


# Genome-wide prediction of activating regulatory elements in rice by combining STARR-seq with FACS

Wei Tian, Xi Huang and Xinhao Ouyang\* 

State Key Laboratory of Cellular Stress Biology, School of Life Sciences, Faculty of Medicine and Life Sciences, Xiamen University, Xiamen, China

Received 19 April 2022;

revised 23 July 2022;

accepted 3 August 2022.

\*Correspondence (Tel +86-0592-2182563;

fax +86-0592-2181015; email

ouyangxinhao@xmu.edu.cn)

## Summary

Self-transcribing active regulatory region sequencing (STARR-seq) is widely used to identify enhancers at the whole-genome level. However, whether STARR-seq works as efficiently in plants as in animal systems remains unclear. Here, we determined that the traditional STARR-seq method can be directly applied to rice (*Oryza sativa*) protoplasts to identify enhancers, though with limited efficiency. Intriguingly, we identified not only enhancers but also constitutive promoters with this technique. To increase the performance of STARR-seq in plants, we optimized two procedures. We coupled fluorescence activating cell sorting (FACS) with STARR-seq to alleviate the effect of background noise, and we minimized PCR cycles and retained duplicates during prediction, which significantly increased the positive rate for activating regulatory elements (AREs). Using this method, we determined that AREs are associated with AT-rich regions and are enriched for a motif that the AP2/ERF family can recognize. Based on GC content preferences, AREs are clustered into two groups corresponding to promoters and enhancers. Either AT- or GC-rich regions within AREs could boost transcription. Additionally, disruption of AREs resulted in abnormal expression of both proximal and distal genes, which suggests that STARR-seq-revealed elements function as enhancers *in vivo*. In summary, our work provides a promising method to identify AREs in plants.

**Keywords:** STARR-seq, FACS, protoplast, enhancer, promoter.

## Introduction

Core promoters define the transcription start sites and typically support basal transcription. Activating regulatory elements (AREs) harbouring transcription factors can influence the recruitment of RNA polymerase II in core promoters and thereby modulate transcription rates. The activating signals of AREs may be either local or distal with respect to core promoter genomic proximity. If the ARE and core promoter are close to each other, the region encompassing both would be defined as the 'promoter'. Regulation can also be achieved by integrating signals from a distal ARE, which would then be referred to as an 'enhancer'.

Interaction between promoters and enhancers allows exquisite control of increases in transcription. For instance, two upstream enhancers containing 'CCAAT' motifs associate with the promoter of *FLOWERING LOCUS T* to boost transcription during floral initiation (Cao *et al.*, 2014). In addition, jasmonate enhancers exert different effects on *MYC2* expression during short- or long-term jasmonate responses (Wang *et al.*, 2019). As enhancers play a critical role in the regulation of gene expression, they also could be valuable for agronomic applications. However, only a handful of enhancers have been characterized in plants.

In one assessment of enhancers co-localizing with DNase I hypersensitive sites, 10 044 putative enhancers were predicted in *Arabidopsis* (Zhu *et al.*, 2015). Notably, ten of 14 candidates (71.4%) were validated using a  $\beta$ -glucuronidase gene reporter assay, which supports the applicability of the open-signature-based approach in plants. However, this approach limits prediction to intergenic regions, which could overlook many enhancers. To overcome this limitation, a whole-genome prediction was carried out in rice (Sun *et al.*, 2019) using self-transcribing active

regulatory region sequencing (STARR-seq; Arnold *et al.*, 2013), a massively parallel reporter assay to screen potential enhancers based on the principle that enhancers function independently of their positions, distance and orientations to the core promoters of target genes. The analysis results suggested that, contrary to traditional consensus, rice might have unique features in enhancers. For example, enhancers in rice showed pronounced enrichment for repetitive sequences. A later application of STARR-seq (Jores *et al.*, 2020) in *Nicotiana benthamiana* demonstrates that the traditional method which places candidates in the 3' untranslated region (3'UTR) suffers from a relatively low signal-to-noise ratio and therefore is not suitable for direct use in plants. That group modified the approach by inserting enhancers upstream of the core promoter and linking them with unique barcodes. The new version worked well to identify reported enhancers and important cis-elements. However, linking sequences with barcodes is complex and limits genome-wide usage.

The conflicting reports attracted our interest in applying STARR-seq in plants. Here, we confirmed that the traditional method for STARR-seq can be used in rice protoplasts, although with unsatisfactory performance. We optimized the protocol by carrying out fluorescence activating cell sorting (FACS) to separate GFP-positive cells and performing predictions covering duplicates. As a result, our improved method predicted AREs more precisely.

## Results

### Traditional STARR-seq works in rice protoplasts

We revisited whether the traditional STARR-seq method can be directly used in plants. A vector (Figure S1) with an expression

cassette including a minimal 35S promoter, the first intron of castor bean catalase gene (*Cat1*), an *EGFP* gene, a multiple cloning site and a nopaline synthase gene (*NOS*) terminator was generated, which shows the similar structure with reported (Jores *et al.*, 2020; Sun *et al.*, 2019). We selected four STARR-seq-predicted enhancers (SPEs; Sun *et al.*, 2019) that appeared in both replicates with fold enrichment (FE) ranging from 1.3 to 2.3 in replicate 1 to test their ability to boost gene expression in rice protoplasts (Figure 1a). The positive control constructs generated strong fluorescence signals, unlike the negative control constructs; among the four SPEs, only SPEd generated fluorescence signals; SPEd could activate gene expression in both orientations, although the reverse-oriented construct showed much weaker strength (Figure S2). These findings demonstrated that our vector successfully worked in rice protoplast system and indicated that traditional STARR-seq can be directly used in plants to identify enhancers.

To further test this conclusion, a total of 26 predicted enhancers present in both reported replicates and within the remaining top 100 in replicate 1 were cloned for validation (Figure 1a). For an estimation of activating strength, flow cytometry was used to count fluorescent cells (Figure S3). It should be noted that intact cells could not be distinguished from cell debris based on cell size in our case, and therefore both of the events were recorded. Based on fluorescence signals and fluorescent cell numbers, the candidates were divided into four classes: CI (defined as generating fluorescence signals, average positive cells >50/20 000 events and  $P < 0.05$  compared with negative control), CII (defined as generating fluorescence signals, 50/20 000 events > average positive cells >10/20 000 events and  $P < 0.05$  compared with negative control), CIII (defined as generating fluorescence signal but showing no significant difference compared with negative control) and CIV (defined as lacking fluorescence signal and showing no significant difference compared with negative control). Half of the candidates were validated by reporter assays; in detail, four validated candidates were from CI (15.4%, FSPE1-4), four were from CII (15.4%, FSPE5-8), five were from CIII (19.2%, FSPE9-13) and 13 were from CIV (50.0%, FSPE14-26; Figures 1b,c and 4e). These results demonstrate that traditional STARR-seq works in rice protoplasts.

### The rice protoplast-based STARR-seq system can reveal conserved enhancers from *Arabidopsis*

Enhancers found in one plant lineage are often active in other species. Open chromatin signatures were previously used to predict enhancers in *Arabidopsis*, and ten of 14 candidates were validated (Zhu *et al.*, 2015). Here, we cloned 13 candidates and two intergenic region controls to test their enhancer activity in rice protoplasts (Figure 1a). The two positive sites, C5 and L2, can activate gene expression in both directions, consistent with their ability in *Arabidopsis*; however, the other eight positive enhancers in *Arabidopsis* could not be identified in the rice protoplast system (Figure 2a–c). These results suggest that the system could be used to identify conserved but not all enhancers from other plant genomes.

### STARR-seq identifies not only enhancers but also active promoters

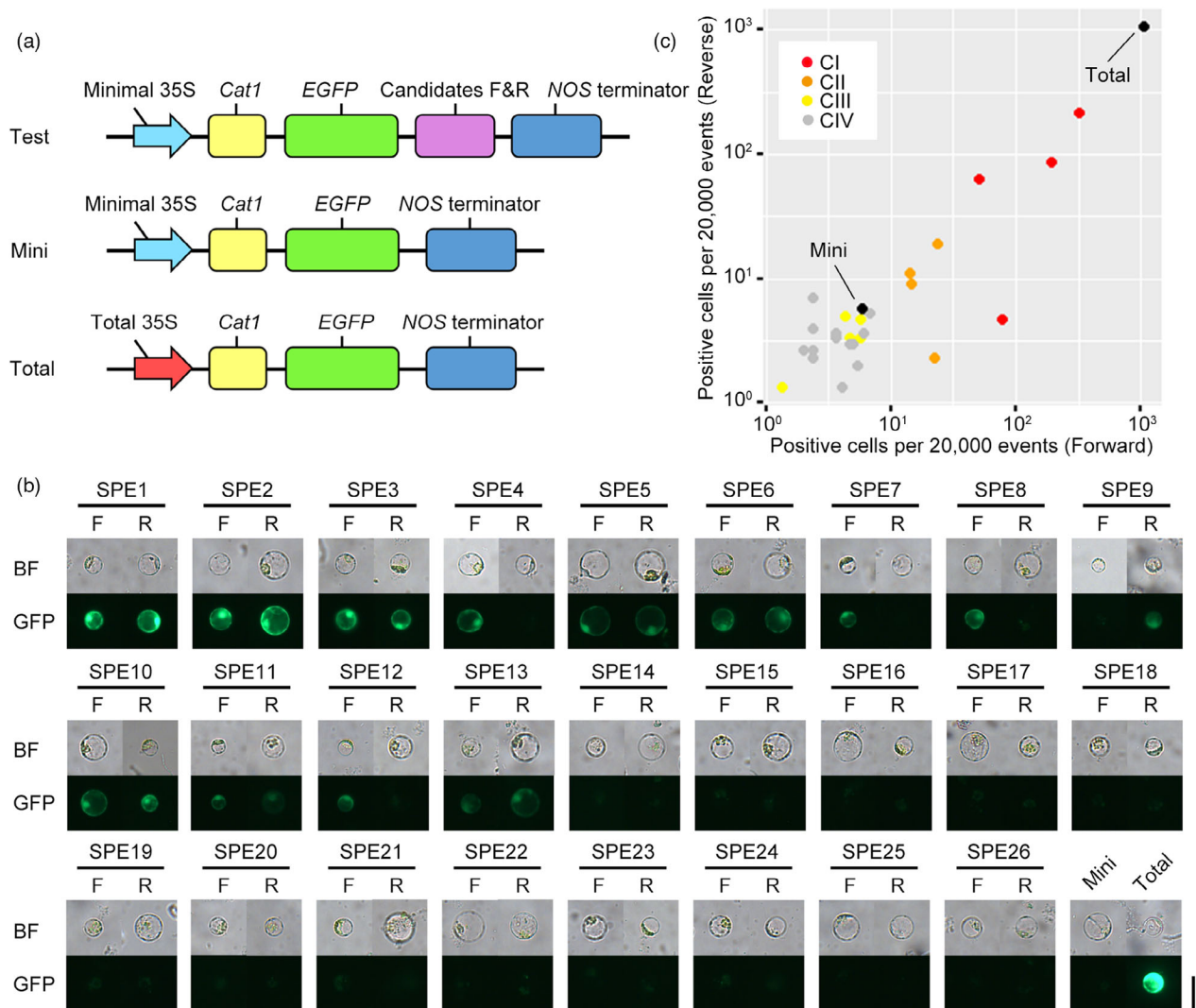
In general, enhancers function independently of their positions and orientations to target genes while promoters modulate their

