme (PCR-RE assay). When the gene mutation is induced by the CRISPR/Cas9 system, the restriction enzyme site in the gene is destroyed. The results of PCR-RE



**Figure 4.** Detection of mutants using the PCR-restriction enzyme (PCR-RE) assay. Detection of mutations using the PCR-restriction enzyme (PCR-RE) assay. Lanes 1–33: the digested DNA of the PCR products amplified from the independent hairy root samples; The monoallelic and biallelic mutants are shown with black and red arrows, respectively. w1 and w2: the undigested and digested DNA, respectively, from the PCR products amplified from wild-type controls. (a) Targeted mutations induced by the pCas9-AtU6-sgRNA vector. (b) Targeted mutations induced by the pCas9-GmU6-sgRNA vector.

Target gene	pCas9-GmU6-sgRNA vector				pCas9-AtU6-sgRNA vector			
	examined hairy roots	monoallelic mutants	biallelic mutants	Mutation efficiency	examined hairy roots	monoallelic mutants	biallelic mutants	Mutation efficiency
<i>Glyma06g14180</i>	95	12	2	14.7%	91	6	0	6.6%
<i>Glyma08g02290</i>	94	7	12	20.2%	93	1	2	3.2%
<i>Glyma12g37050</i>	84	15	0	17.9%	72	7	0	9.7%

**Table 1.** Gene mutations in three target genes using different vectors.

assay for the non-mutation show two digested bands. For the monoallelic mutation, the results are three bands with one undigested band from the mutated gene and two digested bands from non-mutated allelic gene. For the biallelic mutation, both of the two allelic genes are mutated and the PCR-RE assay shows only a single undigested band (Supplementary Figure S1). The PCR-RE assay shows that gene mutations were induced using all six vectors (Fig. 4 and Supplementary Figures S2–S7). The undigested bands from the PCR-RE assay were cloned and sequenced to confirm the mutations. Sequence analysis indicated that the types of mutations differed between the genes (Supplementary Figures S2–S7). Most *Glyma06g14180* mutations were single nucleotide insertions, whereas the majority of the detected mutations in *Glyma08g02290* and *Glyma12g37050* involved multiple-nucleotide deletions. Although rare in the soybean hairy roots, nucleotide substitutions were the major type of mutation induced in soybean protoplasts using the CRISPR-Cas9 system. Mutation efficiencies differed between the pCas9-GmU6-sgRNA and pCas9-AtU6-sgRNA vectors (Table 1), with markedly higher efficiencies obtained with all three genes using the pCas9-GmU6-sgRNA vector. Mutation efficiencies with the pCas9-GmU6-sgRNA vector

*Glyma06g14180*

GTACTCTGCTGGCTGTGAAATTAAC CAGCTGCAG**TGG**TCCGCCGCCAGCCCGAT WT

## Mutants

- 1 GTACTCTGCTGGCTGTGAAATTAAC CAGCTG**GC**AGTGGTCCGC CGCCAGCCCGAT +1 (X1)  
GTACTCTGCTGGCTGTGAAATTAAC CAGCTG**TC**AGTGGTCCGC CGCCAGCCCGAT +1 (X1)  
GTACTCTGCTGGCTGTGAAATTAAC CAGCTG**AC**AGTGGTCCGC CGCCAGCCCGAT +1 (X1)  
GTACTCTGCTGGCTGTGAAATTAAC CAGCTG-AGTGGTCCGCCGCCAGCCCGAT -1 (X1)  
GTACTCTGCTGGCTGTGAAATTAAC CAGCT---GTGGTCCGCCGCCAGCCCGAT -3 (X1)
- 2 GTACTCTGCTGGCTGTGAAATTAAC CAGCTG**TC**AGTGGTCCGC CGCCAGCCCGAT +1 (X1)  
GTACTCTGCTGGCTGTGAAATTAAC CAGCTG**AC**AGTGGTCCGC CGCCAGCCCGAT +1 (X1)  
GTACTCTGCTGGCTGTGAAATTAAC CAGCTG**TAC**AGTGGTCCGCCGCCAGCCCGAT +2 (X3)

