BRIEF REPORT

Agro-STARR-seq enables *in planta* genome-wide enhancer discovery and AI-empowerd enhancer design

Yongzhou Bao[1,2,†], Xiwen Zhang[1,2,†], Yanwei Song[1,2,†], Huiming Li[1,2,4], Junyan Qi[1,2], Hongqiu Cui[1,2,3], Na Li[3], Yun Chen[1,2], Weigang Zheng[1,2], Lingna Xu[1,2], Yuanyuan Zhang[1,2], Choulin Chen[1,2], Fangjie Zhu[5], Shaoqun Zhou[3,*], Zhaohong Li[1,2,6,*], Yuwen Liu[1,2,*]

1State Key Laboratory of Genome and Multi-omics Technologies, Shenzhen Branch, Guangdong Laboratory for Lingnan Modern Agriculture, Key Laboratory of Livestock and Poultry Multi-Omics of MARA, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518124, China.

2 Innovation Group of Pig Genome Design and Breeding, Research Centre for Animal Genome, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China.

3 Shenzhen Branch, Guangdong Laboratory of Lingnan Modern Agriculture, Key Laboratory of Synthetic Biology, Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China.

4 School of Life Science, Henan University, Kaifeng, 475004, China.

5 Haixia Institute of Science and Technology, National Engineering Research Center of JUNCAO, Fujian Provincial Key Laboratory of Haixia Applied Plant Systems Biology, Fujian Agriculture and Forestry University, Fuzhou 350002, Fujian, China.

6 Key Laboratory of Animal Genetics, Breeding and Reproduction of Shaanxi Province, College of Animal Science and Technology, Northwest A&F University, Yangling, Shaanxi, China.

[†] These authors contributed equally to this work.

*Correspondence:

Shenzhen Branch, Guangdong Laboratory for Lingnan Modern Agriculture, Key Laboratory of Livestock and Poultry Multi-Omics of MARA, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China; E-mail: liuyuwen@caas.cn

E-mail: zhaohongli90@nwafu.edu.cn

E-mail: zhoushaoqun@caas.cn

Running title: Yongzhou Bao *et al*. / Identification and engineering enhancers *in planta*

Abstract

Comprehensive enhancer screening is essential for dissecting gene regulatory networks. However, methods for such screening remain limited in plants. We present Agro-STARR-seq, an *in planta* enhancer screening platform using *Agrobacterium*-mediated delivery of STARR-seq libraries into *Nicotiana benthamiana* leaves. In the *Arabidopsis* genome, we identified 14,715 sequences with enhancer activity, of which 91.59% were located outside of open chromatin regions. This largest *in planta* enhancer dataset to date enabled AI-assisted design of highly active synthetic enhancers. For the large *N. benthamiana* genome, we pre-processed the genomic DNA with an Assay for Transposase Accessible Chromatin (ATAC) procedure before the Agro-STARR-seq protocol,and found that only one-third of open chromatin exhibited enhancer activity, 76.00% of which showed no canonical H3K27ac histone marks. Footprinting analysis further enables systematic linkage of transcription factor (TF) binding to enhancer function. Overall, Agro-STARR-seq provides a novel platform for dissecting and engineering plant enhancers.

Keywords: *In planta*, Enhancer Screening, STARR-seq, Deep Learning, Enhancer Design

Background

Gene expression governs plant cell fate determination, drives morphological development, and mediates adaptive responses to environmental stimuli [1]. This process is dynamically regulated by enhancers, which confer spatiotemporal precision and fine-tune transcriptional dynamics across developmental and environmental contexts [2]. Systematic identification of functional enhancers will facilitate the dissection of gene regulatory networks underlying complex traits and the *de novo* design of synthetic enhancers with desired activity.

Traditionally, plant enhancers have been inferred from epigenomic signatures, such as histone modifications and chromatin accessibility [3]. While these approaches

65 are valuable, they rely on correlative biochemical markers and do not directly measure
66 the enhancer activity. Methods based on DNA-binding activity of nuclear TFs were
67 also developed [4], yet DNA-binding does not guarantee transcriptional-regulatory
68 activity. The advent of massively parallel reporter assays (MPRAs) [5], exemplified by
69 Self-Transcribing Active Regulatory Region Sequencing (STARR-seq) [6], enabled
70 genome-wide, quantitative assessment of enhancer activity. Originally developed in
71 *Drosophila melanogaster* [6] and *Homo sapiens* [7], STARR-seq has since been
72 adapted to plants, including *Arabidopsis* [8], rice [9,10], maize [11,12] and barley [13],
73 representing a major advance in functional plant genomics.

74 However, existing plant whole-genome STARR-seq studies have been limited to
75 protoplasts (Supplemental Table S1). While protoplasts enable efficient delivery of
76 enhancer-screening plasmid libraries—an essential requirement for STARR-seq—the
77 enzymatic removal of cell walls and the loss of native tissue architecture could
78 profoundly alter the intricate cellular states transcriptional regulatory networks in the
79 multicellular context. Consequently, enhancers identified in protoplasts may not
80 accurately reflect their activity in intact tissues and may fail to capture many genuine
81 enhancers, leaving a major gap in our understanding of plant gene regulation *in planta*.
82 To date, genome-wide STARR-seq enhancer profiling directly *in planta* has not been
83 achieved.

84 To address this limitation, we developed Agro-STARR-seq (*Agrobacterium*-
85 mediated Self-Transcribing Active Regulatory Region Sequencing, Fig. 1A), a
86 technology that enables high-throughput, genome-wide profiling of enhancer activity
87 directly in plant leaves. This approach preserves the native regulatory environment
88 while providing quantitative assessment of enhancer strength. Applying Agro-STARR-
89 seq to the entire *Arabidopsis thaliana* genome, we demonstrated its capacity for
90 comprehensive enhancer discovery and highlighted the utility of the resulting dataset
91 for AI- empowered *de novo* design of synthetic enhancers that surpass the activity of
92 natural sequences [14]. We further applied Agro-STARR-seq to systematically identify
93 enhancers within the open chromatin regions of *Nicotiana benthamiana* leaves,

94 providing new insights into the epigenomic features and the *cis*-regulatory lexicon of
95 plant enhancers.

96 Results and discussion

97 Development and Validation of Agro-STARR-seq System

98 The Agro-STARR-seq method exploits *Agrobacteria tumefaciens* to deliver
99 STARR-Seq enhancer screening plasmids into intact leaves of *Nicotiana benthamiana.*
100 To enable this, we modified the conventional STARR-seq vector for compatibility with
101 *Agrobacterium*-mediated delivery (Fig. S1A and S1B, detailed in Supplemental notes).
102 As a proof of concept, we constructed an enhancer-screening library from randomly
103 fragmented *Arabidopsis thaliana* Col-0 genomic DNA with an average fragment size
104 of 450 bp, taking advantage of its relatively small genome size (ca. 135 Mbp) to
105 streamline experimental procedures (Fig. S2A). To evaluate library complexity, we
106 sequenced both the plasmid input library and the corresponding *Agrobacterium* input
107 library after transformation. The latter contained at least 4.8 million independent
108 candidate fragments (Supplemental Table S2), achieving 94% coverage of the *A.*
109 *thaliana* Col-0 genome (Fig. S2B). Simulation-based analysis further indicated that the
110 library complexity had reached saturation (Fig. S2C). Fragment abundance between
111 plasmid and *Agrobacterium* input libraries showed high correlation (Pearson's
112 correlation coefficient, PCC = 0.9, P-value < $2.2e^{-16}$, Fig. S3A), confirming minimal
113 sequence bias during transformation. After infiltrating *N. benthamiana* leaves with the
114 screening library–containing *Agrobacteria*, transcript abundance from STARR-seq
115 plasmids showed strong reproducibility across two biological replicates (PCC = 0.92,
116 P-value < $2.2e^{-16}$, Figure S3B). Collectively, these results demonstrate that Agro-
117 STARR-seq achieves library complexity and coverage comparable to established
118 STARR-seq protocols published from our previous work [7,15], validating the
119 robustness of this method.

120

121 Genome-wide Enhancer Atlas of the *Arabidopsis* Genome

122 To quantitatively assess enhancer activity, we compared the enrichment of cDNA
123 fragments of the output libraries relative to *Agrobacterium* input library DNA
124 fragments across genomic bins. Sequences exhibiting increased self-transcription are
125 indicative of stronger enhancer activity (Fig. 1B and S4A-G). Using this approach, we
126 identified 14,715 candidate enhancers *in planta*, with activities highly reproducible
127 across replicates. Notably, only 1,237 (8.4%) of these enhancers were located within
128 previously annotated open chromatin regions (Fig. 1C) [16], consistent with previous
129 *in vitro* studies in rice [11] and human [7]. This finding suggests that a substantial
130 proportion of sequences with *in planta* enhancer potential reside outside open
131 chromatin regions.