proximal downstream genes. In the *N. benthamiana* research, four enhancers showed no activating ability in the 3'UTR (Jores *et al.*, 2020). Because all of the enhancers were located proximally to their core promoters (Fejes *et al.*, 1990; Giuliano *et al.*, 1988; Ow *et al.*, 1987; Simpson *et al.*, 1986), we were curious about whether candidates used in that work were promoters but enhancers. Therefore, we compared the activating ability of four SPEs belonging to group CI with eight rice constitutive promoters (Bang *et al.*, 2015; Park *et al.*, 2010) in upstream region and 3'UTR (Figure 3a). Surprisingly, apart from negative controls, all candidates showed activation in both positions (Figure 3b). However, compared with upstream, placing enhancers or promoters in 3'UTR led to a great reduction of activating strength (Figure 3c). This finding is consistent with the assay in maize (*Zea mays*) protoplasts (Jores *et al.*, 2020). In addition, all AREs functioned independently of their orientations to the minimal 35S promoter when placed in the upstream position (Figure 3b and c). Taken together, these results demonstrate that enhancers and promoters display similar behaviour in the STARR-seq system. Thus, the unsuitable application of the traditional STARR-seq method in *N. benthamiana* is not caused by using promoters. The majority of exogenous DNA in the *Agrobacterium*-mediated method is single-stranded DNA (Albright *et al.*, 1987) coated by bacterial proteins (Herrera-Estrella *et al.*, 1988; Herrera-Estrella *et al.*, 1990). By contrast, the PEG-mediated method transfers double-stranded and naked DNA into cells. Therefore, divergence of transfection methods may underlie the conflicting reports.

It should be noted that, as activating ability was greatly reduced in the 3'UTR, there could still be AREs among the members of group CIV. Nevertheless, as these regions cannot be distinguished from the minimal promoter control in our system, we treat them as negative.

### Improved STARR-seq more precisely predicts AREs

Although our results showed that traditional STARR-seq could be directly performed in rice protoplasts, the positive rate for SPEs was far from satisfactory (Figure 4e and Figure S2). SPE22 was previously identified (Sun *et al.*, 2019) and, in fact, has 101 homologous sequences (Data S1) within the peaks in both replicates and the remaining top 1000 in replicate 1. However, SPE22 displayed no activating ability in our system (Figure 1b). False positives for SPEs may occur for a combination of two reasons. First, the plasmid origin of replication (ORI) constitutively starts transcription, which means sequences that cannot activate transcription will be transcribed in the STARR-seq system. It has been reported that a cryptic promoter within the ORI can produce substantial amounts of RNA in human cells (Lemp *et al.*, 2012). Similarly, we found that cryptic transcription also occurred in the transient expression system of rice protoplasts and insertions lacking activating ability produced considerable transcripts compared with true AREs (Figure S4). To alleviate the effect of leakage from ORIs, a former study removed the core promoter of the human source and set the ORI as the only core promoter to reach a higher signal-to-noise ratio (Muerdter *et al.*, 2018). However, this method cannot reduce the background noise from sequences lacking activating ability. Considering that AREs only account for a small part of whole genome, a large amount of sequencing data should be acquired to predict AREs precisely. Second, duplicates were removed during prediction. To avoid the influence of non-specific amplification, the traditional method



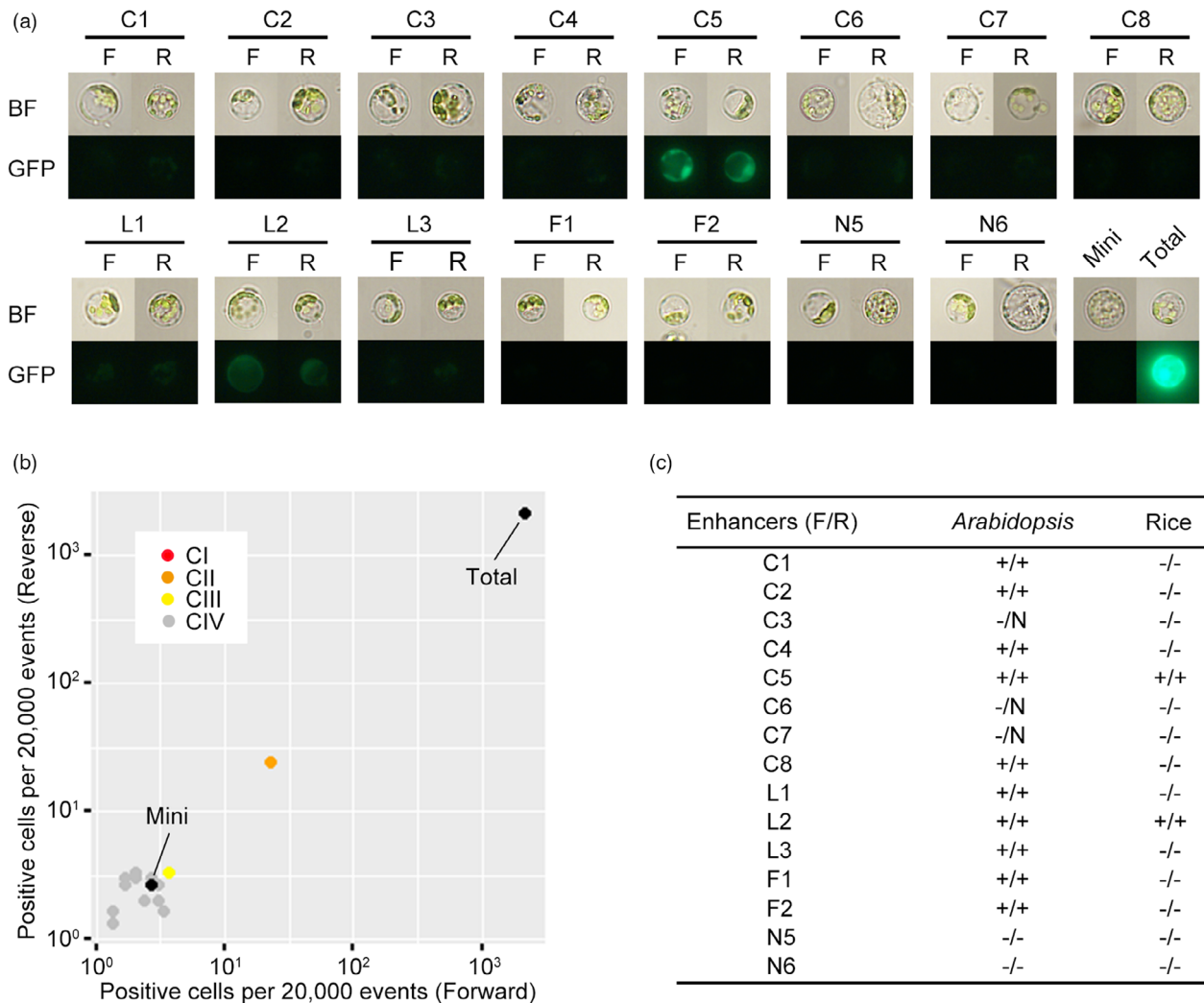
**Figure 1** Traditional STARR-seq method works in rice protoplasts. (a) Vectors used for validation assay. Mini, original STARR-seq vector which served as negative control; Total, full-length 35S promoter STARR-seq vector which served as positive control; Test, original vector carrying candidates in the 3'UTR; F, forward insertion; R, reverse insertion. (b) Fluorescence microscopy of protoplasts transfected by SPEs (Sun *et al.*, 2019). Bar = 15  $\mu$ m. (c) Flow cytometry analysis of protoplasts transfected by SPEs. Candidates were artificially divided into four classes according to their activating ability. Data are shown as means;  $n = 3$ .

removed duplicates after mapping. However, owing to the leakage from ORIs, high-coverage regions are naturally preferred by the analysis approach and this may be why repetitive sequences were highly enriched in the prediction results of the traditional method (Sun *et al.*, 2019). Therefore, we optimized two procedures (Figure 4a) to increase the performance of STARR-seq in plants. First, we took advantage of FACS to separate fluorescent cells (Figure 4b and Figure S5), which in theory harbour more information of AREs. Second, we minimized the number of PCR cycles and retained duplicates for further analysis.

In melanoma cells, the median length of typical enhancers is 1.3 kb (Hnisz *et al.*, 2013). In *Arabidopsis*, 6.4% of common enhancers are longer than 800 bp (Zhu *et al.*, 2015). Sheared DNAs between 500 and 800 bp were used to generate plasmid library in former study (Sun *et al.*, 2019), which might therefore

omit partial larger AREs. To identify AREs more comprehensively, we constructed a plasmid library carrying fragmented genomic DNA mainly between 700 and 1500 bp to transfect protoplasts. After cell sorting, about 100 000 cells were harvested. Then, RNA was extracted and reverse transcribed to cDNA using a specific primer. Insertions were enriched using a 2-step nested PCR within which 10 cycles were carried out in first round with a forward primer spanning intron and 15 cycles were carried out in second round. Sheared insertions were used to generate a PCR-free library.