*Glyma08g02290*

ACGGGTTAAGGCAGTACATTCCAGAAGGTGGATCC**TGG**TCAGACATATATAC WT

## Mutants

- 1 ACGGGTTAAGGCAGTACATTCCAGAA**TGTT**---TCCCTGGTCAGACATATATAC -2, S2 (X2)  
ACGGGTTAAGGCAGTACATTCCAGAAGGTGGAT-- (-36bp) --TCTTATGATA -36 (X2)
- 2 ACGGGTTAAGGCAGTACATTCCAGAAGGTGGAT**TTCC**CTGGTCAGACATATATAC +1 (X3)  
ACGGGTTAAGGCAGTACATTCCAGAAGGTGGAT---CTGGTCAGACATATATAC -2 (X4)
- 3 ACGGGTTAAGGCAGTACATTCCAGAAGGTGGAT**TTCC**CTGGTCAGACATATATAC +1 (X2)  
ACGGGTTAAGGCAGTACATTCCAGAAGGTGGAT---CTGGTCAGACATATATAC -2 (X4)
- 4 ACGGGTTAAGGCAGTACATTCCAGAAGGTGGAT**TTCC**CTGGTCAGACATATATAC +1 (X3)  
ACGGGTTAAGGC**TG**TACATTCCAGAAGGTGGAT**TTCC**CTGGTCAGACATATATAC +1, R1 (X1)  
ACGGGTTAAGGCAGTACATTCCA-----CCCTGGTCAGACATATATAC -10 (X1)  
ACGGGTTAAGGCAGTACATTCC-----TGGTCAGACATATATAC -14 (X1)  
ACGGGTT----- (-33bp) -----CAGACATATATACCCGAGATAAACTGGA -33 (X1)
- 5 ACGGGTTAAGGCAGTACATTCCAGAAGGT--- (-36bp) --TCTTATGATAAT -36 (X4)  
ACGGGTTAAGGCAGTACATTCCAG---TGG---AGA--- (-14bp) --ATAAACT -19 (X1)  
ACGGGTTAAGGCAGTACATT-----CCCTGGTCA**AA**ACATATAT**GC** -12, S2 (X1)
- 6 ACGGGTTAAGGCAGTACATTCCAGAAGGTGGAT**TTCC**CTGGTCAGACATAT**GTAC** +1, S1 (X4)  
ACGGGTTAAGGCAGTACATTCCAGA-----CCCTGGTCAGACATATATAC -8 (X4)
- 7 TCCAGAAAGGTGG**GA**AACATT**CA**AAT**GC**AT**GA**CCT**GT**TTAT**CC**ACCCCTGGTCAGA +30 (X3)  
ACGGGTTAAGGCAGTACATTCC-----CTGGTCA**AA**ACATATATAC -13, S1 (X2)

**Figure 5.** Gene sequences from 9 independent biallelic mutants. Gene sequences are shown for 9 independent biallelic mutants. Wild-type sequences of the target genes are shown with the protospacer-adjacent motif sequence highlighted in red. The change in the number of nucleotides is shown to the right of each sequence. +: insertion; -: deletion; S: substitution. Inserted and substituted nucleotides are shown in green.

for *Glyma06g14180*, *Glyma08g02290* and *Glyma12g37050* were 14.7, 20.2 and 17.9%, respectively, with corresponding efficiencies of 6.6, 3.2 and 9.7% using the pCas9-AtU6-sgRNA vector.

Biallelic mutations can be detected in  $T_0$  transgenic plants using the CRISPR-Cas9 system<sup>16,19</sup>. We detected several biallelic mutations of *Glyma06g14180* and *Glyma08g02290* using the PCR-RE assay (Supplementary Figures S3–S5). A higher frequency of biallelic mutants was observed in *Glyma08g02290*. Twelve of 19 *Glyma08g02290* mutants generated using the pCas9-GmU6-sgRNA vector and 2 of 3 *Glyma08g02290* mutants induced by the pCas9-AtU6-sgRNA vector were biallelic (Table 1). Sequencing of several gene clones from independent biallelic mutant roots revealed a variety of mutations per root (Fig. 5), which suggests that the CRISPR-Cas9 system continued to modify the genes during hairy root development.