132 We next characterized the genetic and epigenetic features of the enhancers. *In*
133 *planta A. thaliana* enhancers (mean peak length = 200 bp; Fig. S5A) were distributed
134 across all five *A. thaliana* chromosomes, with certain genomic hotspots exhibiting
135 higher densities (Fig. S5B and S5C). Genomic annotation revealed that 27.5% of
136 identified enhancers were located promoter-proximal regions, while 24.8% resided in
137 distal intergenic regions (Fig. 1D). This preferential distribution near genes likely
138 reflects the compact genome architecture of *A. thaliana*. Integration with ChIP-seq
139 profiles for five histone modifications (H3K27ac, H3K9ac, H3K27me3, H3K4me1, and
140 H3K4me3; Supplemental Notes) showed that promoter-proximal enhancers co-
141 localized with H3K4me3, H3K27ac, and H3K9ac peaks, while distal enhancers were
142 associated with H3K4me1 (Fig. 1E and S6), consistent with previously reported ChIA-
143 PET results for *A. thaliana* [16].

144 Motif analysis revealed 108 significantly enriched transcription factors (TFs)
145 DNA binding motifs (q-value < 0.05), including members of the ERF family, which are
146 widely involved in the regulation of plant growth, development, and environmental
147 responses (Fig. 1F). Many enhancers co-localized with transposable elements (TEs),
148 predominantly with LINE/L1 (4,181), LTR/Copia class I retrotransposons (941), and
149 class II transposons including DNA (1,781) and DNA/HAT (1,366) (One-sided
150 hypergeometic test; Fig. 1G), indicating specific TE types may contribute significantly

151 to enhancer formation in *A. thaliana*. Furthermore, a pronounced differential
152 enrichment of class II DNA transposons (1,781) between active and inactive accessible
153 chromatin regions was observed, suggesting that specific TEs contribute differently to
154 enhancer activity and chromatin accessibility (Fig. S7). Evolutionary rate analysis
155 demonstrated these enhancers exhibit significantly higher conservation than non-
156 enhancers across different functional genomics categories (Fig. 1H). Collectively, these
157 results highlight the functional and biological significance of the enhancers identified
158 through our Agro-STARR-seq approach *in planta*.

159

160 **AI-empowered *de novo* plant enhancer design from *in planta* activity**

161 Unlike previous protoplast-based plant STARR-seq studies, this proof-of-concept
162 application of Agro-STARR-seq scanned the entire *A. thaliana* genome, generating the
163 largest *in planta* enhancer dataset to date—an invaluable resource for AI-driven
164 enhancer design in plant leaves. Using this dataset, we applied our recently developed
165 computational framework, DREAM (DNA *cis*-Regulatory Elements with Controllable
166 Activity Design PlatforM) (Fig. 1A) [14] to design synthetic enhancers with high
167 regulatory activity in *N. benthamiana* leaves. The enhancer prediction module of
168 DREAM demonstrated robust generalization capability, achieving a PCC of 0.53 (P-
169 value $< 2.2e^{-16}$, Fig. 1I, detailed in Supplemental notes) between predicted and observed
170 enhancer activities in a hold-out chromosome dataset. Its sequence design module uses
171 a genetic algorithm to iteratively maximize DREAM-predicted enhancer activity,
172 generating synthetic enhancers predicted to exhibit exceptionally high regulatory
173 activity. Simulated evolution analysis showed that enhancer activity reached saturation
174 after approximately 100 generations (Fig. S8). From the 100th generation, we randomly
175 selected a synthetic enhancer which contained several key TF motifs, including
176 ERF003 and ERF095, known activators of gene expression. A model interpretability
177 analysis confirmed the substantial contribution of TFs motifs to enhancer activity
178 prediction (Fig. 1J). Furthermore, *in planta* luciferase reporter assays (detailed
179 experimental methods are provided in the Supplemental notes, with vector constructs

180 shown in Fig. S9A and S9B) validated the functionality of this DREAM-designed
181 synthetic enhancer (DREAMer003), demonstrating a 5.8-fold activity increase when
182 positioned upstream of a viral 35S promoter (One-sided t-test, P-value < 0.05, Fig. 1K
183 and Supplemental Table S3).

184

185 **ATAC-Agro-STARR-seq identifies enhancers of the *N. benthamiana* genome**

186 To demonstrate the versatility of our Agro-STARR-seq platform beyond *A.*
187 *thaliana*, we extended its application to identify enhancers in open chromatin regions
188 of *N. benthamiana* leaves (ATAC-Agro-STARR-seq, Fig. 2A). Notably, this setup
189 provides a cellular environment fully matched to the genomic sequences under
190 investigation, offering a robust framework to dissect the genetic, epigenetic, and
191 functional properties of plant enhancers. To enrich for open chromatin regions in the
192 enhancer screening vector, we adapted the ATAC-seq protocol and generated the first
193 high-quality ATAC-seq library from *N. benthamiana* leaves (Supplemental Table S4)
194 [17], comprising ~88 million fragments largely representing nucleosome-free regions
195 (NFRs, Fig. S10A and S10B). This library displayed hallmark epigenetic features,
196 including strong enrichment at transcription start sites (TSSs, Fig. S10C), peak sizes of
197 250–750 bp (Fig. S10D), and predominant localization in distal intergenic and promoter
198 regions (Fig. 2B). Functionally, genes proximal to ATAC-seq peaks exhibited
199 significantly higher expression than distal genes (Fig. S10E and S10F), confirming the
200 relevance of these accessible regions.

201 Cloning this high-quality ATAC-seq library into the Agro-STARR-seq vector
202 produced highly consistent fragment coverages across subsequent plasmid input
203 libraries and *Agrobacterium* input libraries, with an average complexity exceeding 10
204 million unique fragments (Fig. S11 and S12; Supplemental Table S3). We then
205 infiltrated the ATAC-Agro-STARR-seq *Agrobacterium* input libraries into *N.*
206 *benthamiana* leaves, generating output libraries and identifying 39,394 enhancers in
207 open chromatin regions (hereafter termed active accessible regions, Fig. 2C). Notably,
208 only ~33.2% of accessible chromatin peaks contained at least one active enhancer,

209 highlighting the limitations of using ATAC-seq signal alone when annotating plant
210 enhancers [18]. The genomic distribution of active accessible regions closely mirrored
211 that of ATAC-seq peaks, indicating no strong positional bias within open chromatin
212 (Fig. 2B). Active accessible regions were associated with higher expression of
213 neighboring genes, demonstrating their functional relevance in the native chromatin
214 context (Fig. S13A).

215 Interrogation of published histone modification datasets (H3K27ac, H3K4me3,
216 H3K27me3, and H3K9me2) revealed that these active regions were significantly
217 enriched for marks of active chromatin (H3K27ac, H3K4me3) but not the repressive
218 (H3K27me3) or heterochromatin (H3K9me2)-associated mark (Fig. S13B; Extended
219 Data Table). We further observed distinct patterns of histone mark enrichment at
220 proximal versus distal sites: while the promoter-associated H3K4me3 mark exhibited
221 higher signals specifically in proximal active regions, H3K27ac was enriched at both
222 proximal and distal active regions (Fig. 2D).

223 Notably, we found that 76.00% of enhancers identified within open chromatin
224 regions lacked H3K27ac, a canonical enhancer-associated histone modification,
225 revealing tens of thousands of previously hidden enhancers in the plant genome. This
226 is consistent with previous studies in human and mouse [19], and underscores the
227 limitation of relying solely on active histone marks to annotate plant enhancers.

228 **Integration of accessibility and activity reveals TF binding and functional**
229 **networks**

230 To identify TFs contributing to enhancer activity, we analyzed TF DNA binding
231 motifs enriched in all, proximal, and distal active accessible regions. Many TFs,
232 including members of the TCP, ERF, bZIP, and bHLH families, were significantly
233 enriched in these regions (Fig. 2E). Notably, TCPs, previously reported as major
234 contributors to distal chromatin accessibility quantitative trait loci (caQTLs) in maize,
235 were most significantly enriched in distal active accessible regions (Fig. S14) [12].
236 These results demonstrate that our ATAC-Agro-STARR-seq platform provides a robust
237 framework for providing insights into how TFs contribute to plant enhancer function.

238 To further link TF binding with enhancer activity, we integrated the footprinting-
239 based TF binding site predictions from ATAC-seq with the precise enhancer activity
240 measurements of STARR-seq. By combining the strengths of both approaches, Agro-
241 STARR-seq offers a unique platform to systematically investigate the relationship
242 between TF binding and enhancer activity, providing an integral perspective to study
243 concerted *cis-trans* transcriptional regulation in plants.

244 To validate this approach, we generated Tn5-bias–corrected cut site signal files
245 from ATAC-Agro-STARR-seq data and predicted TF binding sites using footprint
246 scores. TFs with available ChIP-seq data in *N. benthamiana* leaves served as a gold
247 standard for evaluating predicted versus actual binding sites. For example, for TCP20,
248 corrected cut site signals clearly distinguished predicted TF-bound from unbound sites
249 (Fig. 2F and S15), a trend consistent across other TFs (Fig. S16). ChIP-seq confirmed
250 preferential binding of TCP20 to the predicted sites, and integration with STARR-seq
251 data demonstrated significantly higher enhancer activity at these bound sites compared
252 to unbound sites, providing genome-wide evidence that TCP20 binding generally
253 activates gene expression.