After sequencing, reads were aligned by Bowtie2 with default settings. About 8.1 million reads (35.0%) of the ARE library were aligned concordantly and exactly 1 time. By contrast, the counterparts in reported data (Sun *et al.*, 2019) were 1.6 million reads (1.9%) and 0.7 million reads (0.9%) respectively (Table S1). Only concordantly and uniquely mapped reads were kept for



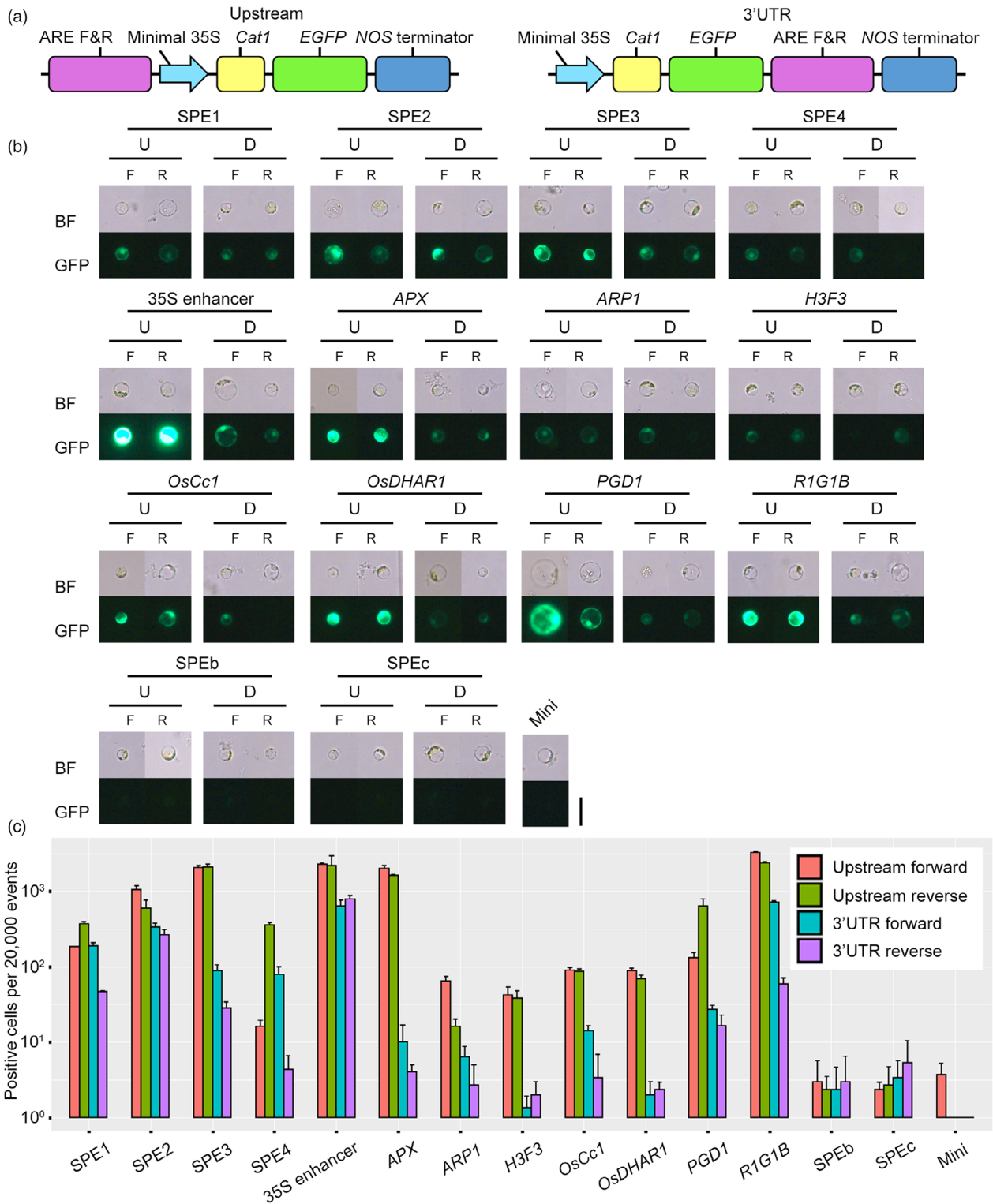
**Figure 2** Conserved enhancers from *Arabidopsis* can be identified by the rice STARR-seq system. (a) Fluorescence microscopy of rice protoplasts transfected by enhancers from *Arabidopsis* (Zhu *et al.*, 2015). Bar = 15  $\mu\text{m}$ . (b) Flow cytometry analysis of rice protoplasts transfected by enhancers from *Arabidopsis*. Candidates were artificially divided into four classes according to their activating ability. Data are shown as means;  $n = 3$ . (c) Comparison of enhancer activating ability in *Arabidopsis* and rice. The activating ability in *Arabidopsis* is summarized from reported assays (Zhu *et al.*, 2015). +, showing activating ability; -, lacking activating ability; N, not available.

further data processing. Duplicates were retained, and 29 172 peaks with  $447.6 > \text{FE} > 1.49$ , ranging from 205 to 2584 bp were finally revealed by MACS2 (Data S2). Within them, only 2288 peaks overlapped with previously reported enhancers (Sun *et al.*, 2019). This divergence was not caused by different analysis methods, as peaks generated from reported data following the same process still had little in common with our results (Figure S6). We randomly cloned 30 FACS-STARR-seq-predicted AREs (FSPEs) within the top 150 peaks (Figure 1a). Twenty-seven of 30 candidates (90%) were detected as showing fluorescence signals (Figure 4c), which was 1.8 fold more than the reported sites. In detail, 12 belonged to CI (40.0%, FSPE1-12), ten belonged to CII (33.3%, FSPE13-22), five belonged to CIII (16.7%, FSPE23-27) and three belonged to CIV (10%, FSPE28-30; Figure 4c-e). Notably, the percentage of relatively more robust classes, CI and CII, were higher than with traditional method (Figure 4e). Seven of 13 validated SPEs (53.8%) are also identified by improved method while four of 27 validated FSPEs (14.8%) can also be found in reported data (Data S3). These

results demonstrate that our effort of optimization successfully increases the performance of STARR-seq.

An independent experiment was carried out and yielded 26 916 peaks (Data S2). Twenty-four of 27 verified FSPEs (88.9%) in replicate 1 can also be found in replicate 2 (Data S3). To determine a reasonable cutoff value, we cloned 24 candidates with relatively low FE (Figure 1a). Within them, eight were present only in replicate 1 ( $18.9 > \text{FE} > 4.4$ ), eight were present only in replicate 2 ( $12.0 > \text{FE} > 4.7$ ) and eight were present in both replicates ( $13.6 > \text{FE} > 4.8$  in replicate 1). As expected, the decrease of FE was accompanied by the lowering of positive rate and activating strength. Four of eight (50.0%), three of eight (37.5%) and five of eight (62.5%) candidates in each group were validated, respectively, which are all lower than that in the top 150 of replicate 1 (90.0%); in detail, one belonged to CI (4.2%), seven belonged to CII (29.2%), four belonged to CIII (16.7%) and 12 belonged to CIV (50.0%), and the overall precision was 50% which is the same as the top 100 of reported data (Figure 4e, Figures S7 and S8). These findings demonstrate that it is difficult

