nature biotechnology

ARTICLES

https://doi.org/10.1038/s41587-021-00891-x

Genome-wide specificity of prime editors in plants

Shuai Jin[1,2,6], Qiupeng Lin[1,2,6], Yingfeng Luo[3,6], Zixu Zhu[1,2,6], Guanwen Liu[4,5], Yunjia Li[1], Kunling Chen[1], Jin-Long Qiu[4,5] and Caixia Gao[1,2] ✉

Although prime editors (PEs) have the potential to facilitate precise genome editing in therapeutic, agricultural and research applications, their specificity has not been comprehensively evaluated. To provide a systematic assessment in plants, we first examined the mismatch tolerance of PEs in plant cells and found that the editing frequency was influenced by the number and location of mismatches in the primer binding site and spacer of the prime editing guide RNA (pegRNA). Assessing the activity of 12 pegRNAs at 179 predicted off-target sites, we detected only low frequencies of off-target edits (0.00–0.23%). Whole-genome sequencing of 29 PE-treated rice plants confirmed that PEs do not induce genome-wide pegRNA-independent off-target single-nucleotide variants or small insertions/deletions. We also show that ectopic expression of the Moloney murine leukemia virus reverse transcriptase as part of the PE does not change retrotransposon copy number or telomere structure or cause insertion of pegRNA or messenger RNA sequences into the genome.

Prime editing[1], a 'search-and-replace' CRISPR-based genome editing technique, which has great potential in gene therapy[2,3] and agriculture[4,5], can introduce desired base conversions, deletions, insertions and combination edits into target genomic sites. Prime editing systems consist of a fusion of Moloney murine leukemia virus reverse transcriptase (MMLV-RT) with the nCas9 (H840A) nickase and a pegRNA that contains a spacer sequence, a primer binding site (PBS) sequence and an RT template sequence[1]. The RT template sequence encodes the desired edits to be reverse transcribed into DNA and inserted into the target site[1]. This technology has been successfully applied in animals[6,7] and plants[8–11]. Off-target effects are one of the major factors affecting the application of CRISPR-based genome editing tools and are composed of two types: single guide RNA (sgRNA) dependent and sgRNA independent[12–15]. These effects result from similarities between on-target and off-target sequences and over-expression of functional elements in the CRISPR-based tools, respectively.

Previous work has suggested that PEs have higher sgRNA-dependent specificity than CRISPR-based knockout systems owing to the three distinct DNA complementarities required[1]: target DNA-pegRNA spacer, target DNA-pegRNA PBS and target DNA-RT template. However, the pegRNA-dependent and pegRNA-independent off-target effects of PEs have not been comprehensively evaluated[1]. Using previously developed plant PEs[8] with optimized codon usage and promoters for plant applications, we performed a comprehensive and genome-wide analysis of pegRNA-dependent and pegRNA-independent off-target effects by targeted deep sequencing and whole-genome sequencing (WGS) in rice protoplast and regenerated rice plants. We found that the PE system produced pegRNA-dependent off-target effects but not pegRNA-independent ones. An additional four reverse transcription-associated analyses also revealed that PEs do not affect the endogenous reverse transcription mechanisms of plant cells.

Results

PE2 tolerance to mismatches in pegRNA. We tested the tolerance of prime editing to mismatches in the spacer and PBS sequence of pegRNA in rice protoplasts by deep amplicon sequencing. We first selected two previously reported sites[8], *OsCDC48-T1* and *OsGAPDH* (Fig. 1a), and measured on- and off-editing efficiencies after co-transfecting mismatched pegRNAs paired with 5 μg of PE plasmid DNA (a saturating concentration) (Supplementary Fig. 1 and Table 1). We found that mismatches located in seed sequence regions (near the protospacer adjacent motif (PAM)) of the spacer greatly decreased the efficiency of prime editing (Fig. 1a). These results were similar to the previously published findings in Cas9-based knockout and base editing systems[16–21]. In the case of PBS mismatch tolerance, we found that prime editing efficiency was decreased by mismatches near the nicking site of nCas9 (H840A) (Fig. 1b). This was probably due to the absence of genomic DNA-PBS pairing, which would lead in failure to initiate reverse transcription. To make the results more generalizable, we measured the on- and off-target editing efficiencies of prime editing with 29 additional mismatched pegRNAs at seven target sites. These pegRNAs harbored single mismatches at different locations in the PBS and spacer sequences (Fig. 1c-i). Only two of the eight pegRNAs with mismatches at the 5′ terminus of the PBS sequence (distal to the nick site) affected editing efficiencies (PBSMM01 and PBSMM02 in Fig. 1c–f). In contrast, all of the 14 pegRNAs with nick-adjacent mismatches in the PBS or spacer sequence greatly reduced prime editing (PBSMM03 and sgMM in Fig. 1c–i). Meanwhile, all seven pegRNAs with mismatches in both PBS and spacer failed almost completely to induce any edits (illustrated as sgPBSMM in Fig. 1c-i).

Using Cas-OFFinder[22], we identified 8 endogenous sites with one or two mismatches to the spacers for 12 pegRNAs and 8,428 endogenous sites with three, four or five mismatches to these spacers (Supplementary Table 2). There are two on-target sites for the *OsCDC48-T1*-pegRNA located at chromosomes 3 and 10. The same

[1]State Key Laboratory of Plant Cell and Chromosome Engineering, Center for Genome Editing, Institute of Genetics and Developmental Biology, Innovation Academy for Seed Design, Chinese Academy of Sciences, Beijing, China. [2]College of Advanced Agricultural Sciences, University of Chinese Academy of Sciences, Beijing, China. [3]State Key Laboratory of Microbial Resources, Institute of Microbiology, Chinese Academy of Sciences, Beijing, China. [4]State Key Laboratory of Plant Genomics, Institute of Microbiology, Chinese Academy of Sciences, Beijing, China. [5]CAS Center for Excellence in Biotic Interactions, University of Chinese Academy of Sciences, Beijing, China. [6]These authors contributed equally: Shuai Jin, Qiupeng Lin, Yingfeng Luo, Zixu Zhu. ✉e-mail: cxgao@genetics.ac.cn

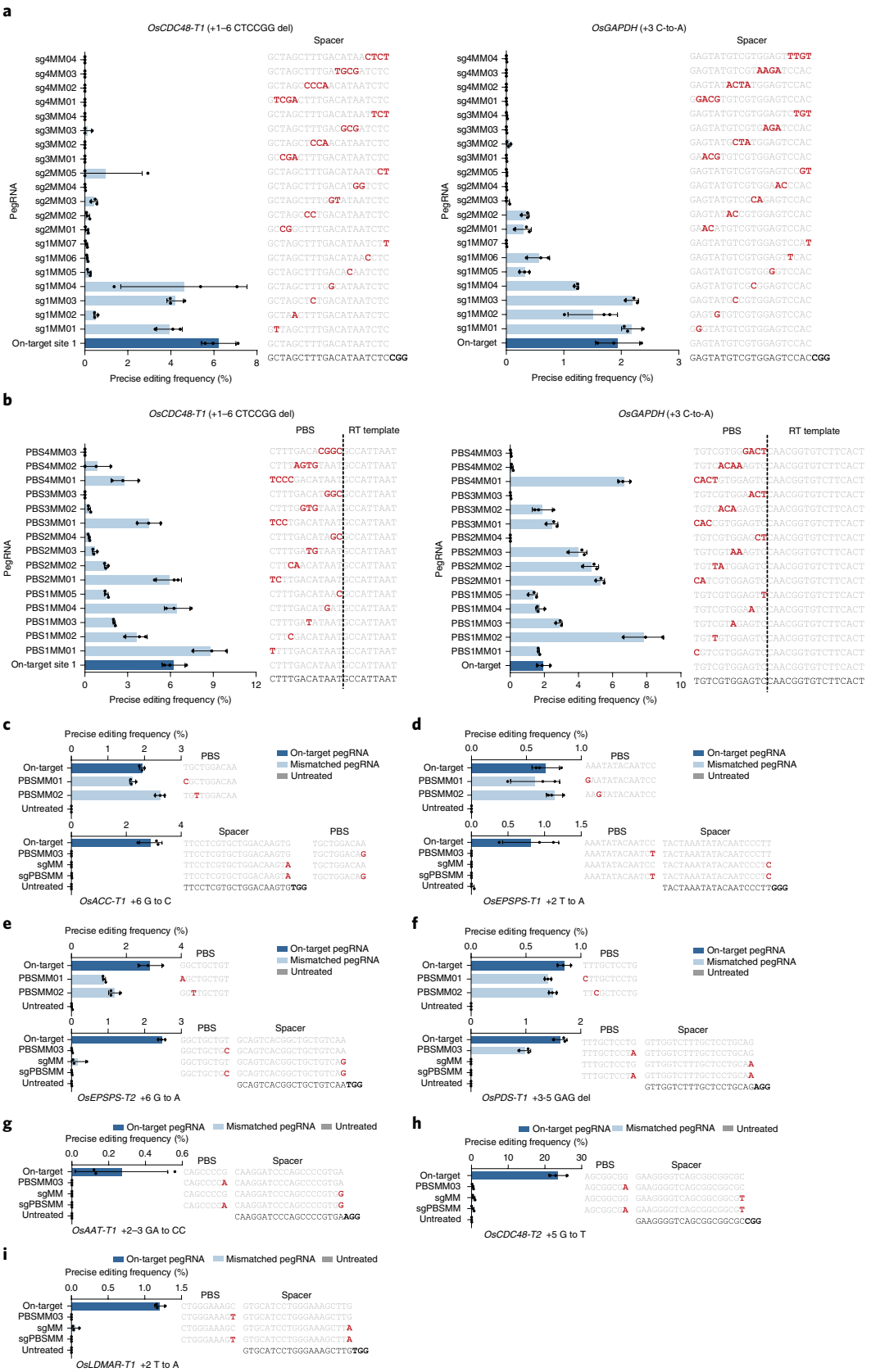


Fig. 1 | Effect of mismatched pegRNA on prime editing in rice protoplasts. a, **b**, Activities of PEs paired with on-target pegRNA or mismatched pegRNA at the target sites *OsCDC48-T1* and *OsGAPDH-T1* in rice protoplasts. Mismatches in the spacer region (**a**) and PBS sequence (**b**) were tested. **c**–**i**, Activities of PEs paired with on-target pegRNA or one base-mismatched pegRNA at seven target sites in rice protoplasts. Mismatched nucleotides are shown in red. The endogenous on-target sequences are shown in black. PAM sequences are shown in bold. Frequencies (means ± s.e.m.) were calculated from three independent experiments ($n = 3$) in **a**–**i**.