254 Subsequently, we focused on the TCP regulatory network and examined the
255 functions of genes associated with predicted TCP-bound active accessible peaks. We
256 found that TCPs regulate genes involved in circadian rhythm and light responses (Fig.
257 2G and Fig. S17). TCPs have been reported to bind directly upstream of the CCA1
258 transcription start site, a morning gene in *A. thaliana*, and activate its expression [20,21].
259 In *N. benthamiana*, CCA1 also functions in circadian rhythm regulation [22]. Although
260 CCA1 is not annotated in the current *N. benthamiana* reference genome (Niben261),
261 we identified and assembled a CCA1-like transcript from RNA-Seq data, mapped to
262 Niben261 Chr12:120,946,470–120,948,272 (Supplemental Table S5 and S6).
263 Strikingly, we found that near the promoter of this transcript, the predicted TCP-bound
264 sites upstream of this putative NbCCA1 (as supported by published TCP ChIP-seq
265 signals, also exhibit Agro-STARR-Seq enhancer activity (Fig. 2H).

266 Collectively, our findings establish *in planta* ATAC-Agro-STARR-seq as a
267 powerful method for delineating regulatory networks, as it directly connects

268 transcriptional activation by upstream TFs to the expression of their downstream target
269 genes.

Conclusions

Here, we present Agro-STARR-seq, a high-throughput method that addresses the challenge of transitioning high-throughput enhancer screening from isolated protoplast systems to living plants. The approach utilizes *Agrobacterium*-mediated delivery into intact *N. benthamiana* leaves—a widely adopted model plant that provides a rapid, scalable, and transient gene expression platform, thus establishing an ideal *in planta* chassis. Whereas previous *in planta* methods screened only thousands of sequences for enhancer activity (Supplemental Table S1) [23], Agro-STARR-seq scales this up by more than 100-fold, enabling genome-wide enhancer discovery across diverse plant species. Beyond uncovering novel genetic and epigenetic features of plant enhancers, Agro-STARR-seq generates large-scale *in planta* enhancer datasets suitable for AI training, offering a unique platform for precise engineering of transcriptional regulatory networks in plant genomes.

For method development, we first screened enhancer activity of the entire *A. thaliana* genome. *A. thaliana* was chosen as the source of test sequences due to its small genome size, which facilitates methodological development and efficient validation. Nonetheless, the substantial increase in the scale of our methodology enabled a transition from assessing enhancer activities in a limited set of candidates to broader genomic regions or genomic regions enriched through functional genomics techniques such as ATAC-seq and CUT&Tag. This versatility establishes Agro-STARR-seq as a broadly applicable platform for functional annotation of plant enhancers.

To further explore the genetic and epigenetic features of plant enhancers, we profiled enhancer activity in the accessible chromatin regions in *N. benthamiana* by ATAC-Agro-STARR-Seq. In this study, we successfully constructed ATAC-seq libraries in *N. benthamiana* for the first time, identifying 99,851 peaks with a FRiP (Fraction of Reads in Peaks) score of 34.90% (Supplemental Table S4). Integration of footprinting signals of ATAC-seq and Agro-STARR-seq enables the simultaneous capture of multiple regulatory layers—including chromatin accessibility, transcription

factor binding, and enhancer activity—within a single assay. As demonstrated in *N. benthamiana*, this integrated profiling offers a powerful strategy to unravel multilayered gene regulatory networks, exemplified by our discovery of a TCP-mediated circadian regulatory circuit. Strikingly, we found that 76.00% of enhancers in *N. benthamiana* leaf accessible chromatin regions lack H3K27ac marks, revealing—for the first time in plants—the widespread presence of 'hidden' enhancers, similar to observations in animals [19]. This highlights the complexity of inferring enhancer activity from histone modifications and underscores the value of Agro-STARR-seq in plant genetics. Current large-scale, histone mark-based enhancer screens may overlook many functionally important enhancers and enhancer-modulating variants that shape plant phenotypes.

The successful integration of our high-throughput screening platform with the DREAM computational framework marks a paradigm shift in plant synthetic biology. By combining the largest *in planta* enhancer activity dataset to date with deep learning–driven design, we engineered synthetic enhancers exhibiting a 5.8-fold increase in regulatory activity over the conventional viral 35S promoter, long a benchmark for transgenic expression. This unprecedented enhancement demonstrates the power of bridging large-scale functional genomics with predictive AI models to surpass the limitations of natural cis-regulatory elements. Since N. *benthamiana* responds to multiple stresses, such as heat and drought, generating Agro-STARR-seq data under diverse conditions will enable AI-guided design of environment-responsive enhancers—crucial for developing ‘smart’ plants that fine-tune critical gene expression to adapt to environmental challenges.

In conclusion, our work opens new avenues for deciphering gene regulatory networks and accelerates the engineering of tailored traits in plants, marking a major step forward in plant functional genomics and plant synthetic biology.

Methods

326 **Construction of Agro-STARR-Seq enhancer screening vector**

327 To construct a vector suitable for Agro-STARR-seq experiments, we modified the
328 reporter gene vector based on the backbone of the pGreen II plasmid (Supplemental
329 Table S7, Fig. S1). In brief, the CaMV 35S promoter was replaced with the CMV 35S
330 minimal promoter, and the CaMV poly (A) signal was removed. Additionally, several
331 key functional elements were incorporated, including an intron, EGFP sequence,
332 terminator elements, and a CmR-ccdB cassette. These sequences were arranged
333 sequentially and are underlined in Fig. S1.

334

335 **Preprocessing of *Arabidopsis* whole-genome plasmid input Library**

336 Three micrograms of *Arabidopsis thaliana* Col-0 genomic DNA was diluted to 10
337 ng/μL and fragmented using ultrasonication (Diagenode). The fragmented DNA was
338 size-selected (450–500 bp) through a 1% agarose gel (D4008; ZYMO). DNA fragments
339 underwent end repair, adapter ligation, and purification (ND607; Vazyme Biotech)
340 (N411; Vazyme Biotech). Adapter-ligated DNA was then amplified using polymerase
341 chain reaction (PCR) with the KAPA HiFi HotStart ReadyMix PCR Kit (KK2602;
342 Roche), with homologous arms incorporated for cloning (Supplemental Table S8). The
343 PCR cycling conditions were: 98°C for 45 s; 15 cycles of 98°C for 15 s, 65°C for 30 s,
344 and 72°C for 70 s. Amplified products were purified and used for downstream cloning
345 into the Agro-STARR-seq vector (T1030S; NEB).

346

347 **Preprocessing of *N. benthamiana* ATAC-seq plasmid input library**

348 Leaves (500 g) were harvested from two vigorously growing *Nicotiana*
349 *benthamiana* plants (approximately one month old, pre-flowering stage). Nuclei were
350 isolated using the Plant Sample Nuclei Isolation Kit (52305-10, BioYou) according to
351 the manufacturer's protocol. Assay for transposase-accessible chromatin (ATAC) was
352 conducted using Chromatin Profile Kit (N248-01A, Novoprotein) following the
353 manufacturer's instructions, respectively.

354 *N. benthamiana* ATAC-seq library was then amplified using polymerase chain
355 reaction (PCR) with the KAPA HiFi HotStart ReadyMix PCR Kit (KK2602; Roche),
356 with homologous arms incorporated for cloning (Supplemental Table S8). The PCR
357 cycling conditions were: 98°C for 45 s; 15 cycles of 98°C for 15 s, 65°C for 30 s, and
358 72°C for 70 s. Amplified products were purified and used for downstream cloning into
359 the Agro-STARR-seq vector (T1030S; NEB).

360

361 **Construction of the plasmid input libraries**

362 The *A. thaliana* and *N. benthamiana* purified PCR products, now containing
363 homologous arms, were inserted into the linearized vector backbone (digested with
364 AgeI and SacI) via Gibson assembly (E5510S; NEB). After recombination, all products
365 were pooled, and 5 µL of DNA constructs were subsequently transformed into MegaX
366 DH10B™ T1R electrocompetent cells (C640003; ThermoFisher Scientific). A total of
367 four Gibson assembly and transformation experiments were performed. All transformed
368 Escherichia coli cultures were pooled and grown in 1 L of LB medium until the optical
369 density (OD) reached 1.0. The plasmid input libraries were extracted using the Plasmid
370 Mega Kit (D6926; Omega) and quantified with a NanoPhotometer (Westburg; Leusden,
371 Netherlands).

372 Illumina HiSeq-compatible primers were used for sequencing. For each biological
373 replicate, the index primers were used in independent final PCR reactions. Each index
374 PCR was performed in 15 replicate PCR reactions (Supplemental Table S8). The PCR
375 products were size-selected (450–500 bp) using a 1% agarose gel (D4008; ZYMO) and
376 submitted for Illumina sequencing.

377

378 **Construction and infiltration of *Agrobacterium* input libraries**

379 The plasmid input libraries were introduced into *Agrobacterium* tumefaciens strain
380 GV3101 carrying the helper plasmid pSoup via electroporation. The transformed
381 *Agrobacterium* culture was grown in 150 mL of YEP medium (1% [w/v] yeast extract,
382 2% [w/v] peptone) until the OD reached 1.0. The bacteria were harvested and
383 resuspended in infiltration buffer (10 mM MES, pH 5.2; 10 mM MgCl2; 0.5 M MES,

384 pH 5.6; and 100 mM AS) to an OD of 1. *N. benthamiana* was grown in soil at 25°C in
385 a long-day photoperiod (16 h light and 8 h dark; cool-white fluorescent lights [Philips
386 TL-D 58W/840]; intensity 300 µmol/m2/s). The suspension was then infiltrated into the
387 two oldest fully expanded leaves of 20 vigorously growing *N. benthamiana* plants
388 (approximately one month old, pre-flowering). The plants were maintained under
389 normal growth conditions for 72 hours.