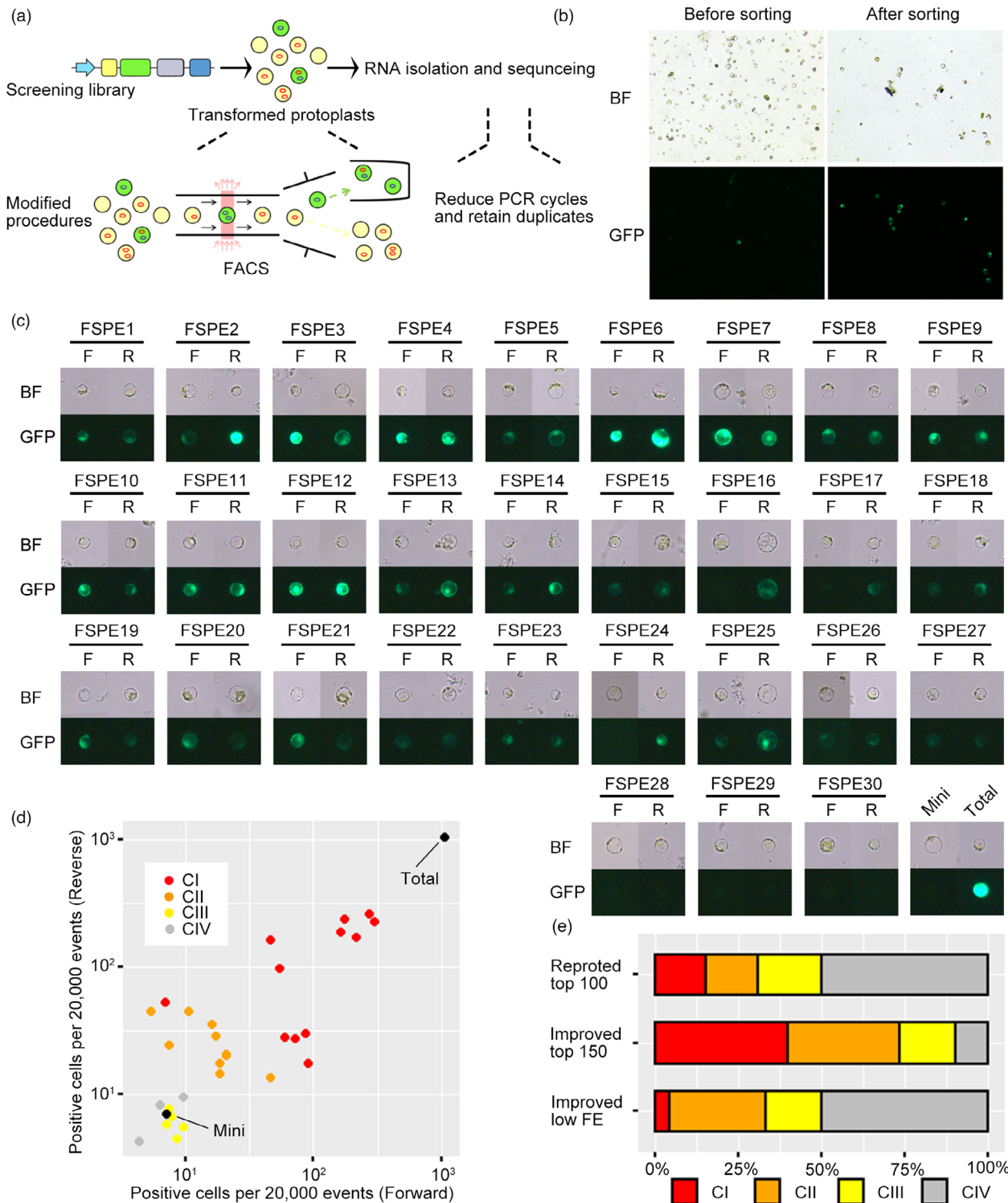




**Figure 3** Both enhancers and promoters show activating ability in the STARR-seq system. (a) Vectors used for comparison assay. SPEs belonging to CI and rice constitutive promoters (Bang *et al.*, 2015; Park *et al.*, 2010) were inserted into proximal upstream regions of core promoters or the 3'UTR. (b) Fluorescence microscopy of protoplasts transfected by ARES. SPEb, SPEc and original vector (Mini) served as negative controls. U, upstream insertion; D, 3'UTR insertion; F, forward insertion; R, reverse insertion. Bar = 15  $\mu$ m. (c) Flow cytometry analysis of protoplasts transfected by ARES. Data are shown as means + SD;  $n = 3$ .

to distinguish vulnerable regions from noise. As a result, only peaks with FE > 20 in each replicate were kept for further analysis. Among them, 372 peaks were only identified in replicate

1 (20.5%), 277 peaks were only identified in replicate 2 (15.2%), and 1169 peaks were identified in both replicates (64.3%; Figure S9).



**Figure 4** Improved STARR-seq significantly improves precision. (a) Scheme of the improved method. Sheared genomic DNA (grey) was inserted into the 3'UTR of the original STARR-seq vector to generate a screening library. Compared with the traditional version, two procedures were modified. First, transformed protoplasts were sorted by FACS, and only fluorescent cells (green) were used for RNA isolation. Second, we minimized PCR cycles to retain duplicates during prediction. (b) Comparison of cell populations before and after sorting. Bar = 100 μm. (c) Fluorescence microscopy of protoplasts transfected by FSPEs. Bar = 15 μm. (d) Flow cytometry analysis of protoplasts transfected by FSPEs. Candidates were artificially divided into four classes according to their activating ability. Data are shown as means;  $n = 3$ . (e) Comparison of positive rates between traditional and improved methods.

## Two groups of AREs with distinct preferences in GC content are identified among FSPEs

The majority of the FSPEs were distributed in promoter (41.0%) and intergenic (38.8%) regions, whereas 17.4% of FSPEs were located in the gene body and 2.8% were downstream (Figure 5a). A 'GGCGGC' motif that is recognized by the AP2/ERF superfamily (Allen *et al.*, 1998; Ohme-Takagi and Shinshi, 1995) and an A/T-rich sequence were enriched in FSPEs (Figure 5b). These findings prompted us to examine the GC content of FSPEs. We calculated the GC percentage of FSPEs in 100-bp bins, and clustered sequences using K-means. The majority of FSPEs had a relatively low GC percentage (around 40%) in their centers (Figure 5c), consistent with the reported data that AT-rich promoters possess transcriptional activity (Delaney *et al.*, 2007; Qian *et al.*, 2007; Tjaden and Coruzzi, 1994). Interestingly, FSPEs could be clustered into two groups where G1 had an apparent GC-rich region in the edge, but G2 showed a low GC percentage throughout the whole body (Figure 5c and Data S2). Distribution analyses of the two groups showed that G1 was more likely to appear in the proximal promoter (promoter <1 kb; 34.6%), whereas G2 was more likely to appear in the distal intergenic region (48.1%; Figure 5a).

We took advantage of reported ChIP-seq data (Zhao *et al.*, 2020) from similar growth stage materials to examine the endogenous histone modification of FSPEs. Active marks, H3K4me3 and H3K27ac, and repressive mark, H3K27me3, were enriched in the GC-rich region of G1 (Figure 5e, f and h). This result was consistent with the reported data that both active and silenced enhancers could be identified by STARR-seq (Arnold *et al.*, 2013). The central region of G1 showed a slightly lower level of H3K4me1 modification (Figure 5d). By contrast, G2 showed no association with any histone modifications (Figure 5d–h).

One marked difference between promoters and enhancers in mammals is the overall GC content. About 50% of promoters overlap with CpG islands, while almost no enhancers do (Andersson *et al.*, 2014). Another key feature that has been used to distinguish promoters from enhancers is the much more likely enrichment of flanking H3K4me3 modification in promoters (Soares *et al.*, 2017). G1 apparently favoured a flanking GC-rich region (Figure 5c) that was enriched with H3K4me3 modification (Figure 5e) and had a preference to distribute in the proximal promoter (Figure 5a). These features demonstrate that G1 members are more likely to be promoters. By contrast, G2 showed AT-rich content throughout whole body (Figure 5c), were depleted of histone modifications (Figure 5d–h) and preferred to distribute in distal intergenic regions (Figure 5a), which suggests G2 members are more likely to be enhancers.

## FSPEs do not show apparent correlation with chromatin structures

It has been reported that chromatin packing is highly associated with epigenomic features and DNA motifs in rice (Liu *et al.*, 2017). Therefore, we analysed the relationship between FSPEs and chromatin structures. The A compartment shows higher levels of active chromatin marks and contains more euchromatin, while the B compartment generally has higher levels of DNA methylation and contains more heterochromatin. We calculated eigenvector values and determined the distribution of FSPEs in different compartments. There were 1004 FSPEs located in A compartment and 814 FSPEs located in B

compartment (55.2% and 44.8%, respectively, Figure S10). Considering B compartments cover around 60% of the *Nip* genome, FSPEs showed a slightly higher preference to distribute in A compartments. Topologically associated domain (TAD) boundaries in rice are characterized by open chromatin signatures and enriched for several motifs including 'GGCGGC' and A/T-rich sequences, which are also prevalent in FSPEs. Therefore, we were curious about whether the STARR-seq signal was associated with TAD boundaries. However, no apparent enrichment of STARR-seq signal could be observed in TAD boundaries (Figure S11). Furthermore, the distribution of FSPEs was not associated with TAD (Figure S12). Taken together, our results suggest that FSPEs do not show apparent correlation with chromatin structures.

We also performed loop calling under 5 kb resolution and obtained 876 potential chromatin interactions within which 583 FSPEs were covered (Data S4). Distribution analysis showed that the regions interacted with FSPEs were mainly located in proximal promoter and distal intergenic regions (53.8% and 24.3%, respectively, Figure S13).

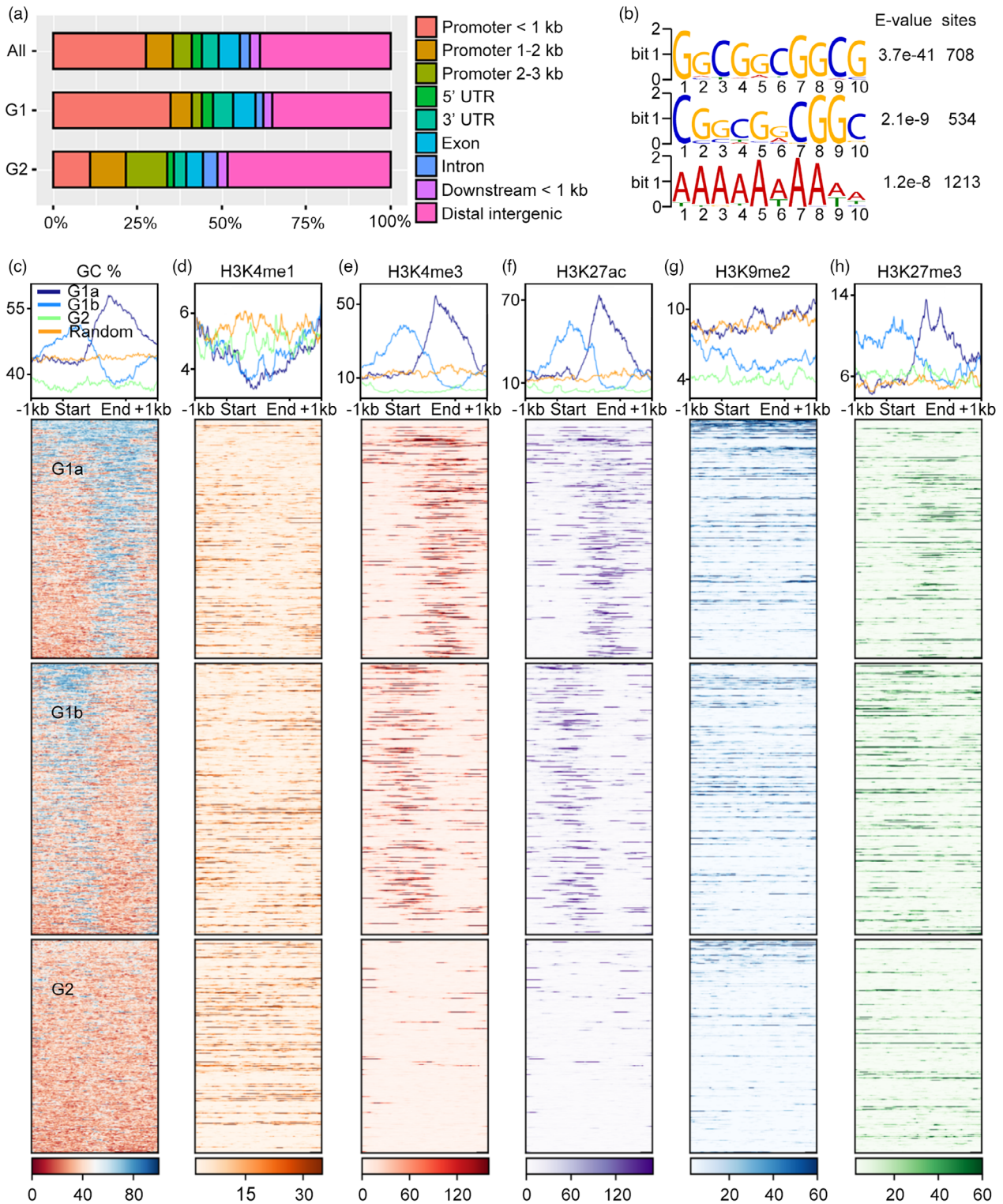
## Both AT- and GC-rich regions within AREs have activating ability