**Off-target activity in soybean.** The CRISPR-Cas9 system can tolerate several mismatches between the sgRNA seed and its target, especially in the first 12 nucleotides at the 5' end of the sgRNA seed<sup>35–37</sup>, which suggests that off-target activity is common with the CRISPR-Cas9 system. We accordingly searched the soybean genome for homologs of the three targeted genes in this study. We found that *Glyma06g14180* and *Glyma04g40610* had the same target sequence and that the sequences of *Glyma08g02290* and *Glyma05g37270* were also identical to one another. *Glyma12g37050* and *Glyma09g00490* differed by a single nucleotide at the PAM site (AGG vs. ATG). Mutations in *Glyma04g40610* and *Glyma09g00490* induced by the CRISPR-CAS9 system using primers for *Glyma06g14180* and *Glyma12g37050* were detected in protoplasts and hairy roots (Fig. 3 and Supplementary Figures S3, S6 and S7).

## Discussion

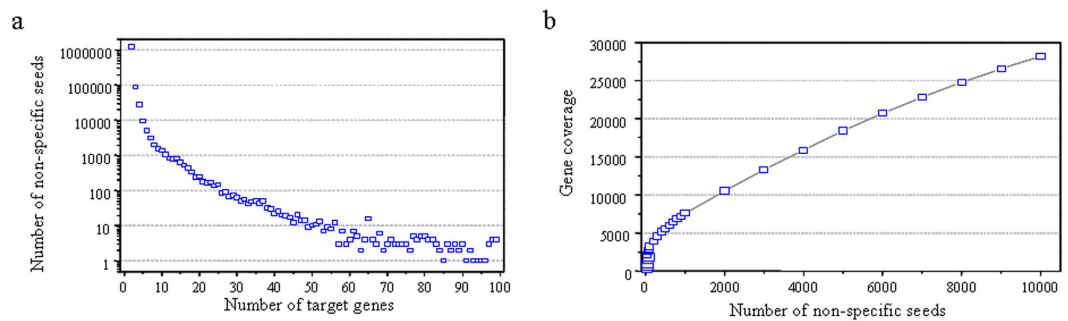
In this study, we used two *U6* promoters, Arabidopsis *U6-26* and soybean *U6-10*, to generate sgRNA. Mutation efficiencies in the three target genes were significantly increased by the use of the soybean *U6-10* promoter (Table 1), which may be related to the *U6* promoter activity. The transcriptional efficiency of the different *U6* promoters varies in Arabidopsis<sup>38</sup>. Eleven *U6* promoters were predicted in soybean, which provided the opportunity to select a suitable *U6* promoter for the expression of sgRNA in soybean. The choice of promoter is critical, as high concentrations of the Cas9-sgRNA complex can increase off-target activity<sup>35,37</sup>.

Mutagenesis is a powerful tool for the studying gene function. The mutations induced by T-DNA insertion, chemical agents and physical treatments are random, which make it difficult to obtain the target mutants. Targeted mutagenesis technologies, such as TALEN, ZFN and CRISPR-Cas9 approaches, are powerful tools to generate target gene mutations. Compared with TALENs and ZFNs, the CRISPR-Cas9 system efficiently produces mutations and is easy to use<sup>1</sup>. In this study, we successfully used the CRISPR-Cas9 system for target gene mutation in soybean. The mutation efficiencies are ranged from 14.7% to 20.2% (Table 1). Sequencing of several gene clones from the mutant roots revealed that the CRISPR-Cas9 system continued to modify the genes during hairy root development, which suggests that the mutation efficiency would be increased given enough time for the development of the transgenic plants. The high efficiency of the target gene mutation can improve the research on gene function in soybean.