pegRNA can not only install (+1−6 CTCCGG del) edit at the target in chromosome 3 (On-target site 1) but also install (+1−7 CTCAGGG del & + T ins) (a combined edit) at the target in chromosome 10 (On-target site 2) (Supplementary Table 2). All 8 sites with one or two mismatches, as well as 171 endogenous sites with three, four or five mismatches, were amplified and sequenced from protoplasts treated with PE2 constructs (Fig. 2 and Supplementary Table 3). Deep sequencing revealed that the PEs induced low levels of editing (0.00~0.02%) at the 8 off-target sites with one or two mismatches (Fig. 2a–h). Editing efficiencies at the 171 off-target sites with three, four or five mismatches were extremely low (0.00~0.23%) (Fig. 2i). Only 3 of the 179 endogenous off-target sites showed significantly higher off-target editing efficiencies than the untreated control. Furthermore, these three off-target editing efficiencies remained extremely low (Supplementary Fig. 2 and Supplementary Table 4). These results are consistent with our finding that the presence of two or more PAM-adjacent spacer mismatches or nick-adjacent PBS mismatches markedly reduces prime editing efficiency.

Genome-wide analysis of pegRNA-dependent off-target effects. To investigate whether overexpression of PEs could induce undesired edits at the genome-wide level, we evaluated their off-target effects in rice by WGS. We used five plant PE constructs[8]—namely *OsALS*-PE3, *OsCDC48*-PE2, *OsCDC48*-PE3, *OsGAPDH*-PE3 and *OsLDMAR*-PE3—targeting the corresponding rice genes (Fig. 3a and Supplementary Table 5). We delivered these PE constructs with pegRNA expression cassettes into rice calli via *Agrobacterium*-mediated transformation (Fig. 3b) and obtained 21 regenerated T0 heterozygous plants with the desired edits, including three independent plants with accompanying byproducts (Supplementary Table 6). The genotypes of these plants were confirmed by polymerase chain reaction (PCR) and Sanger sequencing (Supplementary Fig. 3). These 21 mutant plants were named the PE group (Fig. 3b). Eight regenerated plants that had been transformed with PE vectors without pegRNA expression cassettes were named the PE$^{-pegRNA}$ group (Fig. 3b). In addition, eight plants transformed with a vector expressing only nCas9 (H840A) nickase were used as negative controls and were named the control group (Fig. 3b), whereas five plants (the BE3 group) treated with pH-CBE (a plant codon-optimized version of BE3) vector, which has been reported to induce genome-wide C-to-T off-target edits[12,23], served as positive controls (Fig. 3b). All groups were analyzed by WGS. Thirteen wild-type (WT) plants with the same genetic background were also sequenced to filter background mutations (Fig. 3b). All the plants were sequenced at 50× depth, and, for most of the samples, the mapping ratio (%) of qualified reads exceeded 99.5%, and 95% of the genome was covered by at least 20 non-redundant mapped reads (Supplementary Table 7). Three variant callers—GATK[24], LoFreq[25] and Strelka2[26]—were used to call single-nucleotide variants (SNVs) (Supplementary Fig. 4 and Supplementary Table 8), and two, GATK[27] and Strelka2[26], were used to call insertions/deletions (indels) (Supplementary Fig. 5 and Supplementary Table 9).

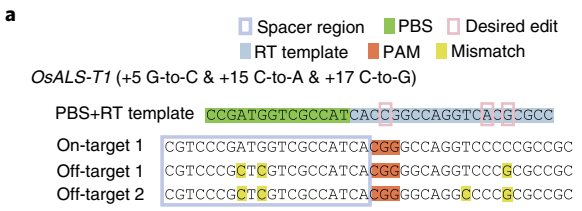
The quality of our variant calling results was confirmed by the fact that all the desired on-target mutations in the 21 edited plants were detected by WGS analysis (Supplementary Table 10). To further check the quality of the calling results, 25 identified variants were randomly selected from the intersection of three variant callers for Sanger sequencing, and all but one of these variants were confirmed (Supplementary Figs. 6 and 7 and Supplementary Table 11). We also assessed the accuracy of the datasets identified by one and two, but not all three, of the variant callers. A total of 63 SNVs and 17 indels were randomly selected for Sanger sequencing, and only 14.3% of the SNVs and 47.1% of the indels not in the intersection of the variant callers were verified (Supplementary Fig. 8). Hence, for the pegRNA-independent off-target analysis, we used the intersection of these calls as true variants to reduce false positives. For detecting pegRNA-dependent off-target effects, false-positive variants can be filtered by the mutation type and by failure to match the RT template. Hence, both the intersection and union of these calls were used for pegRNA-dependent off-target analysis.

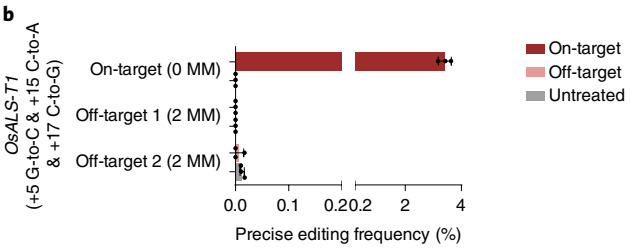
We employed Cas-OFFinder[22] to predict off-target sites (up to five mismatches) in the reference genome and identified 475 potential off-target sites for the five pegRNAs and 1,458 for the four nicking sgRNAs (Fig. 3c,d and Supplementary Table 2). Only two mutations were detected at these sites for the five pegRNAs in the union of these calls, and no mutation was detected in the intersection of these calls (Fig. 3d). We checked these two mutations by Integrative Genomics Viewer[28] and found that they were background mutations with no relation to the RT template sequence and were not located in the RT template region (Supplementary Fig. 9). We found only one on-target non-homologous end joining indel (*OsCDC48*-PE3-T0–3) (Supplementary Fig. 3c,10) but no mutation at 1,458 off-target sites identified by Cas-OFFinder (Supplementary Table 12).

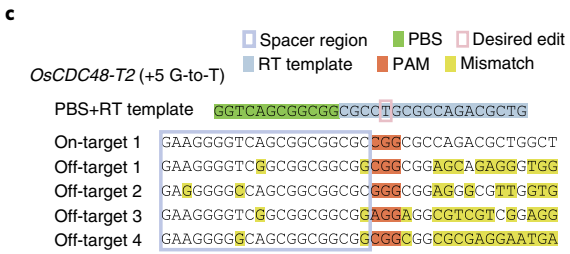
We also envisioned that a different type of off-target event might be induced by pairing between genomic DNA and the PBS and RT template sequences of the pegRNA, which is independent of the spacer sequence. We predicted such pegRNA PBS-RT template sequence-like off-target sites using at least five continuous base pairs and BLAST with default parameters between the PBS-RT template sequence and off-target sites. Using the two strategies, we identified 21,703 and 67 potential sites, respectively. We did not detect any off-target events at all PBS-RT template sequence-like off-target sites, whereas all on-target events were identified (Fig. 2i and Supplementary Tables 13 and 14).

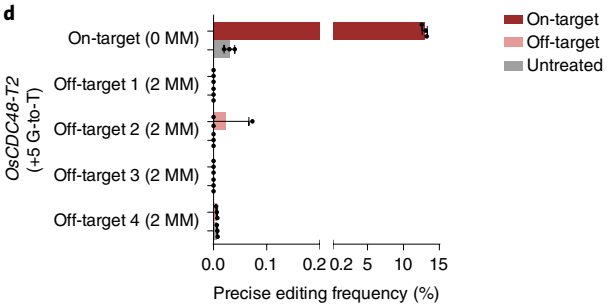
Analysis of genome-wide pegRNA-independent off-target effects. We also investigated the pegRNA-independent off-target effects of PEs using WGS data. We first focused on the ensemble of SNVs that could be identified by WGS analysis across the PE, BE3 and control groups (Fig. 4a). As expected[12,23], significantly more SNVs were identified in the BE3 group than the control group (averages of 547 and 370, respectively; $P=0.0031^{**}$), whereas the number of SNVs in the PE group was not significantly different from that in the control group (averages of 380 and 370, respectively; $P=0.8424$) (Fig. 4a). In addition, total SNVs in the PE groups with and without pegRNAs were not different from the number in the control group (Fig. 4b). When we analyzed the mutation types of these SNVs, the PE group contained similar mutation types as the control group, whereas those in the BE3 group were mainly C-to-T (G-to-A) transitions (average proportion, 44.3%), a significantly higher proportion than in the control group ($P=0.0016^{**}$) (Fig. 4c and Supplementary Fig. 11), in agreement with our previous findings[12,23]. The number of C-to-T SNVs in the PE group did not differ significantly from that in the control group (averages of 105 and 103, respectively; $P=0.8352$), whereas the number in the BE3 group was significantly higher than in the control group (averages of 240 and 103, respectively; $P=0.0016^{**}$) (Fig. 4d). Also, the numbers of C-to-T SNVs in the PE$^{-pegRNA}$ group and the plants prime edited by the various pegRNAs were also similar to the numbers in the control group (Fig. 4e). All these results agree in demonstrating that the PE system does not induce genome-wide off-target SNV changes in plants.