To gain deeper understanding of the relationship between GC content and activating ability, four FSPEs belonging to CI were selected and fragmented according to their GC percentage (Figure 6). Fragments were cloned into STARR-seq vector with forward orientation in upstream position. Activating strength was also roughly represented by GFP-positive cells. AT-rich fragments, FSPE6-F1 (Figure 6b) and FSPE10-F2 (Figure 6c), and GC-rich fragments, FSPE5-F1 (Figure 6a) and FSPE11-F2 (Figure 6d) all showed comparable activating ability to their full-length counterparts. These results demonstrate that activating ability is not correlated with GC content, and both AT- or GC-rich regions in AREs can boost transcription.

## Disturbance of FSPEs leads to abnormal expression of both proximal and distal genes

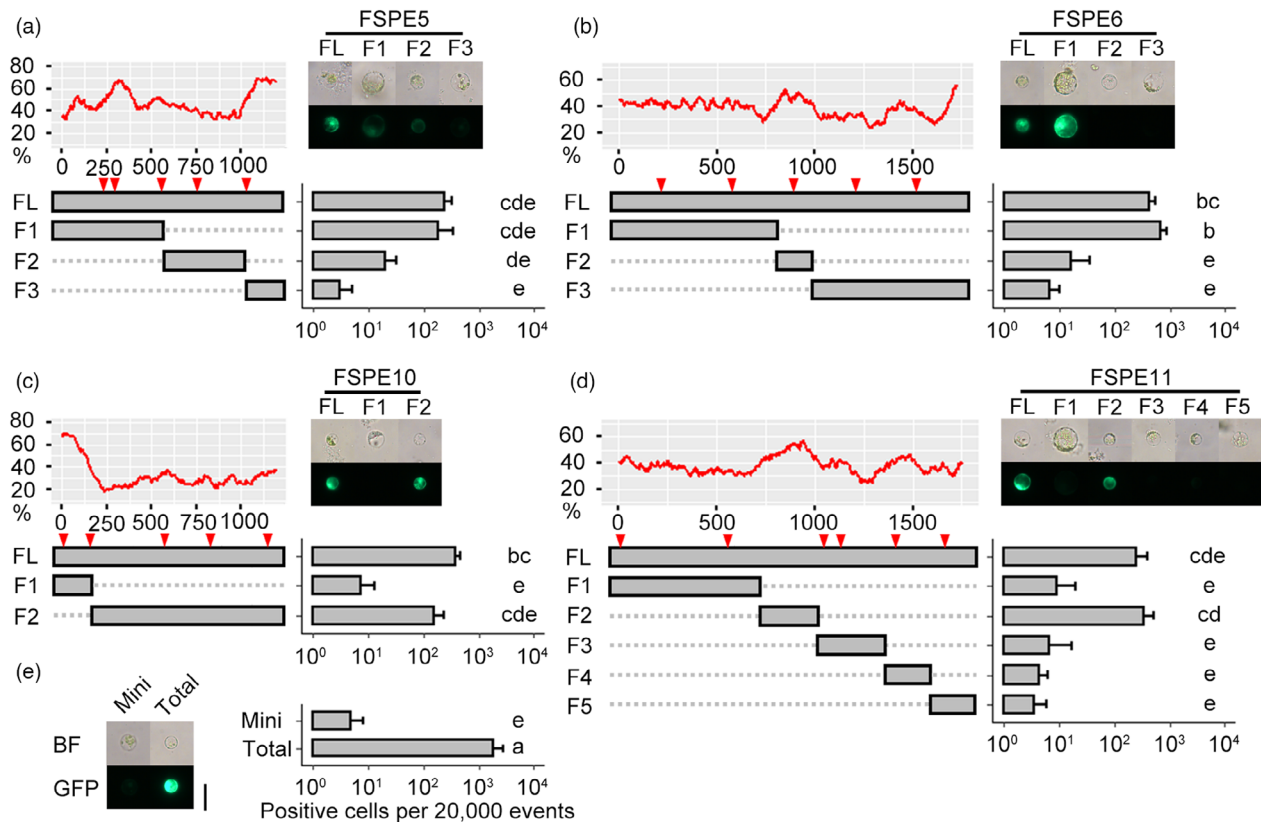
CRISPR/Cas9 system was used to facilitate the determination of FSPEs function *in vivo*. For each FSPE, five specific sgRNA were designed (Figure 6). After transfection, targets were amplified following Sanger sequencing. Mutations around the protospacer adjacent motif could be detected (Figure S14), demonstrating that Cas9 successfully bound to targets. Coding genes within 50 kb from FSPEs were examined. Ten of 25 genes showed consistent up- or down-regulation among four replicates (Figure S15), and five of them showed significant differences with no sgRNA control (Figure 7a–d). For FSPE5 and FSPE6, not only proximal but also distal gene expression was influenced when FSPEs were disrupted (Figure 7a,b, Figures S16 and S17). Both FSPE5 and FSPE6 are located in the proximal intergenic regions of opposite gene pairs. After disruption, a downstream flanking gene of FSPE5, *OsDjC21*, was consistently down regulated. Disruption of FSPE6 led to a consistent up-regulation of the proximal downstream gene *HMA5* but a consistent down-regulation of the proximal upstream gene *SERR*. Notably, interactions between gene pair could be detected in both cases (Figures S16 and S17). This evidence suggests that the two FSPEs potentially serve as bi-directional promoters of their flanking genes. However, whether they regulate distal genes through long-range DNA interactions or influencing signalling should be further explored. As for intron-located FSPE10 and 3'UTR-located FSPE11, they more likely function as enhancers because they





**Figure 5** Characterization of FSPes. (a) Distribution of FSPes in the genome. Upstream 3 kb region of transcription start site was defined as the promoter. Priority: Promoter > 5'UTR > 3'UTR > Exon > Intron > Downstream > Distal Intergenic. (b) Enriched motifs within FSPes. Distributions of GC percentage and epigenetic marks in FSPes. Sequences were normalized to 2 kb. FSPes were clustered into two groups according to their GC-rich region preference. G1 was further divided into G1a and G1b based on the position of the GC-rich region. Two thousand sequences randomly selected from genome served as random control. Random selections were performed ten times, and one typical selection is presented. (c) Distribution of GC percentage in FSPes. (d-h) Distribution of H3K4me1 (d), H3K4me3 (e), H3K27ac (f), H3K9me2 (g), and H3K27me3 (h) in FSPes.





**Figure 6** Both AT- and GC-rich elements within FSPEs show activating ability in the STARR-seq system. Four validated FSPEs belonging to CI were selected for dissection. The GC percentage of each FSPE was calculated in 100-bp bins. According to GC content, FSPEs were fragmented into AT- and GC-rich sequences. Fragments in forward orientation were inserted into original STARR-seq vector at the proximal upstream of core promoter. Activating ability was evaluated by fluorescence observation and flow cytometry analysis. Red inverted triangles indicate the target sites for sgRNA in the disruption assay. Results of flow cytometry analysis are shown as means + SD;  $n = 3$ . Statistical comparison was done by least significant difference tests. Bar = 15  $\mu\text{m}$ . (a–d) Dissection of FSPE5 (a), FSPE6 (b), FSPE10 (c), and FSPE11 (d). (e) Original vector (Mini) served as negative control while full-length 35S promoter vector (Total) served as positive control.

influence the gene expression of their hosts in spite of position (Figure 7c,d, Figures S18 and S19). Active contacts could be found around both FSPE10 and FSPE11 (Figures S18 and S19). Taken together, these results demonstrate that STARR-seq-revealed candidates indeed function as transcriptional regulatory elements *in vivo*.

## Discussion

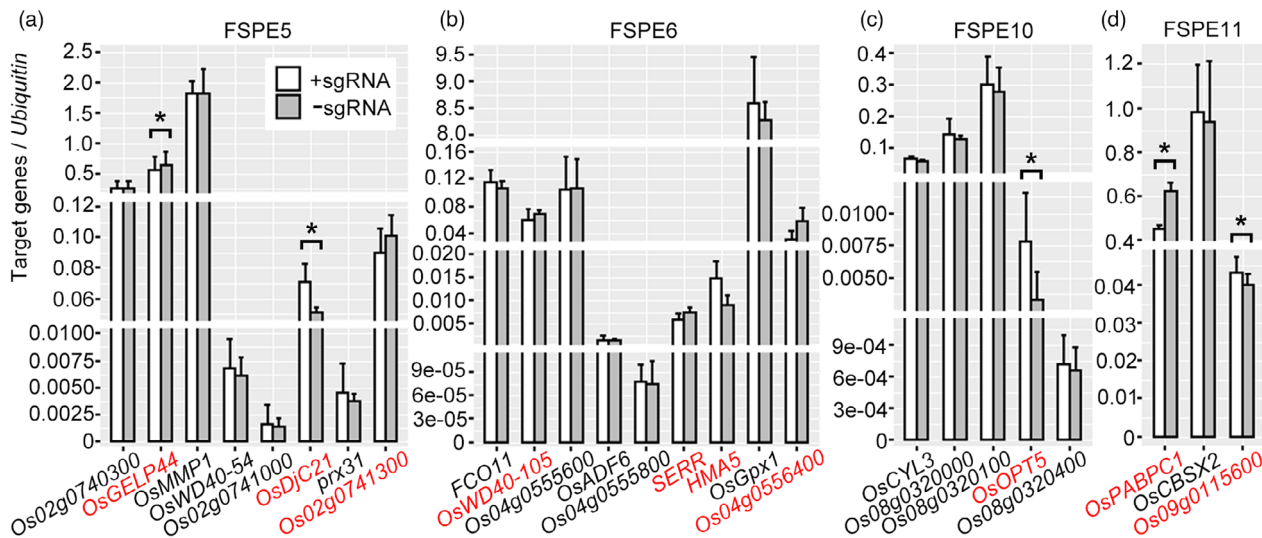
This study presents an improved method that can identify AREs in plants. Compared with the traditional approach, our method displayed higher precision, and compared with the upstream insertion version of the method, ours is more suitable for genome-wide screening. Furthermore, our method relies on generating protoplasts, for which protocols are well established in many plant species (Lin *et al.*, 2018). Sorting plant cells is also routine (Bargmann and Birnbaum, 2010; Ortiz-Ramírez *et al.*, 2018; Petersen *et al.*, 2019; You *et al.*, 2014). Therefore, it is reasonable to believe that our improved method can be adapted for other plant species.