390 Illumina HiSeq-compatible primers were used for sequencing. For each biological
391 replicate, the index primers were used in independent final PCR reactions. Each index
392 PCR was performed in 15 replicate PCR reactions (Supplemental Table S8). The PCR
393 products were size-selected (450–500 bp) using a 1% agarose gel (D4008; ZYMO) and
394 submitted for Illumina sequencing.

395

396 **Preparation of Agro-STARR-seq output Libraries**

397 A total of 2000 µg of total RNA was extracted from the transformed *N.*
398 *benthamiana* leaves using the Plant RNAexr Kit (02160037; HYY). Polyadenylated
399 (poly A) mRNA was then isolated using the Dynabeads® Oligo(dT) kit (61005;
400 ThermoFisher Scientific). The mRNA was treated with TURBO™ DNase (AM2239;
401 ThermoFisher Scientific) and purified using Agencourt RNAClean XP beads (A66514;
402 Beckman Coulter).

403 First-strand cDNA was synthesized using SuperScript III (18080093;
404 ThermoFisher Scientific) with 1.5 µg mRNA per reaction. Target-specific reverse
405 transcription (RT; Supplemental Table S8) was performed using Agro-STARR-seq
406 mRNA-specific RT primers. All mRNA samples were used for the RT reaction. After
407 RT, the reaction was treated with RNase A and H, followed by pooling. In the final PCR
408 amplification, 5 µL of cDNA reaction was used as a template for a 50 µL PCR reaction
409 with the KAPA HiFi HotStart ReadyMix PCR Kit (KK2602; Roche). The PCR cycling
410 conditions were: 98°C for 30 s; 20 cycles of 98°C for 45 s; 15 cycles of 98°C for 15 s,
411 65°C for 30 s, and 72°C for 70 s.

412 Illumina HiSeq-compatible primers were used for sequencing. For each biological
413 replicate, 24 different index primers were used in independent final PCR reactions.
414 Each index PCR was performed in 15 replicate PCR reactions (Supplemental Table S8).
415 The PCR products were size-selected (450–500 bp) using a 1% agarose gel (D4008;
416 ZYMO) and submitted for Illumina sequencing.

417

418 **Luciferase Reporter Assay**

419 For the luciferase reporter assay, the vector backbone used for the reporter and
420 control plasmids was identical to that of the plasmid input library but lacked the CmR-
421 ccdB cassette (Fig. S8A and S8B). The candidate regions for the reporter plasmids were
422 derived from *Arabidopsis* genomic DNA. A dual-luciferase assay was performed in *N.*
423 *benthamiana* with three independent biological replicates. The dual-luciferase reporter
424 plasmid was introduced into *Agrobacterium* strain GV3101 via chemical
425 transformation for transient expression in *N. benthamiana* leaves. After 72 hours, the
426 dual-luciferase activity was measured following the manufacturer's protocol (E1910;
427 Promega).

428

429 **RNA isolation, RT-PCR, and qRT-PCR**

430 RNA was extracted from the *N. benthamiana* leaves transformed with
431 *Agrobacterium*. To identify EGFP genes, the suspension was then infiltrated into the
432 two oldest fully expanded leaves of 20 vigorously growing *N. benthamiana* plants
433 (approximately one month old, pre-flowering). The plants were maintained under
434 normal growth conditions for 72 hours. and subjected to RNA extraction with RNAiso
435 Plus (9109; Takara). First-strand cDNA was generated via reverse transcription reaction
436 with a HiScript ®II 1st Strand cDNA Synthesis kit (R312-01; Vazyme). The transcript
437 levels of LORELEI family genes, ROPs and NADPH oxidases were measured by RT-
438 PCR with the primer pairs listed (Supplemental Table S8). The abundance of RALF
439 transcripts was detected using qRT-PCR with ChamQ Universal SYBR qPCR Master

440 Mix (Q711-02; Vazyme). The following program was used: 95℃ for 5 min, followed
441 by 30 cycles at 95℃ for 10 s, and 60℃ for 30 s. The relative transcript abundance was
442 calculated using the 2–ΔΔCT method. The numbers were presented as means ± SD
443 from three replicates.
444

445 **Bioinformatic Analysis of Agro-STARR-seq in *Arabidopsis***
446 **Agro-STARR-seq Data Analysis**

447 Agro-STARR-seq involves two datasets with different types of readouts. The input
448 libraries were generated by directly amplifying the inserted fragments from the plasmid
449 DNA used for *Agrobacterium* electroporation, representing the original composition of
450 the starting *Agrobacterium* plasmid library. The output libraries measure the abundance
451 of self-transcribed mRNA from the inserted fragments in the transfected *Agrobacterium*
452 plasmid library.
453

454 **Enhancer peak calling and Distribution**

455 Fastp (v0.23.4) was used to filter low-quality reads from raw sequencing data with
456 the parameter "-w 5", while all other parameters remained at their default settings. The
457 filtered reads were then aligned to the TAIR10 reference genome using Bowtie2 (v2.2.8)
458 with the following settings: "--wrapper basic-0 --very-sensitive -X 2000 --rg-id <index
459 id>". Duplicate reads from different index libraries were removed separately using
460 Sambamba. Enhancer peak calling was performed with MACS2 (v2.1.4) using the
461 command: "callpeak -c <STARR-seq input library bam> -t <STARR-seq output library
462 bam> --keep-dup all -q 0.05 --gsize 119667750 –nomodel". R package "ideogram" was
463 used to plot the genome-wide distribution of enhancers.
464

465 **Enhancer annotation and conservation analysis**

466 The distribution of enhancers was analyzed using the function 'annotatePeak' from
467 R package ChIPSeeker [24] (v1.42.1) with genomic annotation provided by R package

468 'TxDb.Athaliana.BioMart.plantsmart51'.

469 PhastCons conservation scores were obtained from PlantRegMap [25], and
470 genomic region annotations were downloaded from Ensembl Plants. The average
471 PhastCons score was calculated for each enhancer and genomic region for comparative
472 analysis.

473

474 **Epigenetic states analysis**

475 Epigenomic datasets for *Arabidopsis thaliana* were obtained from published
476 studies in the GEO database: GSE183957 (mature leaves, Supplemental Table S9) [26]
477 and GSE233525 (young shoots, Supplemental Table S9) [16]. These datasets include
478 H3K27ac, H3K9ac, H3K27me3, H3K4me1, H3K4me3, and open chromatin. Data
479 processing, including trimming, mapping, and filtering, was performed for subsequent
480 analysis.

481

482 **Motif analysis**

483 A total of 619 transcription factors (TFs) of *Arabidopsis thaliana* were
484 downloaded from PlantTFDB v5.0 [25]. Motif enrichment analysis was performed
485 using a custom script. The background for motif enrichment analysis was set as shuffled
486 regions of non-enhancer regions. ERF115 was selected as an TF example and visualized
487 in IGV (Fig. S4E, S4F and S4G), based on data obtained from the published study
488 GSE48793 (Supplemental Table S9) [27].

489

490 **Transposable elements**

491 Annotated transposable elements (TEs) were downloaded from the *Arabidopsis*
492 Information Resource (TAIR, TAIR10 genome release).

493

494 **Deep learning model training and enhancer design (DREAM framework)**

Dataset Composition: To enhance data diversity while ensuring accurate quantification of enhancer activity, we randomly selected background genomic regions equal in number to the peak regions, with input DNA read counts > 30. To capture comprehensive enhancer sequence information, we extended 300 bp in both directions from the midpoint of each enhancer peak. Enhancer activity was quantified as the log2 fold change of RNA reads RPKM mapped to the genomic region over the input DNA reads RPKM. The dataset was further augmented by adding the reverse complement of each sequence with the same regulatory activity. Genomic regions on chromosomes 4 and 5 were used as validation and test datasets respectively, while the remaining regions served as training data. Model training and enhancer design: We employed our previously published DREAM framework for training, with a learning rate of 4e-5 and batch size of 256, while other parameters remained consistent with our prior research [24]. In the genetic algorithm, the initial population size, mutation rate, and other parameters were maintained as in our previous work [14].

Bioinformatic Pipeline and Analysis of ATAC-Agro-STARR-seq in *N. benthamiana*

We generally followed the ATAC-STARR-seq pipeline as described by Hansen TJ. et al. 2022 [17].