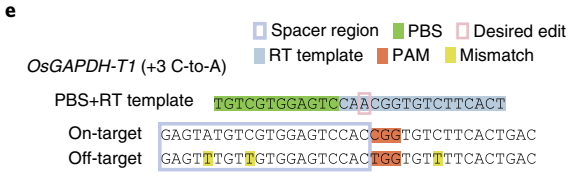
We also investigated whether PEs induce off-target indels (Fig. 4f). To this end, we compared the number of genome-wide indels in the PE, BE3 and control groups (Fig. 4f). The three groups had a similar number of indels (PE, BE3 and control: averages of 142, 159 and 162, respectively; $P=0.7378$ and 0.7545) (Fig. 4f). Numbers of indels in the PE groups with and without pegRNA were also similar to the control group (Fig. 4g). The PE and BE3 groups also contained the same ratio of insertions to deletions as the control group (Fig. 4h and Supplementary Fig. 12). Because BE3 was

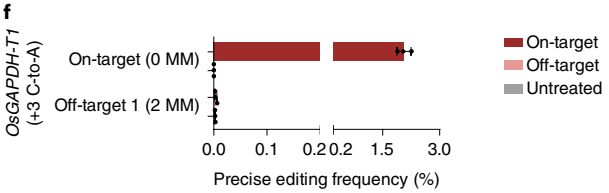


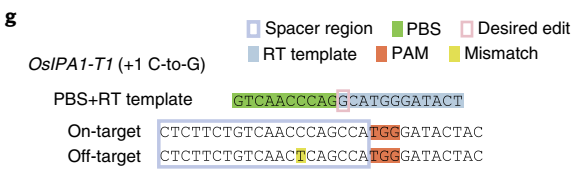


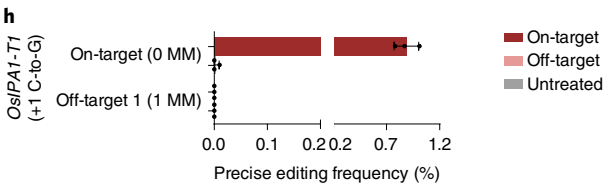












i

OsCDC48-T1

	On-target site		Off-target site																											
	1	2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
Treated	5.82%	4.34%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Untreated	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

	Off-target site																								
	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54
Treated	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Untreated	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

OsGAPDH-T1

	On-target site	Off-target site																												
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
Treated	2.06%	0.00%	0.02%	0.00%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.02%	0.00%	0.00%	0.01%	0.00%	0.00%	0.01%	0.00%	0.00%	0.00%
Untreated	0.00%	0.01%	0.11%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.01%	0.00%	0.00%	0.00%

	Off-target site																						
	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52		
Treated	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.01%	0.00%	0.01%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.03%	0.00%	0.00%		
Untreated	0.00%	0.00%	0.00%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.02%	0.01%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.02%	0.00%	0.00%		

OsACC-T1

	On-target site	Off-target site
		1
Treated	3.56%	0.00%
Untreated	0.00%	0.00%

OsACC-T2

	On-target site	Off-target site				
		1	2	3	4	5	6
Treated	6.23%	0.00%	0.00%	0.14%	0.09%	0.23%	0.15%
Untreated	0.30%	0.00%	0.00%	0.18%	0.12%	0.15%	0.17%

OsACC-T3

	On-target site	Off-target site
		1
Treated	1.52%	0.00%
Untreated	0.00%	0.00%

OsDEP1-T1

	On-target site	Off-target site
		1
Treated	0.28%	0.00%
Untreated	0.01%	0.00%

OsEPSPS-T2

	On-target site	Off-target site
		1
Treated	3.41%	0.02%
Untreated	0.06%	0.02%

OsCDC48-T2

	On-target site	Off-target site																												
	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
Treated	12.96%	0.01%	0.00%	0.00%	0.04%	0.00%	0.01%	0.00%	0.00%	0.01%	0.01%	0.03%	0.01%	0.05%	0.01%	0.01%	0.00%	0.02%	0.00%	0.02%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.02%	0.00%	0.00%
Untreated	0.03%	0.01%	0.01%	0.00%	0.01%	0.00%	0.00%	0.00%	0.01%	0.01%	0.01%	0.05%	0.01%	0.01%	0.03%	0.01%	0.00%	0.01%	0.00%	0.04%	0.00%	0.00%	0.00%	0.00%	0.00%	0.01%	0.00%	0.00%	0.01%	0.00%	0.01%

	Off-target site																				
	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53		
Treated	0.02%	0.00%	0.00%	0.06%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.01%	0.01%		
Untreated	0.00%	0.00%	0.02%	0.00%	0.01%	0.00%	0.00%	0.00%	0.02%	0.00%	0.00%	0.01%	0.00%	0.00%	0.00%	0.02%	0.00%	0.04%	0.01%		

OsCDC48-T3

	On-target site	Off-target site				
		1	2	3	4	5
Treated	13.68%	0.01%	0.06%	0.05%	0.00%	0.00%
Untreated	0.38%	0.00%	0.15%	0.06%	0.00%	0.00%

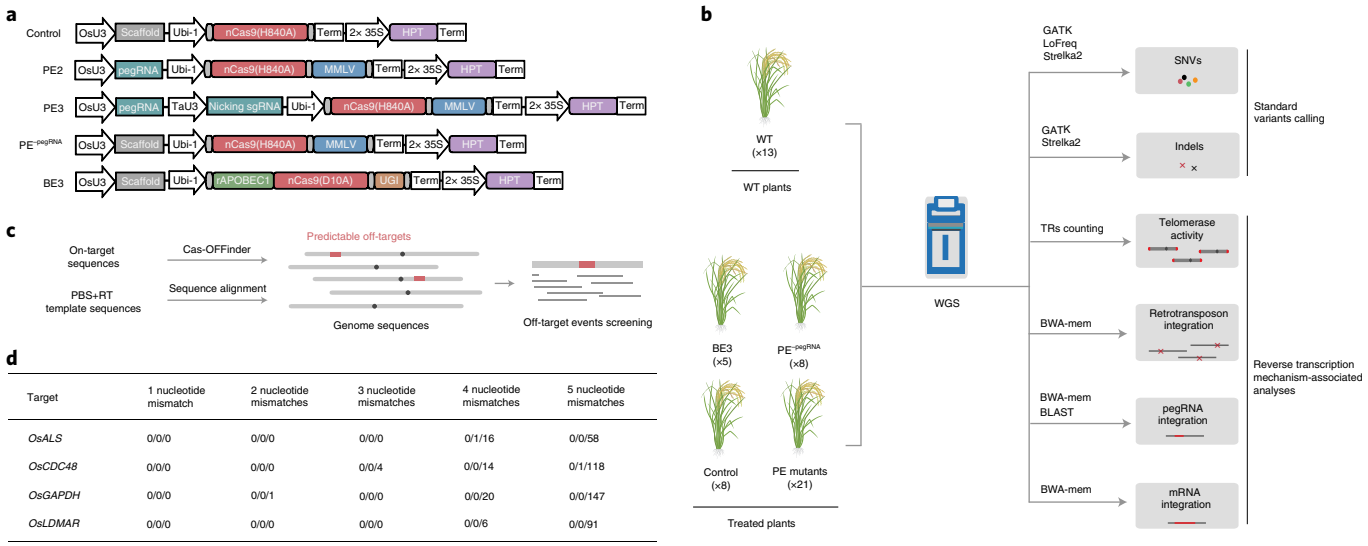
OsCDC48-T4

	On-target site	Off-target site
		1
Treated	2.28%	0.00%
Untreated	0.01%	0.00%

OsIPA1-T1

	On-target site	Off-target site
		2
Treated	0.89%	0.00%
Untreated	0.01%	0.00%

Fig. 2 | Prime editing efficiencies at endogenous on-target and off-target sites. a–**h**, Efficiencies of PEs at endogenous on-target and off-target sites with one or two mismatches in *OsALS-T1* (**a** and **b**), *OsCDC48-T2* (**c** and **d**), *OsGAPDH-T1* (**e** and **f**) and *OsIPA1-T1* (**g** and **h**) in rice protoplasts. Numbers of spacer mismatches are shown in parentheses. Frequencies (means ± s.e.m.) were calculated from three independent experiments ($n = 3$) in **a**–**h**. **i**, Efficiencies of prime editing of 11 pegRNAs at on-target sites and endogenous off-target sites with three, four or five mismatches. Frequencies (means) in **i** were calculated from three independent experiments ($n = 3$).