Although we have developed an improved protocol, several procedures can be further optimized. In the human STARR-seq, several core promoters have been tested (Muerdter *et al.*, 2018). Using the correct promoter enormously improved the signal-to-noise levels in that case. So far, all relevant studies in plants have

used the minimal 35S promoter. However, that promoter has been reported to retain considerable strength to initiate expression (Ow *et al.*, 1987). Thus, it would be wise to seek more suitable promoters.

In our study, to avoid the effect of non-specific amplification, we minimized the number of PCR cycles as much as possible. However, the fluorescent cell population was limited. The number of cycles needed to yield enough DNA for sequencing was considerable, which is unfavourable for accurately assessing activating strength. To quantify activating strength more accurately, unique molecular identifiers could be added to the sequencing library (Neumayr *et al.*, 2019). These are helpful for the removal of PCR duplicates. Nevertheless, reducing PCR cycles was still a fruitful and convenient method and our results demonstrate that it was sufficient for ARE identification.

In the original design of STARR-seq, an artificial terminator was considered to be the only region that provides a polyadenylation site. However, enhancers are co-localized with terminators in some cases, which means self-transcripts from these regions are split before the designated site. Early termination leads to loss of the reverse primer binding site, which is critical in enriching enhancer information, and therefore results in false negatives for enhancers containing polyadenylation sites. An improved STARR-seq method named iSTARR-seq fixed the error by adding a poly(A) sequence after the insertion site and enriched insertions using



**Figure 7** Disturbance of FSPEs resulted in abnormal gene expression. CRISPR/Cas9 system was introduced into protoplasts to disturb the function of FSPEs *in vivo*. Five sgRNA were designed for each FSPE, and target sites are indicated Figure 6. Expression of coding genes that within 50 kb from FSPEs was examined. Expression levels are shown as means + SD;  $n = 4$ . Genes consistently changed in four replicates are coloured in red. Statistical comparison was done by two-tailed  $t$  tests; \* $P < 0.05$ . (a–d) Gene expression with or without disruption of FSPE5 (a), FSPE6 (b), FSPE10 (c), or FSPE11 (d).

modified primers (Niu *et al.*, 2020). In our results, AREs distributed in downstream regions account for only a small portion of the AREs, which may be potentially caused by ignoring the effect of early termination. Combining iSTARR-seq and FACS in the future could be fruitful to identify AREs even more comprehensively and accurately.

FSPEs do not show apparent correlation with chromatin structures in our data. Potential contacts between FSPEs and genes that changed expression after disruption could be observed; however, these interactions were not prominent (Figures S16–S19). The weak interaction strength suggests that the sequencing depth of recent Hi-C data could be further increased to obtain an ideal genome-wide chromatin packing map. Therefore, the conclusions we present here are preliminary and should be further tested when more ideal Hi-C data are acquired. Reported data (Arnold *et al.*, 2013) and our results have found that silenced AREs *in vivo* can also be revealed by the STARR-seq system. Therefore, it is possible that a portion of the AREs analysed in our study were inactive in the plant materials used in Hi-C study (Liu *et al.*, 2017), which might be another potential explanation for why FSPEs did not show correlation with chromatin structures.

There are at least two models proposed for chromatin interaction. Chromatin structural factors, such as Yin Yang 1 (Weintraub *et al.*, 2017), mediate DNA loop formation through binding specific DNA motifs. In addition, transcription increases molecular motion, which facilitates chromatin interaction (Gu *et al.*, 2018). This view of interaction is supported by the finding that RNA polymerase II molecules could recruit each other through their C-terminal domains to specify the localization of active promoters (Lu *et al.*, 2018). In our case, constitutive promoters in the 3'UTR could also activate upstream core promoters, which is consistent with the molecular stirring model. Recently, it has been revealed that enhancers share similar architecture with promoters (Tippens *et al.*, 2020). These results indicate that there might be no distinct boundary between enhancers and promoters. In fact, a sizeable fraction of

promoter-associated regions are other gene promoters (Mifsud *et al.*, 2015; Schoenfelder *et al.*, 2015). It is possible that promoters should be more accurately viewed as one particular enhancer located behind the core promoter. Accordingly, the promoter might participate in regulating a wide range of genes instead of genes just downstream of it.

## Materials and methods

### Plant materials

Rice cultivar *Nipponbare* (*Nip*) was used in this study. Seeds were sown and grown on MS solid culture medium for 10–14 days with a photoperiod 14 h light and 10 h dark. All materials were grown at 28 °C.

### Plasmid construction

The STARR-seq vector was generated from pSP64-poly(A) (Promega) by replacing the sequence between the *AatII* and *EcoRI* sites with an expression cassette, including a minimal 35S promoter, a *Cat1*, an *EGFP* gene, a multiple cloning site and a NOS terminator.

CRISPR/Cas9 vector was generated from total 35S promoter STARR-seq vector by replacing the *Cat1* with a sequence including a 3XFlag tag, a SV40 NLS, a *Cas9* gene, a nucleoplasm NLS and a P2A cleavage site. The sgRNA expression cassettes were yielded from sgRNA intermediate vectors (Ma *et al.*, 2015), and cloned into CRISPR/Cas9 vector digested with *Ascl*.

Detailed information of derivatives is listed in Data S3.

### Screening library preparation

*Nip* genomic DNA was sonicated using a Bioruptor (Diagenode). Fragments were recovered using 0.6X KAPA Pure Beads (KAPA) and subjected to a KAPA Hyper Prep Kit (KAPA) to add consensus adapters following the manufacturer's protocol. During the amplification step, the KAPA Library Amplification Primer Mix was replaced by 'AGTAAGAGCTCAATTACACTCTTCCCTACAC-GACGCTCTCCGATC/CGGGAGGATCCAAGGGTACTGGAGTT

CAGACGTGTGCTTCCGATCT'. After 5 cycles of amplification, the products were size selected on a 2% agarose gel. Fragments mainly between 700–1500 bp were recovered using a QIAquick Gel Extraction Kit (Qiagen).

Recovered DNA and linearized STARR-seq vector (digested by *MfeI* and *StuI*) were recombined using an In-Fusion® HD Cloning Kit (Clontech). Insert and vector were each presented at 1 µg. Products were used to transform homemade *Trans1-T1* competent cells by heat-shock treatment. After 1 h recovery, a small amount of culture was transferred to an LB AMP100 plate to evaluate the transformation efficiency. The transformation efficiency was about  $1 \times 10^7$  cfu/µg, which about 25X coverage of the *Nip* genome. The remaining culture was transferred to LB AMP100 liquid medium and incubated overnight. Overnight culture was recovered and stored at  $-80^\circ\text{C}$  as a stock.

### Protoplast transfection

The isolation and transfection of *Nip* protoplasts were performed as described previously (Zhang *et al.*, 2011) with modifications. Tender stem and sheath tissues were cut into approximately 0.5 mm strips, and directly transferred into enzyme buffer. Twenty millilitre enzyme buffer was used to digest about 150 seedlings. After 7 h digestion, an equal volume W5 solution was added to digestion buffer. The mixture was filtered through a 40 µm nylon mesh. Strips were washed with W5 solution until the released protoplasts became abnormal. Typically, 150 seedlings could yield at least  $1 \times 10^8$  protoplasts. Protoplasts were washed with W5 and MMG solution sequentially, and finally resuspended with MMG solution.

For library and disruption transfections, the concentration was  $1 \times 10^7$  cells mL<sup>-1</sup>. Each 1 mL of cells were mixed with 50–100 µg plasmid extracted using the CsCl density-gradient centrifugation method. PEG-mediated transfection lasted for 15 min. Transfection was terminated by adding 4 mL W5 solution, including 5 mM glucose and 0.1% BSA (W5GB). Protoplasts were precipitated by gravity naturally. Supernatant was discarded, and protoplasts were resuspended with W5GB solution. Finally, protoplasts were cultured under light at  $28^\circ\text{C}$  for 12–24 h.

For normal transfections, several procedures were changed. First, the concentration was about  $2 \times 10^6$  cells mL<sup>-1</sup>. Second, each 200 µL cells were mixed with 10–20 µg plasmid extracted by the spin column method. Third, protoplasts were directly incubated for 12 h after adding W5GB solution.

### Microscopy

Protoplasts were observed using a DM6 B microscope (Leica) following the manufacturer's instructions. Excitation intensity was adjusted to maximum. Exposure time was set at 100.00 ms. Signal gain was set at 2.0. Illumination settings varied slightly with day-to-day operations. Images were acquired under a L5 filter using TL-BF or FLUO channel and processed using LAS X software (Leica).

### Reverse transcription-PCR

Total RNA was extracted from 500 µL transfected protoplasts using an Eastep Super Total RNA Extraction Kit (Promega). Reverse transcription was performed using a Goscript™ Reverse Transcription System (Promega). Oligo(dT) was used as the reverse transcription primer. Input plasmids were extracted from 100 µL transfected protoplasts using E.Z.N.A Plasmid Mini Kit (Omega). Partial coding sequence for *GFP* was amplified for 20 cycles at an annealing temperature of  $58^\circ\text{C}$  using 'GCA-GAACACCCCATCGG/CATGTGATCGCGCTTCTCGT'.