Biallelic mutations can be detected and their phenotypes observed in  $T_0$  transgenic plants using the CRISPR-Cas9 system<sup>16,19</sup>. In a study by Ron *et al.*<sup>39</sup>, the CRISPR-Cas9 system mediated by *A. rhizogenes* was used to produce a targeted mutation in the *SHORT-ROOT (SHR)* gene in tomato transgenic hairy roots. The phenotype of the resulting mutant was consistent with Arabidopsis *shr* mutants. In our study, biallelic mutations in *Glyma06g14180* and *Glyma08g02290* were detected in transgenic hairy roots (Fig. 5 and Supplementary Figures S3–S5). *Glyma08g02290* had a higher number of biallelic mutations, with 12 of 19 root samples showing mutations (Table 1). Biallelic mutants can be detected easily using the PCR-RE assay (Supplementary Figures S3–S5). Biallelic mutants are the ideal materials for researching gene function. Compared to the inefficient and time-consuming transformation mediated by *Agrobacterium tumefaciens (A. tumefaciens)*, transformation mediated by *A. rhizogenes* is easy, quick and efficient in soybean<sup>34</sup>. Transgenic hairy roots can be obtained within one month with transformation efficiencies up to 80%. A large number of genes involved in the roots and nodules have been identified in soybean by next-generation sequencing<sup>40</sup>. It would be easy to generate the target gene mutants using the CRISPR-Cas9 system mediated by *A. rhizogenes*, which would lead to advances in soybean root biology research.

Off-target activity is common using the CRISPR-Cas9 system. In our study, we detected off-target gene mutations for *Glyma06g14180* and *Glyma12g37050* (Fig. 3 and Supplementary Figures S3, S6 and S7). Off-target activity limits the application of the CRISPR-Cas9 system, but several methods are available to reduce this impediment. Decreasing sgRNA-Cas9 concentrations can increase on-target specificity *in vitro*<sup>35,37</sup>. Off-target activity can be reduced 50- to 1500-fold using double-nicking mediated by a *Cas9* nickase mutant (*Cas9n*)<sup>41</sup>. Use of truncated gRNAs (tru-gRNAs), a shorter sgRNA seed (typically 17 or 18 nucleotides) complementary to the target, can also decrease off-target activity by 5000-fold or more<sup>42</sup>. Although these methods can effectively reduce off-target activity, the best strategy is identification of gene-specific sgRNA seeds. Fortunately, 97.3% of annotated transcription units (TUs) have specific sgRNA seeds in soybean; these TU-specific sgRNA seeds can be identified by searching the CRISPR-PLANT database (<http://www.genome.arizona.edu/crispr>)<sup>43</sup>.

Some mutation libraries have been developed by chemical agents and physical treatments in soybean<sup>44–46</sup>, but the mutants induced by these treatments are random and complex. T-DNA-induced mutagenesis has been widely applied in model plants such as Arabidopsis and rice<sup>47,48</sup>. Successful T-DNA insertion mainly depends on efficient of *A. tumefaciens*-mediated transformation. In soybean, the creation of large numbers of mutants using T-DNA insertion is not feasible, as transformation efficiency mediated by *A. tumefaciens* is low in this species. Nevertheless, the acquisition of target mutants is still time-consuming and inefficient because T-DNA-based mutagenesis is random. Off-target activity can be exploited for the construction of a saturated gene mutation library in soybean. The CRISPR-Cas9 system can tolerate several mismatches between the sgRNA seed and its target, especially in the first 12 nucleotides at the 5' end of the sgRNA seed<sup>35,36,49</sup>. With respect to these 12 nucleotides, sgRNA seeds having fewer than four mismatches with other sequences in our study were considered to be non-specific



**Figure 6.** Non-specific synthetic guide RNA (sgRNA) seeds in soybean. (6a) Distribution of non-specific sgRNA seeds and the number of their target genes. More than 1 million sgRNA seeds were associated with two target genes; approximately 100,000 sgRNA seeds were able to target three genes. sgRNA seeds having more than 100 target genes are not shown. (6b) Maximal gene coverage of non-specific sgRNA seeds. The non-specific sgRNA seeds were sorted by their target gene numbers before calculating the maximal gene coverage.