d

Target	1 nucleotide mismatch	2 nucleotide mismatches	3 nucleotide mismatches	4 nucleotide mismatches	5 nucleotide mismatches
OsALS	0/0/0	0/0/0	0/0/0	0/1/16	0/0/58
OsCDC48	0/0/0	0/0/0	0/0/4	0/0/14	0/1/118
OsGAPDH	0/0/0	0/0/1	0/0/0	0/0/20	0/0/147
OsLDMAR	0/0/0	0/0/0	0/0/0	0/0/6	0/0/91

Fig. 3 | pegRNA-dependent off-target prime editing events analyses via WGS in rice plants. a, Schematic representation of the control, PE2, PE3, PE$^{-pegRNA}$ and BE3 vectors were shown. OsU3, the rice small nuclear RNA *U3* promoter; Ubi-1, maize *ubiquitin-1* promoter; Term, terminator; HPT, hygromycin B phosphotransferase; scaffold, sgRNA scaffold sequence without the spacer sequence or the PBS and RT template sequences. **b**, Experimental design and work flow. Values in parentheses represent numbers of individual plants used in WGS. **c**, Schematic representation of pegRNA spacer sequence-like off-target sites and potential pegRNA PBS-RT template sequence-like off-target edits. The black dots represent centromeres, and the red rectangles represent predictable off-target events identified by Cas-OFFinder or sequence alignment. **d**, Analysis of genome-wide pegRNA spacer sequence-like off-target edits identified by Cas-OFFinder. 0/0/0 represents the number of WGS-identified variants located at off-target sites predicted by Cas-OFFinder in the intersection dataset of the variant callers / the number of WGS-identified variants located in Cas-OFFinder-predicted off-target sites in the union of three variant callers / the number of off-target sites predicted by Cas-OFFinder.

previously shown to preferentially induce off-target edits in transcribed regions, in which single-stranded DNA is exposed[23,29], we tested whether prime editing tended to generate undesired edits in highly transcribed regions; C-to-T variant analysis confirmed this expectation (Fig. 4i–n). We also performed an analysis of the enrichment of the total number of SNVs and indels. This showed that SNVs in genic and highly transcribed regions were enriched only in the BE3 group (Fig. 4i–n). Collectively, these results show that the PE system does not induce detectable genome-wide pegRNA-independent off-target edits in plants.

Because the level of Cas9 protein has been shown to influence the off-target activity of CRISPR editing[16,17], we used western blotting to assess six regenerated PE plants and five control BE3 plants for protein expression levels. In the BE3 group, it appears that the protein levels are correlated with the numbers of C-to-T SNVs (Supplementary Fig. 13a). In contrast, there was no obvious correlation between protein expression level and number of detected SNVs or indels in the PE plants (Supplementary Fig. 13b), even though PE protein levels differed considerably (Supplementary Fig. 13b). This result further supports the conclusion that prime editing does not produce genome-wide off-target effects in plant cells.

Reverse transcription-associated off-target effects. Because MMLV-RT is a core element of the PE system, it was unclear whether overexpressing MMLV-RT could interfere with natural reverse transcription mechanisms in the cell. We evaluated the activities of retrotransposons and telomerase in PE plants (Fig. 5a–d). To obtain more convincing results, we included previously published samples to expand the sample size of the analysis[12,23]. *OsTos17* is an extensively studied retrotransposon in rice and has been shown to be activated during tissue culture[30,31]. There are four copies of *OsTos17* in the genome of rice variety Zhonghua11 used in this study. First, all the raw reads were mapped into the *OsTos17* gene sequence with the BWA-mem program[32]. Then, mapped reads containing the 5′ and 3′ terminal sequences of *OsTos17* were re-mapped into the reference genome to identify their points of insertion in the genome, and their copy numbers were calculated. The AGL1 group, in which the plants went through the transformation process but with no integrated foreign DNA, was used as positive control (Fig. 5a). All the PE, BE3, control and AGL1 groups contained increased numbers of *OsTos17* copies (average increase, 1.3, 1.1, 0.8 and 1.2, respectively) compared to the WT group (average increase, 0.4) (Fig. 5a), confirming that *OsTos17* is activated in the tissue culture process. However, the average number of increased *OsTos17* copies in the PE group was similar to those in the BE3 and AGL1 groups (Fig. 5a), and the pattern of *OsTos17* distribution was not changed in the PE plants in comparison to the plants in the AGL1 group (0.73 and 0.79, respectively) (Supplementary Fig. 14). We further analyzed the fidelity of reverse transcription of *OsTos17* by comparing the error rates in the *OsTos17* region in the different groups (Supplementary Fig. 15). Although error frequencies were extremely low, they were similar in the PE and AGL1 groups (Supplementary Fig. 15). These results suggested that the activity of *OsTos17* was unaffected by ectopic expression of MMLV-RT (Fig. 5a).

We also examined whether the telomerase-mediated process was affected by expressing MMLV-RT. The telomere repeats (TRs) and their variants (Methods) were used as telomere region markers[33], and reads with at least five or at least ten TRs were marked as telomere reads. As expected, the telomere reads were found to be enriched at both ends of rice chromosomes (Fig. 5b), confirming the accuracy of the telomere analysis. The numbers of at least five telomere reads (363.2, 317.7, 326.5, 319.8 and 336.4, respectively) or ten telomere reads (269.2, 249.3, 257.0, 247.6 and 266.8, respectively) per million raw reads for each plant were calculated (Fig. 5c,d) and revealed that all four groups (PE, BE3, WT and control) had similar average numbers of telomere reads compared to the AGL1 group. We also estimated the error rates in telomeric regions for each plant by comparing numbers of TR variants (Supplementary Fig. 16). As expected, the PE group showed a similar number of telomere variants as the AGL1 group (Supplementary Fig. 16). These results