### Flow cytometry analysis and fluorescence activated cell sorting

Flow cytometry analysis was performed on a FC500 cytometer (Beckman Coulter) following the manufacturer's instructions.

Cell sorting was performed on a FACS Aria™ III cell sorter (BD Biosciences) fitted with a 100-µm nozzle. Normal saline was used as sheath fluid. To separate GFP-expressing cells, 488 nm laser was used for excitation. For stream setting, amplitude was set to 4.6 V, frequency was set to 87 kHz, and drop delay was set to 44.4. The stream setting will vary slightly with day-to-day operations. The photomultiplier tube voltage was set at 230 V for forward scatter, 280 V for side scatter, and 250 V for GFP.

Suspensions were adjusted to about  $5 \times 10^6$  cells mL<sup>-1</sup> and filtered through 40-µm nylon mesh into cytometer tubes before loading. Samples were analysed at rate about 10 000 events/s. Notably, protoplasts were mixed well about every 10 min by pipetting up and down instead of agitation. Normally, about  $1 \times 10^5$  positive cells could be obtained from a start 150 seedlings.

### Generation of sequencing libraries

For each individual experiment, about  $1 \times 10^5$  cells were used. Total RNA was extracted using an RNeasy Micro Kit (Qiagen) coupled with an RNase-free DNase Set (Qiagen). Reverse transcription was carried out using a Promega Goscript™ Reverse Transcription System (Promega) with a sequence-specific primer 'GCCAATGTTGAACG'. Target sequences were amplified via a 2-step nested PCR using KAPA hotstart mix (KAPA). First-round PCR was carried out for 10 cycles using 'GCAGCAATTTAAATAG-GAACTAGTATGGTGAGC/CGATCGGGAGGATCCAAGGGTGA' at annealing temperature  $55^\circ\text{C}$ . Products were purified using a QIAquick PCR Purification kit (Qiagen). Second-round PCR utilized 15 cycles with 'GCTCAATTACTCTTTCCCTACACGACGCTC/GGTGACTGGAGTTCAGACGTGTGCTCTTC' at annealing temperature  $60^\circ\text{C}$ . Products were size selected on a 2% agarose gel. Fragments mainly between 700–1500 bp were recovered. The resulting DNA was sheared and subjected to a TruSeq DNA PCR-Free kit (Illumina) to construct an ARE sequencing library. One millilitre of cells before sorting were recovered and subjected to the E.Z.N.A Plasmid Mini Kit (Omega) to extract plasmids. Insertion sequences were amplified for 15 cycles using the second-round primers. The remaining steps were carried out as for the ARE library. Libraries were sequenced on Illumina HiSeq X Ten platforms.

### Data processing

STARR-seq data were aligned to the *Nip* genome (IRGSP1.0, downloaded from Gramene) by Bowtie2 (Langmead and Salzberg, 2012). Mapped reads were filtered by SAMtools (Li *et al.*, 2009) with settings '-q 30 -f 2 -F 264'. Transcription peaks were identified using MACS2 (Zhang *et al.*, 2008) with settings '-g 4.4e8 -q 0.01 --keep-dup all' by defining ARE and input libraries as treatment and control, respectively. The detailed peak information is listed in Data S2.

ChIP-seq data (Zhao *et al.*, 2020) were obtained from NCBI GEO and processed followed the same procedures as for STARR-seq except that duplicates were removed using Picard (Broad Institute, 2019).

Distribution of FSPEs in genome was calculated using ChIPSeeker (Yu *et al.*, 2015). DNA motifs were called by The MEME Suite (Bailey *et al.*, 2009). Distributions of GC percentage and epigenetic marks in FSPEs were plotted using deepTools (Ramirez *et al.*, 2016).

Hi-C data of 30 °C treatment (Liu *et al.*, 2017) were obtained from NCBI GEO. Reads were processed using HiC-Pro (Servant *et al.*, 2015) with default settings. Merged file of valid pairs was then converted to the inputs of downstream tools. A/B compartments were determined using CscoreTool (Zheng and Zheng, 2018) in 10 kb resolution. Loop calling was achieved by juicer tools (Durand *et al.*, 2016) using 'hiccups' function under 5 kb resolution. Reported TAD information (Liu *et al.*, 2017) was directly used in our study. DNA interactions were visualized using WashU (Li *et al.*, 2019). The long-range format file was generated from valid pairs output based on digested fragments.

### Quantitative real-time PCR

Total RNA was extracted using an Eastep Super Total RNA Extraction Kit (Promega). Reverse transcription was performed using a Goscript<sup>TM</sup> Reverse Transcription System (Promega). Mixture of oligo(dT) and random six mers were used as the reverse transcription primer. Quantitative real-time PCR assays were performed using iTaq<sup>TM</sup> Universal SYBR® Green Supermix (Bio-Rad) on a CFX Connect Real-Time PCR Detection System (Bio-Rad). The level of gene expression was displayed using Livak method. *Ubiquitin* was used as the internal control for normalization. Amplification efficiency for each primer pair was validated using cDNA or target-containing plasmids.

### Acknowledgements

We thank Jialiang Huang (School of Life Sciences, Xiamen University, China) and Hao Wang (Rice Research Institute, Sichuan Agricultural University, China) for fruitful discussions. We thank Xiaohong Ma (Core Facility of Biomedical Science, Xiamen University, China) for providing us technical support of FACS. We thank Yuchao Cui (School of Life Sciences, Xiamen University, China) for managing growth chambers. We thank Xueke He and Jinpeng Huang (School of Life Sciences, Xiamen University, China) for providing partial rice seeds.

### Funding

This work was supported by the National Science Foundation of Fujian Province of China (No. 2022J02004) and the National Natural Science Foundation of China (32070266).

### Conflict of interest

The authors declare no conflicts of interest.

### Author contributions

WT designed and performed the research; WT, XH and XO analysed the data, and wrote the manuscript.

### Data availability statement

Sequencing data have been deposited in NGDC under BioProject PRJCA011002.

### References

Albright, L.M., Yanofsky, M.F., Leroux, B., Ma, D.Q. and Nester, E.W. (1987) Processing of the T-DNA of *Agrobacterium tumefaciens* generates border nicks and linear, single-stranded T-DNA. *J. Bacteriol.* **169**, 1046–1055.

Allen, M.D., Yamasaki, K., Ohme-Takagi, M., Tateno, M. and Suzuki, M. (1998) A novel mode of DNA recognition by a beta-sheet revealed by the solution structure of the GCC-box binding domain in complex with DNA. *EMBO J.* **17**, 5484–5496.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y. *et al.* (2014) An atlas of active enhancers across human cell types and tissues. *Nature*, **507**, 455–461.

Arnold, C.D., Gerlach, D., Stelzer, C., Boryn, Ł.M., Rath, M. and Stark, A. (2013) Genome-Wide quantitative enhancer activity maps identified by STARR-seq. *Science*, **339**, 1074–1077.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J. *et al.* (2009) MEME SUITE: Tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202–W208.

Bang, S.W., Park, S.H., Kim, Y.S., Choi, Y.D. and Kim, J.K. (2015) The activities of four constitutively expressed promoters in single-copy transgenic rice plants for two homozygous generations. *Planta*, **241**, 1529–1541.

Bargmann, B.O.R. and Birnbaum, K.D. (2010) Fluorescence activated cell sorting of plant protoplasts. *J. Vis. Exp.* **36**, e1673.

Broad Institute. (2019). *Picard Toolkit*. <http://broadinstitute.github.io/picard/>

Cao, S., Kumimoto, R.W., Gnesutta, N., Calogero, A.M., Mantovani, R. and Holt, B.F. (2014) A distal CCAAT/NUCLEAR FACTOR Y complex promotes chromatin looping at the FLOWERING LOCUS T promoter and regulates the timing of flowering in Arabidopsis. *Plant Cell*, **26**, 1009–1017.

Delaney, S.K., Orford, S.J., Martin-Harris, M. and Timmis, J.N. (2007) The fiber specificity of the cotton FSltp4 gene promoter is regulated by an AT-rich promoter region and the AT-hook transcription factor GhAT1. *Plant Cell Physiol.* **48**, 1426–1437.

Durand, N.C., Shamim, M.S., Machol, I., Rao, S.S., Huntley, M.H., Lander, E.S. and Aiden, E.L. (2016) Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98.

Fejes, E., Pay, A., Kanevsky, I., Szell, M., Adam, E., Kay, S. and Nagy, F. (1990) A 268 bp upstream sequence mediates the circadian clock-regulated transcription of the wheat Cab-1 gene in transgenic plants. *Plant Mol. Biol.* **15**, 921–932.

Giuliano, G., Pichersky, E., Malik, V.S., Timko, M.P., Scolnik, P.A. and Cashmore, A.R. (1988) An evolutionarily conserved protein binding sequence upstream of a plant light-regulated gene. *Proc. Natl. Acad. Sci. USA*, **85**, 7089–7093.

Gu, B., Swigut, T., Spencley, A., Bauer, M.R., Chung, M., Meyer, T. and Wysocka, J. (2018) Transcription-coupled changes in nuclear mobility of mammalian cis-regulatory elements. *Science*, **359**, 1050–1055.

Herrera-Estrella, A., Chen, Z.M., Van Montagu, M. and Wang, K. (1988) VirD proteins of *Agrobacterium tumefaciens* are required for the formation of a covalent DNA-protein complex at the 5' terminus of T-strand molecules. *EMBO J.* **7**, 4055–4062.

Herrera-Estrella, A., Van Montagu, M. and Wang, K. (1990) A bacterial peptide acting as a plant nuclear targeting signal: The amino-terminal portion of *Agrobacterium* VirD2 protein directs a beta-galactosidase fusion protein into tobacco nuclei. *Proc. Natl. Acad. Sci. USA*, **87**, 9534–9537.

Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-Andre, V., Sigova, A.A., Hoke, H.A. *et al.* (2013) Super-enhancers in the control of cell identity and disease. *Cell*, **155**, 934–947.