sgRNA seeds. A total of 13,103,481 sgRNA seeds were predicted in soybean genes, of which 5,631,730 were specific and 7,469,546 were non-specific (Supplementary Figure S8). The number of specific sgRNA seeds as well as their coverage (99.5% of soybean genes) is consistent with results obtained by Xie *et al.*<sup>43</sup>. The huge quantity of non-specific sgRNA seeds allows the targeting of two or more genes in one transformation in soybean (Fig. 6a). Off-target activity produces numerous mutations covering different genes in  $T_0$  transgenic soybeans. The resulting mutants can be segregated to produce unique mutations in the progeny, which, similar to the application of *Ac/Ds* transposons or *Tnt1* retrotransposons in T-DNA transformations, improves mutation efficiency<sup>50,51</sup>. In our study, the seeds of *Glyma06g14180* or *Glyma12g37050* were detected to produce two gene mutations (*Glyma06g14180* and *Glyma04g40610*, *Glyma12g37050* and *Glyma09g00490*) respectively in one transgenic plant (Figure S3, S6 and S7). By exploiting off-target activity, the number of transgenic soybean plants required to produce a saturated mutation library can be reduced dramatically (Fig. 6b).

## Methods

**Plant material.** The soybean cultivar Williams 82 was used in this study. In preparation for *A. rhizogenes*-mediated transformation, seeds were sterilized for 7 h with chlorine gas. Seeds were germinated under 16-h light/8-h dark at 25 °C in a humidity chamber. After one week, healthy plants were selected for transformation. To generate protoplasts, seeds were germinated under 16-h light/8-h dark at 25 °C in a low-humidity chamber. Fresh leaves were collected for protoplast preparation from 2-week-old seedlings.

**Vector construction.** A codon-optimized *cas9* gene with a NLS was obtained from Professor Qu (Qu, State Key Laboratory for Protein and Plant Gene Research, Peking-Tsinghua Center for Life Sciences, College of Life Sciences, Peking University). The *cas9* gene was amplified by phusion polymerase (NEB, Massachusetts, USA) using *cas9*-specific primers (Supplementary Table S3) and cloned into pCambia3301 vector by replacing of the *gus* gene.

*Arabidopsis U6-26* and soybean *U6-10* promoters with sgRNA were synthesized (Genscript, Nanjing, China) (Supplementary Figure S9 and S10) and cloned into pUC57-Kan vectors to generate pUC57-AtU6-26-sgRNA and pUC57-GmU6-10-sgRNA plasmids, respectively. These two plasmids were digested completely using *BsaI* (NEB, Massachusetts, USA) and purified with a TIANquick Midi purification kit (Tiagen, Beijing, China). Three target gene oligonucleotides (Supplementary Table S2) were annealed to form sgRNA seeds and were then ligated into the pUC57-AtU6-26-sgRNA and pUC57-GmU6-10-sgRNA vectors. These six vectors and the pCambia3301-Cas9 vector were digested completely using *EcoRI* and *HindIII*. After digestion, the pCambia3301-Cas9 vector and AtU6-26-sgRNAs and GmU6-10-sgRNAs of different genes were purified with a TIANgel Midi purification kit (Tiagen, Beijing, China) and ligated overnight using T4 DNA ligase (Fermentas) to obtain pCas9-AtU6-sgRNA and pCas9-GmU6-sgRNA vectors for different target genes.

**Protoplast isolation and transformation.** Soybean protoplasts were prepared from fresh leaves as described by Yoo *et al.*<sup>52</sup> with some modifications. Briefly, 20 fresh leaves were cut into small strips and immediately transferred into 10 ml digestion solution (0.5% cellulose R10, 0.5% macerozyme R10, 0.1% pectolase Y23, 0.6M mannitol, 10 mM 4-morpholineethanesulfonic acid (MES) pH 5.7, 20 mM KCl, 10 mM  $\text{CaCl}_2$  and 0.1% BSA). The leaf strips were vacuum infiltrated for 30 min in the dark using a vacuum pump at  $-15$  to  $-20$  mm Hg and digested for 6 h with agitation at 30 rpm. The other steps are followed as described by Yoo *et al.*<sup>52</sup>. The protoplasts were re-suspended in MGG solution (4 mM MES



pH 5.7, 0.4 M mannitol, 15 mM MgCl<sub>2</sub>) for the plasmid transformation. Plasmids were transformed into protoplasts mediated by PEG as described by Yoo *et al.*<sup>52</sup>.