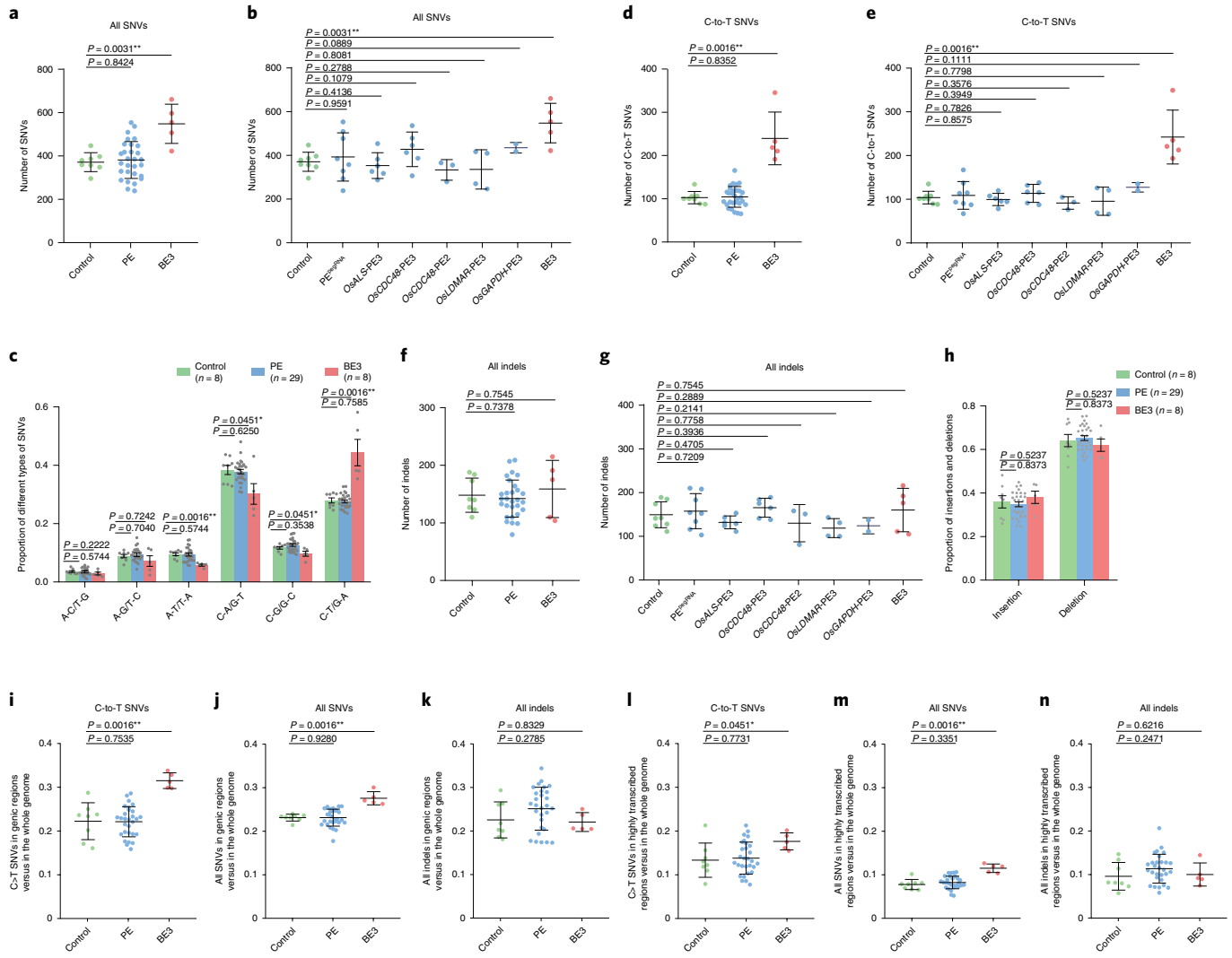


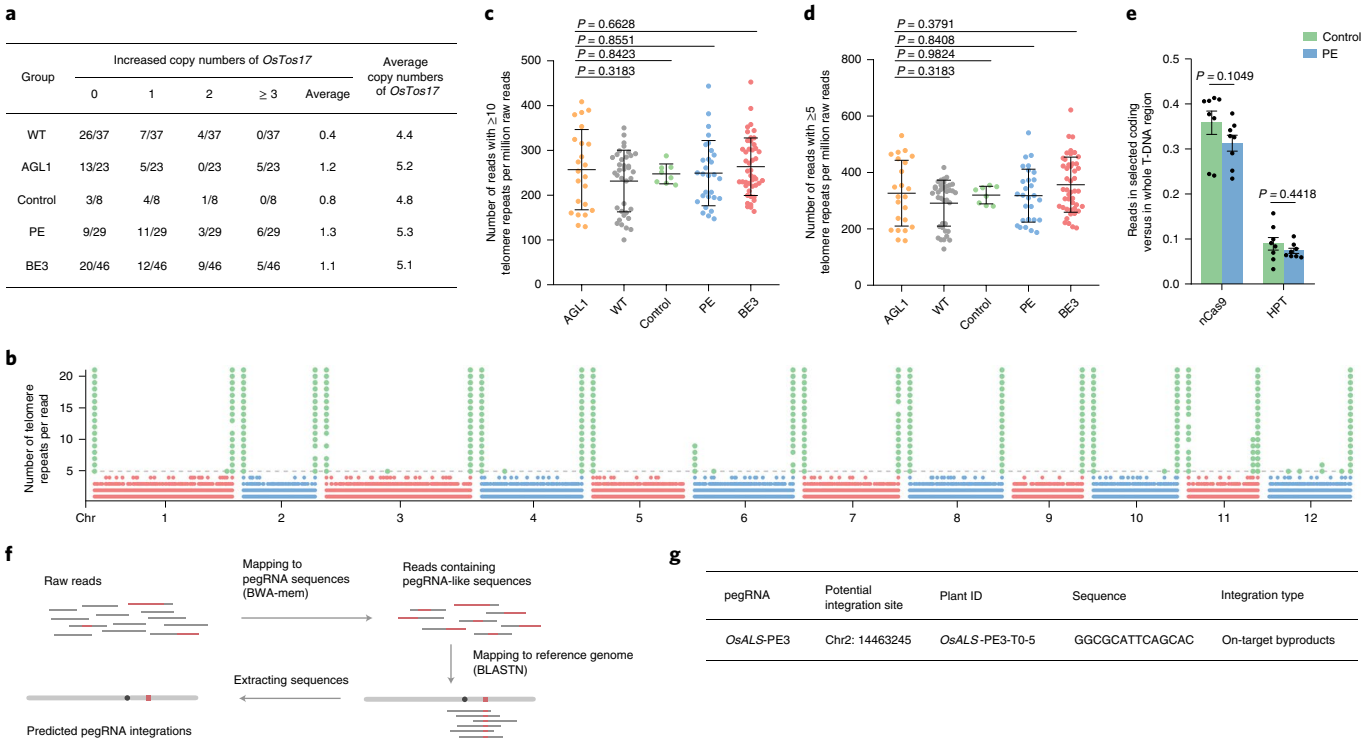
Fig. 4 | pegRNA-independent off-target analyses of prime editing events via WGS in rice plants. a, Numbers of SNVs identified by WGS in the control, PE and BE3 groups. Each dot represents the SNVs from an individual plant. **b**, Comparison of total SNVs in the PE and BE3 groups versus the control group according to target site. **c**, Percentages of the different types of SNV identified in the control, PE and BE3 groups. **d**, Comparison of numbers of C-to-T SNVs in the PE and BE3 groups versus the control group. **e**, Comparison of C-to-T SNVs in the PE and BE3 groups versus the control group according to target site. **f**, Numbers of small indels in the control, PE and BE3 groups. **g**, Comparison of numbers of indels in the PE and BE3 groups and in the control group according to target site. **h**, Proportions of insertions and deletions among the indels identified in the control, PE and BE3 groups. **i**–**k**, Comparison of the proportions of C-to-T SNVs (**i**), total SNVs (**j**) and indels (**k**) located in genic regions versus in the whole genome in the control, PE and BE3 groups. **l**–**n**, Comparison of the proportions of C-to-T SNVs (**l**), total SNVs (**m**) and indels (**n**) located in highly transcribed regions versus in the whole genome in the control, PE and BE3 groups. For samples in **a**, **c**, **d**, **f**, **h** and **i**–**n**, $n = 8$, 29 and 5 in the control, PE and BE3 groups, respectively; for samples in **b**, **e** and **g**, $n = 8$, 8, 6, 6, 3, 4, 2 and 5 in the control, PE^{sgRNA}, *OsALS*-PE3, *OsCDC48*-PE3, *OsCDC48*-PE2, *OsLDMAR*-PE3, *OsDAPDH*-PE3 and BE3 groups, respectively. Data are presented as mean values ± s.d. *P* values were obtained using two-sided Mann–Whitney tests. $^*P < 0.05$ and $^{**}P < 0.01$.

suggested that the activity of telomerase was also unaffected by ectopic expression of the MMLV-RT.

The MMLV-RT is derived from the RNA virus Moloney murine leukemia virus[34,35]; this retrovirus reverse transcribes its RNA genome sequence into DNA for integration into the host genome[34,35]. We envisioned that over-expression of the RT might increase the risk of random reverse transcription of messenger RNAs (mRNAs) and insertion of the resulting products into the rice genome. We checked for such possible off-target effects of PEs by analyzing pegRNA and mRNA insertions. For the mRNA integration analysis, we determined whether MMLV-RT could reverse transcribe highly transcribed mRNAs and integrate their corresponding sequences into the plant genome. Because the nCas9 and HPT (hygromycin B phosphotransferase used for selection) transgene transcripts were driven by constitutive promoters (the maize *ubiquitin-1* promoter and a 2× 35S promoter, respectively)[27,36] (Fig. 3a), we compared the ratios of mapped reads of these two coding regions with those for the whole transfer DNA (T-DNA) region in the PE and control groups (Fig. 5e and Supplementary Fig. 17) and found no significant enrichment for the nCas9 and HPT sequences ($P = 0.1049$ and 0.4418, respectively) (Fig. 5e). For the pegRNA integration analysis, raw reads for each plant were mapped to the various pegRNA sequences (Supplementary Sequences) with the BWA-mem algorithm[32], and any successful hits were mapped to the reference genome (Fig. 5f). Only one pegRNA insertion event was identified in the 21 prime edited plants analyzed (Fig. 5g), and that event could have been due to duplication of the RT template inserted at the target sites (Supplementary Fig. 3).

a

Group	Increased copy numbers of *OsTos17*					Average copy numbers of *OsTos17*
	0	1	2	≥ 3	Average	
WT	26/37	7/37	4/37	0/37	0.4	4.4
AGL1	13/23	5/23	0/23	5/23	1.2	5.2
Control	3/8	4/8	1/8	0/8	0.8	4.8
PE	9/29	11/29	3/29	6/29	1.3	5.3
BE3	20/46	12/46	9/46	5/46	1.1	5.1



g

pegRNA	Potential integration site	Plant ID	Sequence	Integration type
OsALS-PE3	Chr2: 14463245	*OsALS*-PE3-T0-5	GGCGCATTCAGCAC	On-target byproducts

Fig. 5 | Analysis of retrotransposons and telomerase activities in rice. a, Comparison of increased copy numbers of *OsTos17* in the WT, AGL1, control, PE and BE3 groups. **b**, Genome-wide landscape of the distribution of TRs in the Zhonghua11 genome. Each point represents the number of TRs in the reads at different positions. The red and blue points represent different chromosomes, and the green points show reads that include more than five TRs. All the chromosomes have telomere regions at both ends, except for some of poor assembly quality. **c**, **d**, Comparisons of number of reads with more than ten TRs (**c**) and more than five TRs (**d**) per million raw reads in the WT, AGL1, control, PE and BE3 groups. **e**, Comparison of the percentages of reads in the Cas9 and HPT coding regions versus in the whole T-DNA fragment. Proportions of mapped reads in these two coding regions versus in the whole T-DNA region in the control and PE groups. **f**, Schematic representation of the strategy for detecting pegRNA insertions. **g**, Results of the pegRNA insertion analysis. For samples in **c** and **d**, $n = 37$, 8, 45, 29 and 23 in the WT, AGL1, control, PE and BE3 groups, respectively; for samples in **e**, $n = 8$ in both the control and PE groups. Data are presented as mean values ± s.d. *P* values were obtained using two-sided Mann–Whitney tests. $^*P < 0.05$.

Discussion

Rice (*Oryza sativa* L., 2n = 2x = 24) has a relatively small genome (~0.4 Gb) in comparison with the mouse (2.5 Gb) and humans (2.9 Gb)[23]. Hence, more rice individuals can be sequenced at the same cost. In addition, rice is self-pollinated, which reduces the genetic heterogeneity of the progeny of a given population. Moreover, edited plants can be regenerated from calluses derived from a single plant. Therefore, the problem of population heterogeneity is easier to overcome in regenerated rice plants when using WGS. All these factors combined make rice an ideal model organism for assessing the specificity of genome editing tools in higher eukaryotic cells.