Read processing

Reads from ATAC-seq, plasmid, and Agrobacterium-mediated input (DNA) and output (RNA) libraries were processed using fastq_processing.py. The reference genome and gene annotation files were based on Niben261 (downloaded from http://solgenomics.net/), with chloroplast (C_AA066594.1) and mitochondrial (C_AA066595.1) sequences obtained from CNCB. For enhancer identification, duplicate reads were retained in the *Agrobacterium*-mediated DNA and RNA libraries, as suggested in the first ATAC-STARR-Seq paper [17]. The summary statistics are provided in Supplemental Table 3. These duplicate-retained reads were also used to estimate library complexity with the lc_extrap function from Preseq

524 (https://github.com/smithlabcode/preseq, version 3.2.0) and to analyze insert size
525 distribution with the CollectInsertSizeMetrics tool from Picard
526 (https://broadinstitute.github.io/picard/, version 2.26.3).

527

528 **Accessibility Analysis**

529 Genrich (https://github.com/jsh58/Genrich, version 0.6.1) was used to call
530 accessibility peaks for ATAC-seq. The gene distribution of ATAC-seq peaks and
531 enhancer regions as well as active histone modification peaks (described below) was
532 analyzed using the ChIPSeeker package. The promoter region ("tssRegion" parameter)
533 was defined as 2000 bp upstream and 1000 bp downstream of the TSS. The *Niben261*
534 TxDb object was constructed using the "makeTxDbFromGFF" function from the
535 GenomicFeatures R package [28].

536 For TSS enrichment analysis, gene locations and strand orientations were
537 extracted from the GFF annotation file. The "computeMatrix reference-point" and
538 "plotProfile" functions from deepTools (version 3.5.6) were then used to generate TSS
539 enrichment profiles.

540 To compare the expression levels of genes associated with DNA-accessible
541 proximal and distal regions, we downloaded three replicate *N. benthamiana* LAB leaf
542 RNA-seq datasets (SRR25703579, SRR25703580, and SRR25703581) (Supplemental
543 Table S9) [29]. Low-quality reads were filtered using Trim Galore
544 (https://github.com/FelixKrueger/TrimGalore), and the remaining reads were aligned
545 to the reference genome with HISAT2 (version 2.2.1) [30]. Reads with a mapping
546 quality score (MAPQ) $\geq$ 30 were retained for expression quantification. Gene-level read
547 counts were obtained using HTSeq-count (version 0.11.3) [31], and expression values
548 were normalized as FPKM using the fpkm function in edgeR [32]. The mean FPKM
549 across the three replicates was calculated for each gene.

550

551 **Enhancer Identification and Characterization**

552 Duplicate-retained DNA and RNA libraries were used as input for the call_ATAC-
553 STARR_regulatory-regions.py script to identify active bins and active regions, with
554 ATAC-seq peaks provided as the input peak set. Active regions were classified as
555 proximal or distal using the "annotatePeak" function from the ChIPSeeker package.
556 Active regions located within promoter regions were considered proximal active
557 regions, while those outside this region were considered distal active regions.

558 For histone modification analysis in active regions, we downloaded *Nicotiana*
559 *benthamiana* histone modification ChIP-seq data [29], including SRR27031034
560 (H3K27ac), SRR27031032 (H3K4me3), SRR27031033 (H3K27me3), SRR27031030
561 (H3K9me2), and their input SRR27031035, SRR27031031 (Supplemental Table S9).
562 And we used the "computeMatrix reference-point" and "plotHeatmap" functions from
563 deepTools to generate heatmaps for proximal and distal regulatory regions.

564 To compute transcription factor (TF) footprints within accessible peaks, we
565 applied the TOBIAS software to the DNA library. The analysis followed a similar
566 approach to Hansen et al. [17], except that *Arabidopsis thaliana* TF binding motifs were
567 used, obtained from PlantTFDB. Bound and unbound transcription factors (TFs) were
568 defined using the BINDetect step. TCP ChIP-seq datasets were obtained from
569 SRR15036661 and SRR15036676.

570 For motif enrichment analysis, TF motifs from *Arabidopsis thaliana* were first
571 scanned in all active regions, proximal active regions, distal active regions, and
572 background regions using FIMO [33]. Background sequences consisted of 50,000
573 randomly selected regions from the reference genome, excluding active regions, with
574 an average length of 115 bp, corresponding to the mean active region length. Then, the
575 number of regions containing each motif was counted for each type of region. Finally,
576 we used a chi-square test to identify motifs that were significantly enriched compared
577 to the background.

578 To analyze the functions of genes associated with TCP-bound active accessible
579 peaks, we extracted gene sets annotated to the bound peaks for each TCP transcription
580 factor. We performed GO enrichment analysis for these gene sets using clusterProfiler

581 [34]. GO term annotations were obtained from the eggNOG-mapper website [35]
582 (http://eggnog-mapper.embl.de/).

583 Because no CCA1-like gene is annotated in Niben261, we extracted RNA-seq
584 reads from the region surrounding Niben261Chr12:120,946,470–120,948,272 and used
585 StringTie [36] (version v3.0.0) to predict transcripts. The resulting exon sequences were
586 then aligned to *Arabidopsis* Araport11 coding sequences using blastn in website:
587 (https://www.arabidopsis.org/) to confirm its homologous gene.

588

Acknowledgements

We would like to thank Yang Fu, Hu Yang, Yantao Wang, Renyu Deng, Sheng Yu and Zhenghang Zhu for their technical assistance.

Author's contributions

Y.L. conceived the experiments. Y.B. conducted the experiments. H.L., J.Q., H.C., N.L., Y.C., L.X., Y.Z., F.Z and C.C. provided the technical support. Z.L. and X.Z. led the bioinformatics analysis; Z.L., X.Z., Y.S., and W.Z. performed the bioinformatics analysis; S.Z. supervised the work; Y.B., X.Z., Z.L. and Y.L. wrote the manuscript, and all authors approved the final manuscript.
Y.B., X.Z. and Y.S. should be considered co-first authors.

Data availability

Raw sequencing data will be made available upon publication.

Declarations

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare no conflict of interest with respect to the content of this manuscript. Methods for screening active regulatory elements in plants based on STARR-seq and active regulatory elements has filed a provisional patent application (202411137906.8) related to the work described here. The enhancer design and optimization framework DREAM has filed a provisional patent application (2023114102122) related to the work described here.

Reference

1. Schmitz RJ, Grotewold E, Stam M. Cis-regulatory sequences in plants: Their importance, discovery, and future challenges. Plant Cell. 2021;34:718–41. https://doi.org/10.1093/plcell/koab281

2. Lim F, Solvason JJ, Ryan GE, Le SH, Jindal GA, Steffen P, et al. Affinity-optimizing enhancer variants disrupt development. Nature. 2024;626:151–9. https://doi.org/10.1038/s41586-023-06922-8

3. Li T, Zeng W, Zhu F, Lü P. Cis-regulatory elements: systematic identification and horticultural applications. aBiotech. 2025;https://doi.org/10.1007/s42994-025-00237-0. https://doi.org/10.1007/s42994-025-00237-0

4. Wen C, Yuan Z, Zhang X, Chen H, Luo L, Li W, et al. Sea-ATI unravels novel vocabularies of plant active cistrome. Nucleic Acids Research. 2023;51:11568–83. https://doi.org/10.1093/nar/gkad853

5. Voichek Y, Hristova G, Mollá-Morales A, Weigel D, Nordborg M. Widespread position-dependent transcriptional regulatory sequences in plants. Nat Genet. Nature Publishing Group; 2024;56:2238–46. https://doi.org/10.1038/s41588-024-01907-3

6. Arnold CD, Gerlach D, Stelzer C, Boryń ŁM, Rath M, Stark A. Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science. 2013;339:1074–7. https://doi.org/10.1126/science.1232542

7. Liu Y, Yu S, Dhiman VK, Brunetti T, Eckart H, White KP. Functional assessment of human enhancer activities using whole-genome STARR-sequencing. Genome Biol. 2017;18:219. https://doi.org/10.1186/s13059-017-1345-5

8. Tan Y, Yan X, Sun J, Wan J, Li X, Huang Y, et al. Genome-wide enhancer identification by massively parallel reporter assay in Arabidopsis. Plant J. 2023;116:234–50. https://doi.org/10.1111/tpj.16373

9. Sun J, He N, Niu L, Huang Y, Shen W, Zhang Y, et al. Global Quantitative Mapping of Enhancers in Rice by STARR-seq. Genomics Proteomics Bioinformatics. 2019;17:140–53. https://doi.org/10.1016/j.gpb.2018.11.003

10. Tian W, Huang X, Ouyang X. Genome-wide prediction of activating regulatory elements in rice by combining STARR-seq with FACS. Plant Biotechnol J. 2022;20:2284–97. https://doi.org/10.1111/pbi.13907

11. Ricci WA, Lu Z, Ji L, Marand AP, Ethridge CL, Murphy NG, et al. Widespread long-range cis-regulatory elements in the maize genome. Nat Plants. 2019;5:1237–49. https://doi.org/10.1038/s41477-019-0547-0

12. Marand AP, Jiang L, Gomez-Cano F, Minow MAA, Zhang X, Mendieta JP, et al. The genetic architecture of cell type-specific cis regulation in maize. Science. 2025;388:eads6601. https://doi.org/10.1126/science.ads6601

13. Zhou W, Shi H, Wang Z, Huang Y, Ni L, Chen X, et al. Identification of Highly Repetitive Enhancers with Long-range Regulation Potential in Barley via STARR-seq. Genomics Proteomics Bioinformatics. 2024;22:qzae012. https://doi.org/10.1093/gpbjnl/qzae012