**Transformation mediated by *A. rhizogenes*.** The binary vectors were transformed into soybean by *A. rhizogenes* as described by Kereszt *et al.*<sup>34</sup>.

**Detection of mutations in target genes.** Genomic DNA was extracted using a DNAquick Plant System (Tiangen, Beijing, China) according to the manufacturer's protocol with a minor modification: genomic DNA from soybean hairy roots was precipitated using Dr.GenTLE Precipitation Carrier (Takara, Dalian, China). To detect mutations in soybean protoplasts, the genomic DNA was digested with restriction enzyme (*Pst*I, *Bam*HI and *Eco*RI for mutant detection of *Glyma06g14180*, *Glyma08g02290* and *Glyma12g37050* respectively). After digestion, the target genes were amplified with gene-specific primers, and the PCR fragments were ligated to an pEASY-T1 vector (Transgen, Beijing, China) for sequencing. To detect mutations in hairy roots, the target genes were amplified by PCR using gene-specific primers (Supplementary Table S3). The PCR products were purified using TIANquick N96 Purification kit (Transgen, Beijing, China) and digested for three hours with *Pst*I, *Bam*HI and *Eco*RI, respectively. The undigested bands were purified using a TIANgel Midi purification kit (Tiangen, Beijing, China) and then ligated to a pEASY-T1 vector (Transgen, Beijing, China). Several clones were randomly selected and sequenced to detect gene mutations.

**Bioinformatic analysis.** Soybean genome and annotation data were downloaded from the plant-GDB database (<http://www.plantgdb.org/>). The bioinformatic analysis pipeline was primarily constructed using customized Perl scripts and the USEARCH program<sup>53</sup>. For specificity assessment of sgRNA seeds, 20-nt long sgRNA spacer sequences adjacent to NGG PAM sites were excluded from both strands of the soybean chromosome sequences. For specificity analysis, sgRNA seeds were first grouped according to the identity of the eight nucleotides at the 3'-, end. The first 12 nucleotides at the 5'-, end were then compared among members of the same group. sgRNA seeds with no less than four mismatches were regarded as specific candidates; the remaining seeds, including repeat sequences, were considered to be non-specific.