Jores, T., Tonnes, J., Dorrity, M.W., Cuperus, J.T., Fields, S. and Queitsch, C. (2020) Identification of plant enhancers and their constituent elements by STARR-seq in tobacco leaves. *Plant Cell*, **32**, 2120–2131.

Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.

Lemp, N.A., Hiraoka, K., Kasahara, N. and Logg, C.R. (2012) Cryptic transcripts from a ubiquitous plasmid origin of replication confound tests for cis-regulatory function. *Nucleic Acids Res.* **40**, 7280–7290.

Li, D., Hsu, S., Purushotham, D., Sears, R.L. and Wang, T. (2019) WashU Epigenome Browser update 2019. *Nucleic Acids Res.* **47**, W158–W165.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Lin, C.S., Hsu, C.T., Yang, L.H., Lee, L.Y., Fu, J.Y., Cheng, Q.W., Wu, F.H. *et al.* (2018) Application of protoplast technology to CRISPR/Cas9 mutagenesis: From single-cell mutation detection to mutant plant regeneration. *Plant Biotechnol. J.* **16**, 1295–1310.



- Liu, C., Cheng, Y.J., Wang, J.W. and Weigel, D. (2017) Prominent topologically associated domains differentiate global chromatin packing in rice from Arabidopsis. *Nat. Plants*, **3**, 742–748.
- Lu, H., Yu, D., Hansen, A.S., Ganguly, S., Liu, R., Heckert, A., Darzacq, X. et al. (2018) Phase-separation mechanism for C-terminal hyperphosphorylation of RNA polymerase II. *Nature*, **558**, 318–323.
- Ma, X., Zhang, Q., Zhu, Q., Liu, W., Chen, Y., Qiu, R., Wang, B. et al. (2015) A robust CRISPR/Cas9 system for convenient, high-efficiency multiplex genome editing in monocot and dicot plants. *Mol. Plant*, **8**, 1274–1284.
- Mifsud, B., Tavares-Cadete, F., Young, A.N., Sugar, R., Schoenfelder, S., Ferreira, L., Wingett, S.W. et al. (2015) Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.* **47**, 598–606.
- Muerdter, F., Boryn, L.M., Woodfin, A.R., Neumayr, C., Rath, M., Zabidi, M.A., Pagani, M. et al. (2018) Resolving systematic errors in widely used enhancer activity assays in human cells. *Nat. Methods*, **15**, 141–149.
- Neumayr, C., Pagani, M., Stark, A. and Arnold, C.D. (2019) STARR-seq and UMI-STARR-seq: assessing enhancer activities for genome-wide-, high-, and low-complexity candidate libraries. *Curr. Protoc. Mol. Biol.* **128**, e105.
- Niu, L., Wan, J., Sun, J., Huang, Y., He, N., Li, L., and Hou, C. (2020) Resolving a systematic error in STARR-seq for quantitative enhancer activity mapping. *bioRxiv*. <https://doi.org/10.1101/2020.10.20.346908>.
- Ohme-Takagi, M. and Shinshi, H. (1995) Ethylene-inducible DNA binding proteins that interact with an ethylene-responsive element. *Plant Cell*, **7**, 173–182.
- Ortiz-Ramírez, C., Arevalo, E.D., Xu, X., Jackson, D.P. and Birnbaum, K.D. (2018) An efficient cell sorting protocol for maize protoplasts. *Curr. Protoc. Plant Biol.* **3**, e20072.
- Ow, D.W., Jacobs, J.D. and Howell, S.H. (1987) Functional regions of the cauliflower mosaic virus 35S RNA promoter determined by use of the firefly luciferase gene as a reporter of promoter activity. *Proc. Natl. Acad. Sci. USA*, **84**, 4870–4874.
- Park, S.H., Yi, N., Kim, Y.S., Jeong, M.H., Bang, S.W., Choi, Y.D. and Kim, J.K. (2010) Analysis of five novel putative constitutive gene promoters in transgenic rice plants. *J. Exp. Bot.* **61**, 2459–2467.
- Petersen, B.L., Möller, S.R., Mravec, J., Jørgensen, B., Christensen, M., Liu, Y., Wandall, H.H. et al. (2019) Improved CRISPR/Cas9 gene editing by fluorescence activated cell sorting of green fluorescence protein tagged protoplasts. *BMC Biotechnol.* **19**, 36.
- Qian, W., Tan, G., He, S., Zhang, X., Zhang, X., Gao, Y., Liu, H. et al. (2007) Identification of a new AT-rich-element binding factor PsATF1 and its combined effect with PsGBF on the activation of PsCHS1 promoter. *Front. Biosci.* **12**, 1670–1679.
- Ramirez, F., Ryan, D.P., Gruning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S. et al. (2016) deepTools2: A next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–W165.
- Schoenfelder, S., Furlan-Magaril, M., Mifsud, B., Tavares-Cadete, F., Sugar, R., Javierre, B.M., Nagano, T. et al. (2015) The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Res.* **25**, 582–597.
- Servant, N., Varoquaux, N., Lajoie, B.R., Viara, E., Chen, C.J., Vert, J.P., Heard, E. et al. (2015) HiC-Pro: An optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259.
- Simpson, J., Schell, J., Montagu, M.V. and Herrera-Estrella, L. (1986) Light-inducible and tissue-specific pea lhpc gene expression involves an upstream element combining enhancer- and silencer-like properties. *Nature*, **323**, 551–554.
- Soares, L.M., He, P.C., Chun, Y., Suh, H., Kim, T. and Buratowski, S. (2017) Determinants of histone H3K4 methylation patterns. *Mol. Cell*, **68**, 773–785.
- Sun, J., He, N., Niu, L., Huang, Y., Shen, W., Zhang, Y., Li, L. et al. (2019) Global quantitative mapping of enhancers in rice by STARR-seq. *Genom. Proteom. Bioinf.* **17**, 140–153.
- Tippens, N.D., Liang, J., Leung, A.K., Wierbowski, S.D., Ozer, A., Booth, J.G., Lis, J.T. et al. (2020) Transcription imparts architecture, function and logic to enhancer units. *Nat. Genet.* **52**, 1067–1075.
- Tjaden, G. and Coruzzi, G.M. (1994) A novel AT-Rich DNA binding protein that combines an HMG I-like DNA binding domain with a putative transcription domain. *Plant Cell*, **6**, 107–118.
- Wang, H., Li, S., Li, Y., Xu, Y., Wang, Y., Zhang, R., Sun, W. et al. (2019) MED25 connects enhancer-promoter looping and MYC2-dependent activation of jasmonate signalling. *Nat. Plants*, **5**, 616–625.
- Weintraub, A.S., Li, C.H., Zamudio, A.V., Sigova, A.A., Hannett, N.M., Day, D.S., Abraham, B.J. et al. (2017) YY1 is a structural regulator of enhancer-promoter loops. *Cell*, **171**, 1573–1588.
- You, M.K., Lim, S., Kim, M., Jeong, Y.S., Lee, M. and Ha, S. (2014) Improvement of the fluorescence intensity during a flow cytometric analysis for rice protoplasts by localization of a green fluorescent protein into chloroplasts. *Int. J. Mol. Sci.* **16**, 788–804.
- Yu, G., Wang, L.G. and He, Q.Y. (2015) ChIPseeker: An R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics*, **31**, 2382–2383.
- Zhang, Y., Liu, T., Meyer, C.A., Eeckhoutte, J., Johnson, D.S., Bernstein, B.E., Nussbaum, C. et al. (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137.
- Zhang, Y., Su, J., Duan, S., Ao, Y., Dai, J., Liu, J., Wang, P. et al. (2011) A highly efficient rice green tissue protoplast system for transient gene expression and studying light/chloroplast-related processes. *Plant Methods*, **7**, 30.
- Zhao, L., Xie, L., Zhang, Q., Ouyang, W., Deng, L., Guan, P., Ma, M. et al. (2020) Integrative analysis of reference epigenomes in 20 rice varieties. *Nat. Commun.* **11**, 2658.
- Zheng, X. and Zheng, Y. (2018) CscoreTool: Fast Hi-C compartment analysis at high resolution. *Bioinformatics*, **34**, 1568–1570.
- Zhu, B., Zhang, W., Zhang, T., Liu, B. and Jiang, J. (2015) Genome-Wide prediction and validation of intergenic enhancers in arabidopsis using open chromatin signatures. *Plant Cell*, **27**, 2415–2426.

## Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

**Data S1** SPE22-homologous sequences shown in both reported replicates and the remaining top 1000 for replicate 1.

**Data S2** Prediction results of the improved method.

**Data S3** Detailed information of primers used in this study.

**Data S4** Potential chromatin loops covering FSPEs.

**Figure S1** Structure of the original STARR-seq vector.

**Figure S2** Validation of four SPEs with relatively low FE.

**Figure S3** Strategy for flow cytometry analysis.

**Figure S4** Detection of cryptic transcription from ORIs in the transient expression system of rice protoplasts.

**Figure S5** Strategy for FACS.

**Figure S6** Overlap of peaks between improved and traditional method.

**Figure S7** Fluorescence microscopy of protoplasts transfected by FSPEs with low FE.

**Figure S8** Flow cytometry analysis of protoplasts transfected by FSPEs with low FE.

**Figure S9** Overlap of peaks generated from two replicates following the improved method.

**Figure S10** Distribution of FSPEs in A/B compartments.

**Figure S11** Distribution of STARR-seq signal in TAD boundaries.

**Figure S12** Distribution of FSPEs within and outside of TADs.

**Figure S13** Distribution of regions interacted with FSPEs in genome.

**Figure S14** Sanger sequencing results of the region targeted by FSPE11 sgRNA-4.

**Figure S15** Changes of gene expression after disrupting FSPEs.

**Figure S16** Scale diagrams and contact maps of chromatin around FSPE5.

**Figure S17** Scale diagrams and contact maps of chromatin around FSPE6.

**Figure S18** Scale diagrams and contact maps of chromatin around FSPE10.

**Figure S19** Scale diagrams and contact maps of chromatin around FSPE11.

**Table S1** Status of reads after mapping.