14. Li Z, Zhang Y, Peng B, Qin S, Zhang Q, Chen Y, et al. A novel interpretable deep learning-based computational framework designed synthetic enhancers with broad cross-species activity. Nucleic Acids Res. 2024;52:13447–68. https://doi.org/10.1093/nar/gkae912

15. Zhu X, Huang L, Wang C, Li G, Deng B, Kong D, et al. Uncovering the whole genome silencers of human cells via Ss-STARR-seq. Nat Commun. 2025;16:723. https://doi.org/10.1038/s41467-025-55852-8

16. Deng L, Zhou Q, Zhou J, Zhang Q, Jia Z, Zhu G, et al. 3D organization of regulatory elements for transcriptional regulation in Arabidopsis. Genome Biol. 2023;24:181. https://doi.org/10.1186/s13059-023-03018-4

17. Hansen TJ, Hodges E. ATAC-STARR-seq reveals transcription factor–bound activators and silencers within chromatin-accessible regions of the human genome. Genome Res. 2022;32:1529–41. https://doi.org/10.1101/gr.276766.122

18. EGDB: A comprehensive multi-omics database for energy grasses and the epigenomic atlas of pearl millet - Luo - 2025 - iMeta - Wiley Online Library [Internet]. [cited 2025 Mar 21]. https://onlinelibrary.wiley.com/doi/10.1002/imt2.263. Accessed 21 Mar 2025

19. Mannion BJ, Tran S, Plajzer-Frick I, Novak CS, Afzal V, Akiyama JA, et al. Uncovering hidden enhancers through unbiased in vivo testing. Nat Commun. 2025;16:7313. https://doi.org/10.1038/s41467-025-62497-0

20. Alabadí D, Oyama T, Yanovsky MJ, Harmon FG, Más P, Kay SA. Reciprocal regulation between TOC1 and LHY/CCA1 within the Arabidopsis circadian clock. Science. 2001;293:880–3. https://doi.org/10.1126/science.1061320

21. Wu J-F, Tsai H-L, Joanito I, Wu Y-C, Chang C-W, Li Y-H, et al. LWD-TCP complex activates the morning gene CCA1 in Arabidopsis. Nat Commun. 2016;7:13181. https://doi.org/10.1038/ncomms13181

22. Yon F, Joo Y, Cortés Llorca L, Rothe E, Baldwin IT, Kim S-G. Silencing Nicotiana attenuata LHY and ZTL alters circadian rhythms in flowers. New Phytol. 2016;209:1058–66. https://doi.org/10.1111/nph.13681

682 23. Jores T, Tonnies J, Dorrity MW, Cuperus JT, Fields S, Queitsch C. Identification
683 of Plant Enhancers and Their Constituent Elements by STARR-seq in Tobacco Leaves.
684 Plant Cell. 2020;32:2120–31. https://doi.org/10.1105/tpc.20.00155

685 24. Yu G, Wang L-G, He Q-Y. ChIPseeker: an R/Bioconductor package for ChIP peak
686 annotation, comparison and visualization. Bioinformatics. 2015;31:2382–3.
687 https://doi.org/10.1093/bioinformatics/btv145

688 25. Tian F, Yang D-C, Meng Y-Q, Jin J, Gao G. PlantRegMap: charting functional
689 regulatory maps in plants. Nucleic Acids Res. 2020;48:D1104–13.
690 https://doi.org/10.1093/nar/gkz1020

691 26. Zhao L, Zhou Q, He L, Deng L, Lozano-Duran R, Li G, et al. DNA methylation
692 underpins the epigenomic landscape regulating genome transcription in Arabidopsis.
693 Genome Biol. 2022;23:197. https://doi.org/10.1186/s13059-022-02768-x

694 27. Heyman J, Cools T, Vandenbussche F, Heyndrickx KS, Van Leene J, Vercauteren
695 I, et al. ERF115 controls root quiescent center cell division and stem cell replenishment.
696 Science. 2013;342:860–3. https://doi.org/10.1126/science.1240667

697 28. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al.
698 Software for computing and annotating genomic ranges. PLoS Comput Biol.
699 2013;9:e1003118. https://doi.org/10.1371/journal.pcbi.1003118

700 29. Ranawaka B, An J, Lorenc MT, Jung H, Sulli M, Aprea G, et al. A multi-omic
701 Nicotiana benthamiana resource for fundamental research and biotechnology. Nat
702 Plants. 2023;9:1558–71. https://doi.org/10.1038/s41477-023-01489-8

703 30. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment
704 and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37:907–15.
705 https://doi.org/10.1038/s41587-019-0201-4

706 31. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-
707 throughput sequencing data. Bioinformatics. 2015;31:166–9.
708 https://doi.org/10.1093/bioinformatics/btu638

709 32. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for
710 differential expression analysis of digital gene expression data. Bioinformatics.
711 2010;26:139–40. https://doi.org/10.1093/bioinformatics/btp616

712 33. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif.
713 Bioinformatics. 2011;27:1017–8. https://doi.org/10.1093/bioinformatics/btr064

714 34. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing
715 biological themes among gene clusters. OMICS. 2012;16:284–7.
716 https://doi.org/10.1089/omi.2011.0118

35. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Mol Biol Evol. 2021;38:5825–9. https://doi.org/10.1093/molbev/msab293

36. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33:290–5. https://doi.org/10.1038/nbt.3122

725 # Figure legend

726 **Figure 1. *In planta* Identification and Engineering of *Arabidopsis***
727 **Enhancers Using Agro-STARR-seq and DREAM Framework.**

728 (A) Schematic overview of Agro-STARR-seq workflow and DREAM
729 framework. Agro-STARR-seq generates high quality *in planta* enhancer
730 data which is subsequently used as training data by DREAM to design
731 synthetic enhancers. Agro-STARR-seq input libraries include the plasmid
732 input library and *Agrobacterium* input library. The plasmid input library:
733 Genomic DNA is fragmented and cloned into a modified Agro-STARR-
734 seq vector containing a 35S mini promoter driving basal EGFP expression.
735 The *Agrobacterium* input library: The plasmid input library is transformed
736 into *Agrobacterium* by electroporation. Agro-STARR-seq output library:
737 The *Agrobacterium* input library is infiltrated into *N. benthamiana* leaves.
738 After 72 hours, RNA is isolated for next-generation sequencing. Enhancer
739 activity is quantified by comparing sequence enrichment in the output
740 library to the input library. DREAM comprises two integral modules: the
741 state-of-the-art SENet that predicts enhancer activity using DNA
742 sequences as the input and the evolutionary optimization module that
743 designs strong synthetic enhancers by employing a genetic algorithm,
744 which iteratively maximizes regulatory activity as predicted by the SENet-
745 derived model.

746 (B) Representative genomic regions show the signal profiles from both
747 output and input libraries, alongside ATAC-seq and H3K27ac tracks. The
748 first genome browser track displays fragment densities from the Agro-
749 STARR-seq output library (red), the plasmid input library (blue), and the
750 *Agrobacterium* input library (green).

751 (C) Venn diagram depicting overlap between Agro-STARR-seq enhancers
752 and ATAC-seq peaks [16].

753 (D) Annotation of regulatory regions relative to the transcriptional start site
754 (TSS) (UTR: untranslated region; Promoter-proximal: ±500 bp TSS-
755 proximal regions; Distal intergenic: all regions outside the promoter-
756 proximal, gene body, and 300 bp downstream regions).

757 (E) Heatmaps show signal enrichment (±2 kb from enhancer centers) for
758 STARR-seq, H3K27ac, H3K9ac, H3K27me3, H3K4me1, and H3K4me3
759 marks, categorized by TSS proximity, with the figure highlighting TSS
760 distal regions. Signal intensity represents fold enrichment of ChIP-seq over
761 input (normalized by library size, detailed in Supplemental notes).

762 (F) TFs motifs enriched in the Agro-STARR-Seq enhancers compared to
763 genomic regions of non-enhancers (background). The x -axis represents the
764 $-\log_{10}$p-value, and the y -axis represents the Enrichment score.

765 (G) Enrichment profile of enhancers overlapping with TEs.

766 (H) Average PhastCons scores of Agro-STARR-seq enhancers and non-
767 enhancers in different genomic categories.