## References

- Chen, K. & Gao, C. Targeted genome modification technologies and their applications in crop improvements. *Plant Cell Rep.* **33**, 575–583 (2014).
- Wyman, C. & Kanaar, R. DNA double-strand break repair: all's well that ends well. *Annu. Rev. Genet.* **40**, 363–383 (2006).
- Gaj, T., Gersbach, C. A. & Barbas, C. F., 3rd. ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends Biotechnol.* **31**, 397–405 (2013).
- Puchta, H. & Fauser, F. Synthetic nucleases for genome engineering in plants: prospects for a bright future. *Plant J.* **78**, 727–741 (2013).
- Koonin, E. V. & Makarova, K. S. CRISPR-Cas: an adaptive immunity system in prokaryotes. *F1000 Biol. Rep.* **1**, 95 (2009).
- Makarova, K. S. *et al.* Evolution and classification of the CRISPR-Cas systems. *Nat. Rev. Microbiol.* **9**, 467–477 (2011).
- Bhaya, D., Davison, M. & Barrangou, R. CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation. *Annu. Rev. Genet.* **45**, 273–297 (2011).
- Barrangou, R. CRISPR-Cas systems and RNA-guided interference. *WIREs RNA* **4**, 267–278 (2013).
- Deltcheva, E. *et al.* CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471**, 602–607 (2011).
- Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
- Kuscu, C., Arslan, S., Singh, R., Thorpe, J. & Adli, M. Genome-wide analysis reveals characteristics of off-target sites bound by the Cas9 endonuclease. *Nat. Biotechnol.* **32**, 677–683 (2014).
- Wu, X. *et al.* Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. *Nat. Biotechnol.* **32**, 670–676 (2014).
- Anders, C., Niewoehner, O., Duerst, A. & Jinek, M. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature* **513**, 569–573 (2014).
- Hsu, P.D., Lander, E. S. & Zhang, F. Development and applications of CRISPR-Cas9 for genome engineering. *Cell* **157**, 1262–1278 (2014).
- Nekrasov, V., Staskawicz, B., Weigel, D., Jones, J. D. & Kamoun, S. Targeted mutagenesis in the model plant *Nicotiana benthamiana* using Cas9 RNA-guided endonuclease. *Nat. Biotechnol.* **31**, 691–693 (2013).
- Shan, Q. *et al.* Targeted genome modification of crop plants using a CRISPR-Cas system. *Nat. Biotechnol.* **31**, 686–688 (2013).
- Li, J. F. *et al.* Multiplex and homologous recombination-mediated genome editing in *Arabidopsis* and *Nicotiana benthamiana* using guide RNA and Cas9. *Nat. Biotechnol.* **31**, 688–691 (2013).
- Feng, Z. *et al.* Efficient genome editing in plants using a CRISPR/Cas system. *Cell Res.* **23**, 1229–1232 (2013).
- Miao, J. *et al.* Targeted mutagenesis in rice using CRISPR-Cas system. *Cell Res.* **23**, 1233–1236 (2013).
- Jiang, W. *et al.* Demonstration of CRISPR/Cas9/sgRNA-mediated targeted gene modification in *Arabidopsis*, tobacco, sorghum and rice. *Nucleic Acids Res.* **41**, e188 (2013).
- Jia, H. & Wang, N. Targeted genome editing of sweet orange using Cas9/sgRNA. *PLoS One* **9**, e93806 (2014).
- Fauser, F., Schiml, S. & Puchta, H. Both CRISPR/Cas-based nucleases and nickases can be used efficiently for genome engineering in *Arabidopsis thaliana*. *Plant J.* **79**, 348–359 (2014).
- Xie, K. & Yang, Y. RNA-guided genome editing in plants using a CRISPR-Cas system. *Mol. Plant* **6**, 1975–1983 (2013).
- Xiao, A. *et al.* CasOT: a genome-wide Cas9/gRNA off-target searching tool. *Bioinformatics* **30**, 1180–1182 (2014).
- Bae, S., Park, J. & Kim, J. S. Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. *Bioinformatics* **30**, 1473–1475 (2014).
- Xie, S. S., Shen, B., Zhang, C. B., Huang, X. X. & Zhang, Y. L. sgRNACas9: A Software Package for Designing CRISPR sgRNA and Evaluating Potential Off-Target Cleavage Sites. *PLoS One* **9**, e100448 (2014).