In this study, we used rice to evaluate off-target effects and to provide genome-wide insight into the specificity of PEs. WGS-based specificity assessment in edited plants provides an efficient method for comprehensive evaluation of off-target effects in vivo. Nevertheless, WGS of individuals is not as sensitive as other in vitro or cell-level methods (Supplementary Table 15), such as Digenome-seq[20] and Guide-seq[37], for detecting pegRNA-dependent off-target effects. For this reason, we used mismatch pegRNA analysis and examination of endogenous off-target effects in rice protoplasts to complement and support our WGS analyses (Figs. 1 and 2). We found that plant PEs showed tolerance to single mismatch in the 5′ terminal regions of spacer sequences and to multiple mismatches in the 5′ termini of PBSs, and PEs did not induce detectable genome-wide pegRNA-independent off-target SNVs and small indels.

It is conceivable that ectopically expressed MMLV-RT might interfere with endogenous reverse transcription mechanisms in plant cells or lead to various RT-mediated insertion events[34,35]. However, we observed that numbers of copies of the *OsTos17* retrotransposon and of telomere reads were not affected by PEs. It is also possible that MMLV might interact with abundant cellular RNA sequences. It has been previously reported that prime editing did not markedly alter the transcriptome at the whole-cell level[1]. In this study, we examined the possibility of reverse transcription of abundant RNAs (that is, mRNAs and pegRNAs) and subsequent insertion of their complementary DNAs into the rice genome, but no such events were detected, further indicating that the MMLV-RT in PEs does not have non-specific effects in plant cells.

In summary, our data reveal that PEs do not induce detectable pegRNA-independent off-target edits in plants. However, they do generate pegRNA-dependent off-target edits, and the frequency of these might be reduced by engineering the pegRNA and/or Cas9 (refs. [38,39]) or designing pegRNAs with reduced numbers of off-target sites[40–43].

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-021-00891-x.

Received: 28 May 2020; Accepted: 12 March 2021;
Published online: 15 April 2021

References

1. Anzalone, A. V. et al. Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* **576**, 149–157 (2019).
2. Li, G., Liu, Y. G. & Chen, Y. Genome-editing technologies: the gap between application and policy. *Sci. China Life Sci.* **62**, 1534–1538 (2019).
3. Anzalone, A. V., Koblan, L. W. & Liu, D. R. Genome editing with CRISPR–Cas nucleases, base editors, transposases and prime editors. *Nat. Biotechnol.* **38**, 824–844 (2020).
4. Gao, C. Genome engineering for crop improvement and future agriculture. *Cell* **184**, 1621–1635 (2021).
5. Zhu, H., Li, C. & Gao, C. Applications of CRISPR–Cas in agriculture and plant biotechnology. *Nat. Rev. Mol. Cell Bio.* **21**, 661–677 (2020).
6. Liu, Y. et al. Efficient generation of mouse models with the prime editing system. *Cell Discov.* **6**, 27 (2020).
7. Bosch, J. A., Birchak, G. & Perrimon, N. Precise genome engineering in *Drosophila* using prime editing. *Proc. Natl Acad. Sci. USA* **118**, e2021996118 (2021).
8. Lin, Q. et al. Prime genome editing in rice and wheat. *Nat. Biotechnol.* **38**, 582–585 (2020).
9. Veillet, F. et al. Prime editing is achievable in the tetraploid potato, but needs improvement. Preprint at *bioRxiv* https://doi.org/10.1101/2020.06.18.159111 (2020).
10. Jiang, Y. et al. Prime editing efficiently generates W542L and S621I double mutations in two *ALS* genes of maize. *Genome Biol.* **21**, 257 (2020).
11. Wang, L. et al. Spelling changes and fluorescent tagging with prime editing vectors for plants. Preprint at *bioRxiv* https://doi.org/10.1101/2020.07.16.206276 (2020).
12. Jin, S. et al. Cytosine, but not adenine, base editors induce genome-wide off-target mutations in rice. *Science* **364**, 292–295 (2019).
13. Zuo, E. et al. Cytosine base editor generates substantial off-target single-nucleotide variants in mouse embryos. *Science* **364**, 289–292 (2019).
14. Li, J., Hong, S., Chen, W., Zuo, E. & Yang, H. Advances in detecting and reducing off-target effects generated by CRISPR-mediated genome editing. *J. Genet. Genomics* **46**, 513–521 (2019).
15. Tang, X. et al. A large-scale whole-genome sequencing analysis reveals highly specific genome editing by both Cas9 and Cpf1 (Cas12a) nucleases in rice. *Genome Biol.* **19**, 59 (2018).
16. Fu et al. High-frequency off-target mutagenesis induced by CRISPR–Cas nucleases in human cells. *Nat. Biotechnol.* **31**, 822–826 (2013).
17. Hsu, P. D. et al. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* **31**, 827–832 (2013).
18. Mali, P. et al. CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat. Biotechnol.* **31**, 833–838 (2013).
19. Pattanayak, V. et al. High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nat. Biotechnol.* **31**, 839–843 (2013).
20. Kim, D. et al. Digenome-seq: genome-wide profiling of CRISPR–Cas9 off-target effects in human cells. *Nat. Methods* **12**, 237–243 (2015).
21. Kim, D. et al. Genome-wide target specificities of CRISPR RNA-guided programmable deaminases. *Nat. Biotechnol.* **35**, 475–480 (2017).
22. 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).
23. Jin, S. et al. Rationally designed APOBEC3B cytosine base editors with improved specificity. *Mol. Cell.* **79**, 728–740 (2020).
24. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
25. Wilm, A. et al. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res.* **40**, 11189–11201 (2012).
26. Kim, S. et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* **15**, 591–594 (2018).
27. Mitsuhara, I. et al. Efficient promoter cassettes for enhanced expression of foreign genes in dicotyledonous and monocotyledonous plants. *Plant Cell Physiol.* **37**, 49–59 (1996).
28. Robinson, J. T. et al. Integrative Genomics Viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
29. Gaudelli et al. Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. *Nature* **551**, 464–471 (2017).
30. Hirochika, H., Sugimoto, K., Otsuki, Y., Tsugawa, H. & Kanda, M. Retrotransposons of rice involved in mutations induced by tissue culture. *Proc. Natl Acad. Sci. USA* **93**, 7783–7788 (1996).
31. Piffanelli, P. et al. Large-scale characterization of Tos17 insertion sites in a rice T-DNA mutant library. *Plant Mol. Biol.* **65**, 587–601 (2007).
32. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
33. Mizuno, H. et al. Characterization of chromosome ends on the basis of the structure of TrsA subtelomeric repeats in rice (*Oryza sativa* L.). *Mol. Genet. Genomics* **280**, 19–24 (2008).
34. Demeulemeester, J., De Rijck, J., Gijsbers, R. & Debyser, Z. Retroviral integration: site matters: mechanisms and consequences of retroviral integration site selection. *Bioessays* **37**, 1202–1214 (2015).
35. Tsuruyama, T., Hiratsuka, T. & Yamada, N. Hotspots of MLV integration in the hematopoietic tumor genome. *Oncogene* **36**, 1169–1175 (2017).
36. Toki, S. et al. Expression of a maize ubiquitin gene promoter-*bar* chimeric gene in transgenic rice plants. *Plant Physiol.* **100**, 1503–1507 (1992).
37. Tsai, S. Q. et al. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR–Cas nucleases. *Nat. Biotechnol.* **33**, 187–197 (2015).
38. Tycko, J., Myer, V. E. & Hsu, P. D. Methods for optimizing CRISPR–Cas9 genome editing specificity. *Mol. Cell* **63**, 355–370 (2016).
39. Schmid-Burgk, J. L. et al. Highly parallel profiling of Cas9 variant specificity. *Mol. Cell* **78**, 794–800 (2020).
40. Doench, J. G. et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR–Cas9. *Nat. Biotechnol.* **34**, 184–191 (2016).
41. Chuai, G. H., Wang, Q. L. & Liu, Q. In silico meets in vivo: towards computational CRISPR-based sgRNA design. *Trends Biotechnol.* **35**, 12–21 (2017).
42. Kim, N. et al. Prediction of the sequence-specific cleavage activity of Cas9 variants. *Nat. Biotechnol.* **38**, 1328–1336 (2020).
43. Liu, X. et al. SeqCor: correct the effect of gRNA sequences in CRISPR/Cas9 screenings by machine learning algorithm. *J. Genet. Genomics* https://doi.org/10.1016/j.jgg.2020.10.007 (2020).

Methods

Plasmid construction. The pegRNA constructs were made as reported previously[8] by amplifying the spacer-scaffold-RT-PBS fragment by one-step PCR and cloning it into OsU3-sgRNA vectors using a ClonExpress II One Step Cloning Kit (Vazyme Biotech). The binary vectors were used for *Agrobacterium*-mediated transformation. The PE and BE3 plant expression vectors used were pH-CBE[12] and pH-nCas9-PPE[8]. To construct the pH-nCas9 (H840A) vector for the control group, the M-MLV sequence of pH-nCas9-PPE was removed using a ClonExpress II Cloning Kit. The PE3 constructs were made as reported previously[8]. To construct PE2, the pegRNA expression cassette was integrated into *Hind* III-digested pH-nCas9-PPE by ClonExpress II One Step Cloning. PCR was performed using TransStart FastPfu DNA Polymerase (TransGen Biotech). The primers used in this study were synthesized by the Beijing Genomics Institute (Supplementary Data 1).