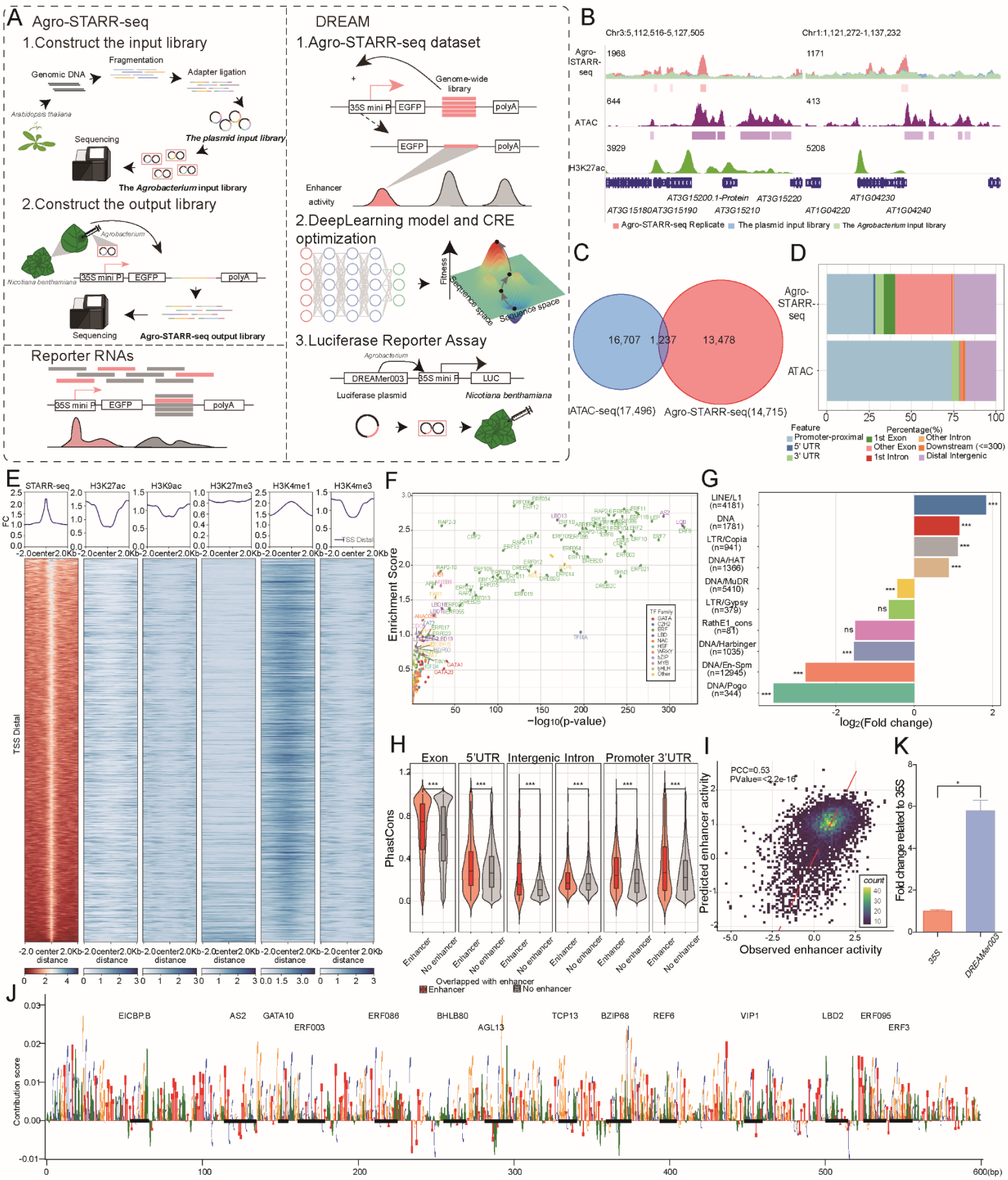
768 (I) Performance evaluation of DREAM enhancer activity prediction
769 module using hold-out chromosome dataset from *A. thaliana*.

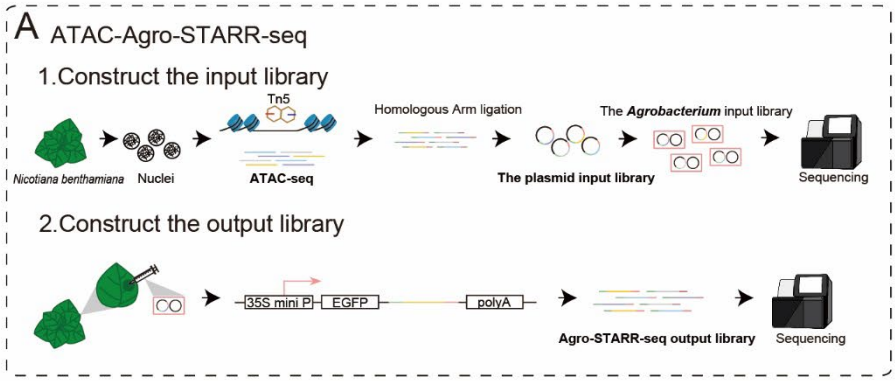
770 (J) Nucleotide contribution scores for the optimized enhancers derived
771 from the enhancer activity models using DeepExplainer. Instances of
772 motifs identified by DREAM are emphasized, with known motifs indicated
773 in black.

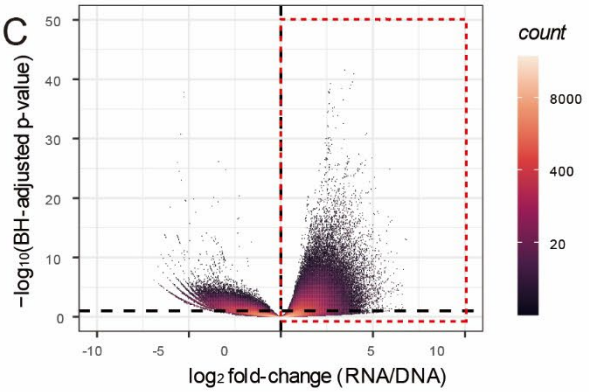
774 (K) Luciferase reporter assay validation of a DREAM-optimized synthetic
775 enhancer. Values normalized to Renilla luciferase signal. Error bars:
776 Standard error of the mean (n = 3 biological replicates). This synthetic
777 enhancer was named DREAMer003, in line with the strong enhancer
778 DREAMer001 and silencer DREAMer002 designed using Drosophila
779 STARR-seq data in our previous study [14].

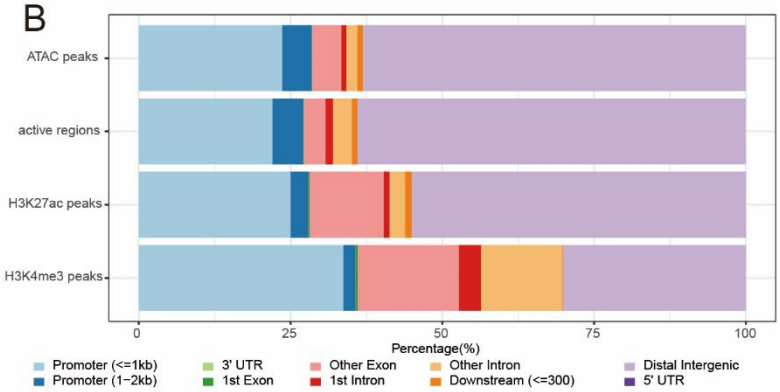
780 **Figure 2. *In planta* enhancer screening and characterization of *N.***
781 ***benthamiana* leaf open chromatin regions using ATAC-Agro-STARR-**
782 **seq.**
783 (A) Schematic overview of ATAC-Agro-STARR-seq workflow. ATAC-
784 Agro-STARR-seq input libraries include the plasmid input library and
785 *Agrobacterium* input library. The plasmid input library: Open chromatin is
786 isolated from *N. benthamiana* leaf nuclei with the cut-and-paste
787 transposase Tn5. The open chromatin fragments are cloned into the Agro-
788 STARR-seq plasmid. The *Agrobacterium* input library: The plasmid input
789 library is transformed into *Agrobacterium* by electroporation. Agro-
790 STARR-seq output library: The *Agrobacterium* input library is infiltrated
791 into *N. benthamiana* leaves. After 72 hours, RNA is isolated for next-
792 generation sequencing. Enhancer activity is quantified by comparing
793 sequence enrichment in the output library to the input library.
794 (B) Annotation of open chromatin peaks, their active regions (Agro-
795 STARR-seq enhancer in open chromatin regions), and H3K27ac and
796 H3K4me3 peaks relative to the transcription start site (TSS). Promoters
797 were defined as regions extending 2 kb upstream and 1 kb downstream of
798 the TSS.
799 (C) A scatter plot showing the $\log_2$ fold change ($\log_2$FC) of normalized
800 output to input coverage, and the $-\log_{10}$ (BH-adjusted p-value) for all bins,
801 as identified by DESeq2 analysis. Red boxes indicate active bins.