27. Belhaj, K., Chaparro-Garcia, A., Kamoun, S. & Nekrasov, V. Plant genome editing made easy: targeted mutagenesis in model and crop plants using the CRISPR/Cas system. *Plant Methods* **9**, 39 (2013).
28. Feng, Z. *et al.* Multigeneration analysis reveals the inheritance, specificity, and patterns of CRISPR/Cas-induced gene modifications in Arabidopsis. *Proc. Natl Acad. Sci. USA* **111**, 4632–4637 (2014).
29. Zhang, H. *et al.* The CRISPR/Cas9 system produces specific and homozygous targeted gene editing in rice in one generation. *Plant Biotechnol. J.* **12**, 797–807 (2014).
30. Jiang, W. Z., Yang, B. & Weeks, D. P. Efficient CRISPR/Cas9-Mediated Gene Editing in Arabidopsis thaliana and Inheritance of Modified Genes in the T2 and T3 Generations. *PLoS One* **9**, e99225 (2014).
31. Wang, Y. *et al.* Simultaneous editing of three homoeoalleles in hexaploid bread wheat confers heritable resistance to powdery mildew. *Nat. Biotechnol.* **32**, 947–951 (2014).
32. Curtin, S. J. *et al.* Targeted mutagenesis of duplicated genes in soybean with zinc-finger nucleases. *Plant Physiol* **156**, 466–473 (2011).
33. Waibel, F. & Filipowicz, W. U6 snRNA genes of Arabidopsis are transcribed by RNA polymerase III but contain the same two upstream promoter elements as RNA polymerase II-transcribed U-sRNA genes. *Nucleic Acids Res.* **18**, 3451–3458 (1990).
34. Kereszt, A. *et al.* Agrobacterium rhizogenes-mediated transformation of soybean to study root biology. *Nat. Protoc.* **2**, 948–952 (2007).
35. Hsu, P. D. *et al.* DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* **31**, 827–832 (2013).
36. Fu, Y. *et al.* High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat. Biotechnol.* **31**, 822–826 (2013).
37. Pattanayak, V. *et al.* High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nat. Biotechnol.* **31**, 839–843 (2013).
38. Li, X., Jiang, D. H., Yong, K. & Zhang, D. B. Varied Transcriptional Efficiencies of Multiple Arabidopsis U6 Small Nuclear RNA Genes. *J. Integr. Plant Biol.* **49**, 222–229 (2007).
39. Ron, M. *et al.* Hairy Root Transformation Using Agrobacterium rhizogenes as a Tool for Exploring Cell Type-Specific Gene Expression and Function Using Tomato as a Model. *Plant Physiol* **166**, 455–469 (2014).
40. Severin, A. J. *et al.* RNA-Seq Atlas of Glycine max: a guide to the soybean transcriptome. *BMC Plant Biol.* **10**, 160 (2010).
41. Ran, F. A. *et al.* Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell* **154**, 1380–1389 (2013).
42. Fu, Y., Sander, J. D., Reyon, D., Cascio, V. M. & Joung, J. K. Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nat. Biotechnol.* **32**, 279–284 (2014).
43. Xie, K., Zhang, J. & Yang, Y. Genome-wide prediction of highly specific guide RNA spacers for CRISPR-Cas9-mediated genome editing in model plants and major crops. *Mol. Plant* **7**, 923–926 (2014).
44. Cooper, J. L. *et al.* TILLING to detect induced mutations in soybean. *BMC Plant Biol.* **9**, 9 (2008).
45. Bolon, Y. T. *et al.* Phenotypic and genomic analyses of a fast neutron mutant population resource in soybean. *Plant Physiol* **156**, 240–253 (2011).
46. Anai, T. Potential of a mutant-based reverse genetic approach for functional genomics and molecular breeding in soybean. *Breed Sci* **61**, 462–467 (2012).
47. Fu, F. E., Ye, R., Xu, S. P. & Xue, H. W. Studies on rice seed quality through analysis of a large-scale T-DNA insertion population. *Cell Res.* **19**, 380–391 (2009).
48. Alonso, J. M. *et al.* Genome-wide insertional mutagenesis of Arabidopsis thaliana. *Science* **301**, 653–657 (2003).
49. Pattanayak, V. *et al.* High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nat. Biotechnol.* **31**, 839–843 (2013).
50. Cui, Y. *et al.* Tnt1 retrotransposon mutagenesis: a tool for soybean functional genomics. *Plant Physiol* **161**, 36–47 (2013).
51. Mathieu, M. *et al.* Establishment of a soybean (Glycine max Merr. L) transposon-based mutagenesis repository. *Planta* **229**, 279–289 (2009).
52. Yoo, S. D., Cho, Y. H. & Sheen, J. Arabidopsis mesophyll protoplasts: a versatile cell system for transient gene expression analysis. *Nat. Protoc.* **2**, 1565–1572 (2007).
53. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).

## Acknowledgments

We thank Dr. Lijia Qu for kindly providing the *Cas9* gene used in this study. We thank Dr. Jin Miao for helpful advice on vector construction. This study was supported by the National Transgenic Key Project from the Ministry of Agriculture of China (2014ZX08011-003) and the Agricultural Science and Technology Innovation Program.

## Author Contributions

Conceived and designed the experiments: H.Z., Y.X. and Z.H. Performed the experiments: Z.H., X.S., R.C., Q.J. and G.S. Analyzed the data: X.S., R.C. and Z.H. Wrote the paper: H.Z., Y.X. and Z.H.

## Additional Information

**Supplementary information** accompanies this paper at <http://www.nature.com/srep>

**Competing financial interests:** The authors declare no competing financial interests.

**How to cite this article:** Sun, X. *et al.* Targeted mutagenesis in soybean using the CRISPR-Cas9 system. *Sci. Rep.* **5**, 10342; doi: 10.1038/srep10342 (2015).



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>