PEG-mediated transformation of rice protoplasts. Zhonghua11 (a Japonica rice variety) was used for protoplast isolation and transformation. Transformation of isolated rice protoplasts was described previously[44]; plasmids were introduced by PEG-mediated transfection. The mean transformation efficiency was 28–45%. Transfected protoplasts were incubated at 26 °C for 48 h. After incubation, genomic DNA was extracted with a DNAquick Plant System (Tiangen Biotech) and used for deep amplicon sequencing. The transformation efficiencies are normally arranged from 30% to 50% as analyzed by flow cytometry[8,23,44,45]. For each transformation, about 5×10^5 edited cells are sequenced[45].

***Agrobacterium*-mediated transformation of rice callus cells.** DNA of binary vector plasmids (plasmids with *cis*-acting T-DNA border sequences and *trans*-acting virulence function (*vir*) genes in two separate replicons) was introduced into the *Agrobacterium tumefaciens* strain AGL1 by electroporation (1 μg per transformation). *Agrobacterium*-mediated transformation of callus cells of Zhonghua11 was performed as reported previously[44,46]. Hygromycin (50 μg ml^{-1}) was used to select transgenic plants.

Flow cytometry analysis. An FACSAria III (BD Biosciences) was used for flow cytometry as previously reported[45]. Rice protoplasts were transfected with pegRNA expression plasmids, fluorophore expression plasmids and PE expression plasmids. The percentage of green fluorescent protein-positive cells was calculated for each sample. Gating of all samples can be found in Supplementary Data 2.

DNA extraction. A leaf was removed from each plant (10 × 5 mm) after 4–5 weeks of regeneration, and genomic DNA was extracted with a DNAquick Plant System (Tiangen Biotech). The extracted genomic DNA was quantified with a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific).

Deep sequencing. Target regions were amplified from protoplast genomic DNA with site-specific primers using nested PCR. Forward and reverse barcodes for library construction were added to the ends of the PCR products in the second-round PCR. Equal amounts of PCR product were pooled and sequenced commercially (Novogene) using the NovaSeq platform, and the pegRNA target sites in the sequenced reads were examined for desired edits and indels. Amplicon sequencing was repeated three times for each target site using genomic DNA extracted from three independent protoplast samples. Prime editing processivity and indels were analyzed as previously described[8].

Prediction of pegRNA spacer-like off-target edits. PegRNA spacer-like off-target sites were predicted with an offline version of Cas-OFFinder[22]. The high-quality Zhonghua11 genome was used as reference genome[23]. The maximum mismatch was set at five.

Prediction of pegRNA PBS-RT template-like off-target edits. We assessed the specificity of pegRNA PBS-RT template-like based on sequence alignment by two strategies. The first strategy was based on BLAST (v2.2.25). The PBS-RT sequences (Supplementary Sequences) were aligned with the reference genome using BLAST (v2.2.25) with the default parameters. For the second strategy, we randomly selected five continuous base pairs in PBS-RT sequence to simulate as 'PBS seed sequence' and compared the 5-nt flanking of the sequence at the variants identified by WGS. If the PBS seed sequence was identical to either side of the flank sequence, and the variation and the other side of the flank sequence could also match the 5-nt RT template sequence, it would be treated as a PBS-RT-dependent off-target event.

Sanger sequencing. PCR and Sanger sequencing was used to verify on-target mutations, pegRNA-dependent off-target mutations, and mutants identified by WGS analysis. National Center for Biotechnology Information (NCBI) primer blast was used to design specific primers (https://www.ncbi.nlm.nih.gov/tools/primer-blast/). Target sequences were amplified with 2× Rapid Taq Master Mix (Vazyme Biotech), and amplicons were purified with an EasyPure PCR Purification Kit (TransGen Biotech) and sequenced by the Beijing Genomics Institute.

WGS and variant calling. A total of 55 plants, including 13 WT, 8 control, 5 BE3, 8 PE-sgRNA and 21 PE plants with pegRNA-induced mutations, were used to analyze genome-wide variants. They were sequenced using a NovaSeq platform (Novogene). An average of 22 Gb of data (~50×) was generated per plant. Raw reads were processed with BIGpre (v2.0.2)[47] and mapped to the Zhonghua11 assembly with BWA-mem (v0.7.15)[32]. Picard (v1.119) (https://github.com/broadinstitute/picard) was used to mark duplicated reads, which were then realigned by the RealignerTargetCreator and IndelRealigner modules in GATK (v3.8.1)[24]. Three variant callers—GATK (v4.1.6.0)[24], LoFreq (v2.1.2)[25] and Strelka2 (v2.9.10)[26]—were used to call the variants with default parameters. To obtain the possible variants more extensively and accurately, we used the HaplotypeCaller module to integrate candidate variants of all individuals but not the previous population-based UnifiedGenotyper module[24]. We uploaded more notes about this method to https://github.com/ReiGao/GWSBE/blob/master/Script/globCallUseGATK.pl. Genomic regions with depth >100 or <20, and genomic regions overlapping with background variants from the WT plants, were excluded. Indels with ≥30% alternate (ALT) allele frequencies were used for further analysis. The intersections of the three programs (GATK/LoFreq/Strelka2) were considered high-confidence SNVs, whereas the intersections of two programs (GATK/Strelka2) were considered high-confidence indels. A simple binomial probability calculation showed that only mutations occurring at frequencies >5.9% would be detected with >95% probability at a WGS coverage of 50×. We checked our WGS results and found only one indel variant occurring at a frequency <5.9% (Supplementary Table 16); this variant was filtered out and did not show up in our high-confidence indels when 'Indels with ≥30% ALT allele frequencies' were used in further analysis.

RNA-seq data analysis. Leaves of WT plants were used for RNA extraction and RNA sequencing (RNA-seq) library construction. After sequencing by NovaSeq, the raw reads were processed with BIGpre (v2.0.2) and mapped to the Zhonghua 11 assembly using HISAT2 (v2.0.4)[48]. Samtools was used to sort the SAM format alignment results and output base resolution depth information. Genes for which >10% of the exonic regions were covered by at least 100 RNA-seq reads were considered to be highly transcribed genes.

Identification of Tos17 insertion events. The sequence of *OsTos17* (NCBI accession numbers AP005292 and AE017097) was used to locate the positions of *OsTos17* insertions in the Zhonghua11 genome using BLAST (v2.2.25), and the results showed that the Zhonghua11 genome has four copies of *OsTos17* (chr1:19,357,676–19,361,790, chr2:985,831–989,945, chr7:28,258,047–28,262,161 and chr10:15,748,265–15,752,468). Paired-end raw reads of WGS data were mapped to the *OsTos17* sequence using BWA-mem (v0.7.15), and the overhanging reads (reads partially mapped to the terminal end of *OsTos17* or unmapped paired reads with one read completely mapped to *OsTos17*) were mapped to the genome using BLASTN. Events with at least four overhanging reads mapped within 500-bp genomic regions were taken as *OsTos17* insertion events. Insertions into all four copies of *OsTos17* in the reference genome could be detected in all samples with this pipeline. The error rates of *OsTos17* were analyzed by Samtools 'stats' command.

Detection of telomere reads. The TRs (TTTAGGG) and their variants (ATTAGGG, CTTAGGG, GTTAGGG, TATAGGG, TTCAGGG and TTGAGGG) were taken as rice telomere markers. For the Zhonghua11 genome, the sequence of each chromosome was divided into overlapping 150-nt 'reads' (window size = 150 nt; step size = 10 nt), and numbers of TRs and their variants were counted for each 'read' and positionally plotted across the chromosome. For raw reads of each sample, the numbers of telomere reads (with ≥5 or ≥10 TRs) were counted per million raw reads.

Analysis of pegRNA insertion. Raw reads of each plant were mapped to the pegRNA sequence using BWA-mem (v0.7.15). Then, the mapped reads were mapped onto the reference genome to identify their insertion points using BLAST (v2.2.25).

Analysis of mRNA integration. The T-DNA sequences between the left border and right border in all PE vectors and control vectors include nCas9 and HPT expression cassettes (Fig. 1a). These expression cassettes are driven by two constitutive promoters: maize *ubiquitin-1* and 2× 35S, respectively. The ratios of reads mapped to these two coding regions to reads in the whole T-DNA region were compared in the control and PE groups.

Protein extraction and immunoblotting. Rice tissue and protoplasts were thoroughly ground in liquid nitrogen, and total protein was extracted as previously described[49]. Immunoblotting was performed by standard procedures. Antibodies used were as follows: anti-plant actin (ABclonal, cat. no. AC009, 1:2,000 dilution), anti-Cas9 (Millipore, cat. no. MAC133, 1:2,000 dilution) and peroxidase-conjugated goat anti-mouse IgG secondary antibody (Sigma-Aldrich, cat. no. A4416, 1:10,000 dilution).

T-A cloning and sequencing. The extracted rice genomic DNA was PCR amplified and purified. Then, the purified PCR products were ligated into *pEASY*-Blunt vector (TransGen Biotech) by T4 DNA ligase. Next, the recombinant plasmid

DNA was transformed into *Escherichia coli*-competent cells and cultivated in ampicillin-resistant LB medium (100 μg ml^{-1}) at 37 °C for 16 h. The monoclonal colonies were selected and sequencing to confirm their genotype. Ampicillin (100 μg ml^{-1}) was added into the LB medium.

Statistical analysis. All numerical values are presented as means ± s.d. Significant differences between controls and treatments were tested using the two-sided Mann–Whitney test. $P < 0.05$ was considered statistically significant, and $P < 0.01$ was considered statistically extremely significant.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All data supporting the findings of this study are available in the article and supplementary figures and tables or are available from the corresponding author upon reasonable request. For sequence data, rice LOC_Os identifiers (http://rice.plantbiology.msu.edu/) are as follows: LOC_Os03g54790 (*OsALS*), LOC_Os03g05730 (*OsCDC48*), LOC_Os08g03290 (*OsGAPDH*), LOC_Os01g55540 (*OsAAT*), LOC_Os05g22940 (*OsACC*), LOC_Os09g26999 (*OsDEP1*), LOC_Os06g04280 (*OsEPSPS*), LOC_Os08g39890 (*OsIPA1*), LOC_Os08g03290 (*OsGAPDH*) and LOC_Os03g08570 (*OsPDS*). The NCBI GenBank identifiers are AP005292 and AE017097 (*OsTos17*). The deep sequencing and genome sequencing data have been deposited in two NCBI BioProject databases (accession codes PRJNA702625 and PRJNA636219). Source data are provided with this paper.

References

44. Shan, Q. et al. Targeted genome modification of crop plants using a CRISPR–Cas system. *Nat. Biotechnol.* **31**, 686–688 (2013).
45. Zong, Y. et al. Efficient C-to-T base editing in plants using a fusion of nCas9 and human APOBEC3A. *Nat. Biotechnol.* **36**, 950–953 (2018).
46. Shan, Q. et al. Rapid and efficient gene modification in rice and *Brachypodium* using TALENs. *Mol. Plant* **6**, 1365–1368 (2013).
47. Zhang, T. et al. BIGpre: a quality assessment package for next-generation sequencing data. *Genom. Proteom. Bioinf.* **9**, 238–244 (2011).
48. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
49. Su, L., Li, A., Li, H., Chu, C. & Qiu, J. L. Direct modulation of protein level in *Arabidopsis*. *Mol. Plant* **6**, 1711–1714 (2013).

Acknowledgements

This work was supported by grants from the National Natural Science Foundation of China (31788103 and 31971370), the National Key Research and Development Program of China (2016YFD0100602), the Strategic Priority Research Program of the Chinese Academy of Sciences (Precision Seed Design and Breeding, XDA24020100), the Chinese Academy of Sciences (QYZDY-SSW-SMC030) and the Youth Innovation Promotion Association of the Chinese Academy of Sciences (2017140).

Author contributions

C.G. supervised the project. C.G., S.J. and Y.F.L. designed the experiment. S.J., Q.L., Z.Z., G.L. and K.C. performed the experiments. Y.F.L., S.J. and Y.J.L. performed the bioinformatics analyses. C.G., J.-L.Q., S.J. and Q.L. wrote the manuscript.

Competing interests

The authors declare no competing financial interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-021-00891-x.

Correspondence and requests for materials should be addressed to C.G.

Peer review information *Nature Biotechnology* thanks Nicole Gaudelli and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

nature research

Corresponding author(s): Caixia Gao

Last updated by author(s): Mar 11, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Illumina NovaSeq platform was used to collect the WGS sequencing data.
Data analysis	Graphpad prism 7 was used to analyze the data. The custom perl script to analyze whole genome sequencing data will be made available upon request.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The authors declare that all data supporting the findings of this study are available in the article and its Supplementary Information files or are available from the corresponding author on request. Datasets of whole genome sequencingexperiments can be deposited with the National Center for Biotechnology Information (NCBI) before publication.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	A total of 55 plants, including 13 wide-type (WT) plants, eight Control plants, five BE3 plants, eight PE without pegRNA plants and 21 PE with pegRNA mutants were used to analysis the genome-wide variants.
Data exclusions	No data exclusion.
Replication	All attempts for replication were successful.
Randomization	All plants were regenerated from one group of callus.
Blinding	Not applicable. As samples were processed identically through standard and in some cases automated procedures (DNA sequencing, transformation, DNA isolation) that should not bias outcomes.

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	*Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study).*
Research sample	*State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For studies involving existing datasets, please describe the dataset and source.*
Sampling strategy	*Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed.*
Data collection	*Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection.*
Timing	*Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.*
Data exclusions	*If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Non-participation	*State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.*
Randomization	*If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.*

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	*Briefly describe the study. For quantitative data include treatment factors and interactions, design structure (e.g. factorial, nested, hierarchical), nature and number of experimental units and replicates.*
Research sample	*Describe the research sample (e.g. a group of tagged Passer domesticus, all Stenocereus thurberi within Organ Pipe Cactus National Monument), and provide a rationale for the sample choice. When relevant, describe the organism taxa, source, sex, age range and any manipulations. State what population the sample is meant to represent when applicable. For studies involving existing datasets,*

describe the data and its source.

Sampling strategy	*Note the sampling procedure. Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.*
Data collection	*Describe the data collection procedure, including who recorded the data and how.*
Timing and spatial scale	*Indicate the start and stop dates of data collection, noting the frequency and periodicity of sampling and providing a rationale for these choices. If there is a gap between collection periods, state the dates for each sample cohort. Specify the spatial scale from which the data are taken*
Data exclusions	*If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Reproducibility	*Describe the measures taken to verify the reproducibility of experimental findings. For each experiment, note whether any attempts to repeat the experiment failed OR state that all attempts to repeat the experiment were successful.*
Randomization	*Describe how samples/organisms/participants were allocated into groups. If allocation was not random, describe how covariates were controlled. If this is not relevant to your study, explain why.*
Blinding	*Describe the extent of blinding used during data acquisition and analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.*

Did the study involve field work? ☐ Yes ☐ No

Field work, collection and transport

Field conditions	*Describe the study conditions for field work, providing relevant parameters (e.g. temperature, rainfall).*
Location	*State the location of the sampling or experiment, providing relevant parameters (e.g. latitude and longitude, elevation, water depth).*
Access and import/export	*Describe the efforts you have made to access habitats and to collect and import/export your samples in a responsible manner and in compliance with local, national and international laws, noting any permits that were obtained (give the name of the issuing authority, the date of issue, and any identifying information).*
Disturbance	*Describe any disturbance caused by the study and how it was minimized.*

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	The anti-plant actin (abclonal, Cat#AC009, 1:2000 dilution), anti-Cas9 (millipore, Cat#MAC133, 1:2000 dilution), and peroxidase-conjugated goat anti-mouse IgG secondary antibody (Sigma, Cat# A4416, 1:10 000 dilution) are used in this study.
Validation	All reagents used were commercially available.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	*State the source of each cell line used.*

Authentication	*Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.*
Mycoplasma contamination	*Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.*
Commonly misidentified lines (See ICLAC register)	*Name any commonly misidentified cell lines used in the study and provide a rationale for their use.*

Palaeontology

Specimen provenance	*Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information).*
Specimen deposition	*Indicate where the specimens have been deposited to permit free access by other researchers.*
Dating methods	*If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.*

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	*For laboratory animals, report species, strain, sex and age OR state that the study did not involve laboratory animals.*
Wild animals	*Provide details on animals observed in or captured in the field; report species, sex and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.*
Field-collected samples	*For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.*
Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about studies involving human research participants

Population characteristics	*Describe the covariate-relevant population characteristics of the human research participants (e.g. age, gender, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."*
Recruitment	*Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.*
Ethics oversight	*Identify the organization(s) that approved the study protocol.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	*Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.*
Study protocol	*Note where the full trial protocol can be accessed OR if not available, explain why.*
Data collection	*Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.*
Outcomes	*Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.*

ChIP-seq

Data deposition

☐ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	*For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.*
Files in database submission	*Provide a list of all files available in the database submission.*
Genome browser session (e.g. UCSC)	*Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.*

Methodology

Replicates	*Describe the experimental replicates, specifying number, type and replicate agreement.*
Sequencing depth	*Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.*
Antibodies	*Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Peak calling parameters	*Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.*
Data quality	*Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.*
Software	*Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.*

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Rice protoplasts were isolated from the stem of rice seedlings, transfected as described in the Mehtods and incubated in 1 ml WI solution for 2 days.
Instrument	BD FACSAriaIII
Software	FACSDiva Version 6.1.3 software was used for analysis.
Cell population abundance	The abundance of cells for flow cytometry analysis was 10,000 for each sample.
Gating strategy	Negative control (untreated) and fluorophore-positive cells were used to establish gates for each cell type. a were drawn to collect cells expressing either fluorophore. See the provided examples for gates used.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type	*Indicate task or resting state; event-related or block design.*

Design specifications

Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.

Behavioral performance measures

State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).

Acquisition

Imaging type(s)

Specify: functional, structural, diffusion, perfusion.

Field strength

Specify in Tesla

Sequence & imaging parameters

Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.

Area of acquisition

State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.

Diffusion MRI ☐ Used ☐ Not used

Preprocessing

Preprocessing software

Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).

Normalization

If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.

Normalization template

Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.

Noise and artifact removal

Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).

Volume censoring

Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.

Statistical modeling & inference

Model type and settings

Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).

Effect(s) tested

Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference
(See Eklund et al. 2016)

Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.

Correction

Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).

Models & analysis

n/a	Involved in the study
☐	☐ Functional and/or effective connectivity
☐	☐ Graph analysis
☐	☐ Multivariate modeling or predictive analysis

Functional and/or effective connectivity

Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).

Graph analysis

Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).

Multivariate modeling and predictive analysis

Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.