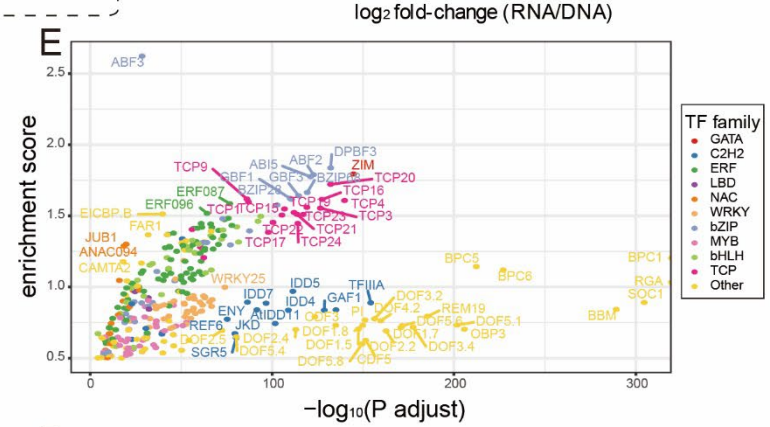
802 (D) Heatmaps of histone modification ChIP-seq signals in ATAC-Agro-
803 STARR-seq-defined active regulatory elements (proximal and distal). The
804 columns (left to right) are as follows: 1-2) H3K27ac and H3K4me3 signals;
805 3) the corresponding input control for the active marks; 4) H3K9me2 signal,
806 a heterochromatin mark shown for comparison; 5) input control for
807 H3K9me2.
808 (E) Enrichment analysis of transcription factor motifs in ATAC-Agro-
809 STARR-Seq enhancer regions.
810 (F) Comparison of ATAC-STARR-seq cut count signal, ATAC-seq cut
811 count signal, TCP ChIP-seq signal, and enhancer activity between
812 predicted motif-bound and predicted motif-unbound regions of TCP20.
813 G) GO enrichment analysis of nearest-neighbor gene sets associated with
814 TCP23 bound active accessible peaks (Supplemental Table S10).
815 H) Chromatin accessibility, TCP binding sites, TCP ChIP signals, active
816 regions, and RNA-seq profiles at the CCA1-like gene.
817

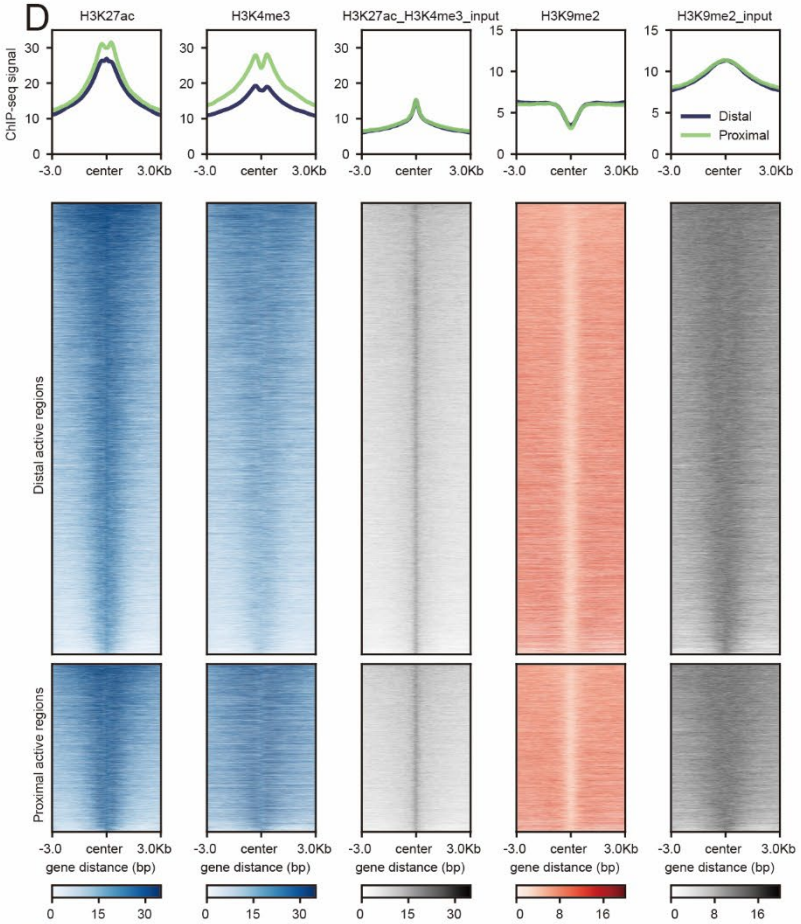

A
Agro-STARR-seq
1.Construct the input library
Genomic DNA
Fragmentation
Adapter ligation
Arabidopsis thaliana
Sequencing
The plasmid input library
The Agrobacterium input library
2.Construct the output library
Agrobacterium
Nicotiana benthamiana
35S mini P
EGFP
polyA
Sequencing
Agro-STARR-seq output library
Reporter RNAs
DREAM
1.Agro-STARR-seq dataset
Genome-wide library
Enhancer activity
2.DeepLearning model and CRE optimization
Fitness
Sequence space
3.Luciferase Reporter Assay
DREAMer003
LUC
Luciferase plasmid
B
Chr3:5,112,516-5,127,505
Chr1:1,121,272-1,137,232
Agro-STARR-seq
ATAC
H3K27ac
AT3G15200.1-Protein
AT3G15220
AT1G04230
AT3G15180
AT3G15190
AT3G15210
AT1G04220
AT1G04240
Agro-STARR-seq Replicate
The plasmid input library
The Agrobacterium input library
C
16,707
1,237
13,478
ATAC-seq(17,496)
Agro-STARR-seq(14,715)
D
Agro-STARR-seq
ATAC
Feature
Percentage(%)
Promoter-proximal
5' UTR
3' UTR
1st Exon
Other Exon
1st Intron
Other Intron
Downstream (<=300)
Distal Intergenic
E
STARR-seq
H3K27ac
H3K9ac
H3K27me3
H3K4me1
H3K4me3
FC
TSS Distal
distance
F
Enrichment Score
−log10(p-value)
TF Family
G
LINE/L1 (n=4181)
DNA (n=1781)
LTR/Copia (n=941)
DNA/HAT (n=1366)
DNA/MuDR (n=5410)
LTR/Gypsy (n=379)
RathE1_cons (n=81)
DNA/Harbinger (n=1035)
DNA/En-Spm (n=12945)
DNA/Pogo (n=344)
log2(Fold change)
H
Exon
5'UTR
Intergenic
Intron
Promoter
3'UTR
PhastCons
Enhancer
No enhancer
Overlapped with enhancer
I
PCC=0.53
PValue=<2.2e-16
Predicted enhancer activity
Observed enhancer activity
count
K
Fold change related to 35S
35S
DREAMer003
J
Contribution score
EICBP.B
AS2
GATA10
ERF003
ERF086
BHLB80
AGL13
TCP13
BZIP68
REF6
VIP1
LBD2
ERF095
ERF3
600(bp)

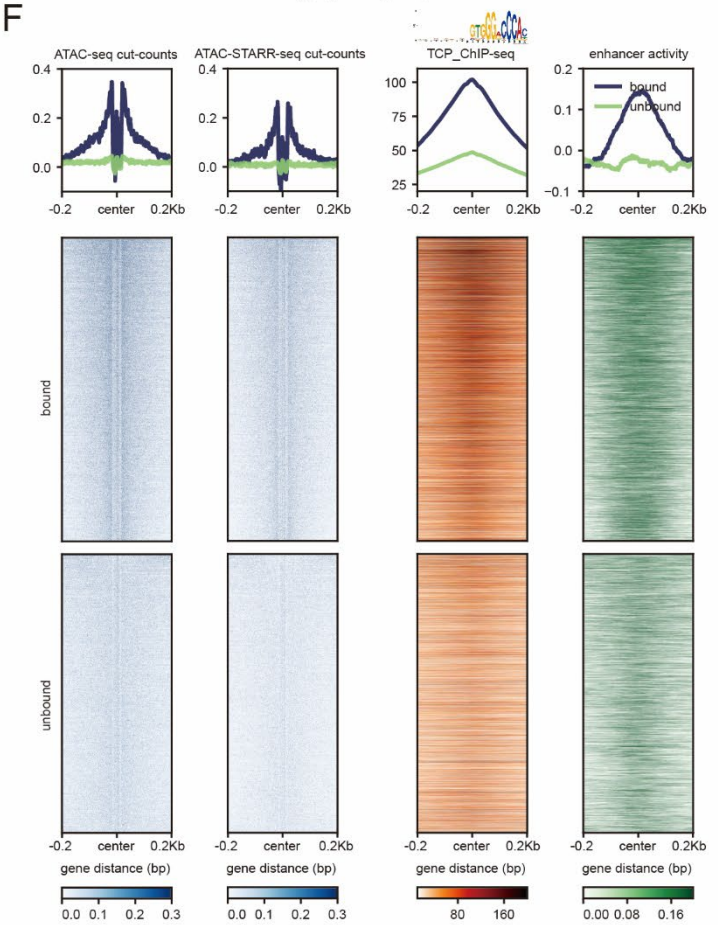

A
ATAC-Agro-STARR-seq
1.Construct the input library
Tn5
Homologous Arm ligation
The Agrobacterium input library
Nicotiana benthamiana
Nuclei
ATAC-seq
The plasmid input library
Sequencing
2.Construct the output library
35S mini P
EGFP
polyA
Agro-STARR-seq output library
Sequencing

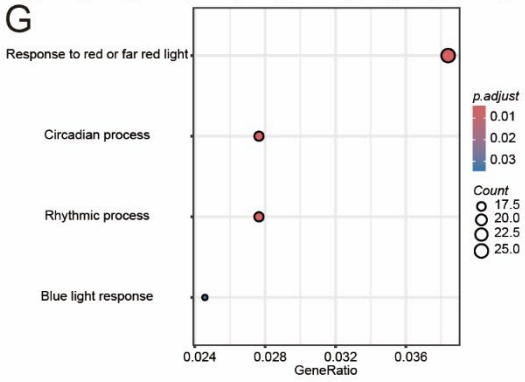

C
count
8000
400
20
−log10(BH-adjusted p-value)
log2 fold-change (RNA/DNA)

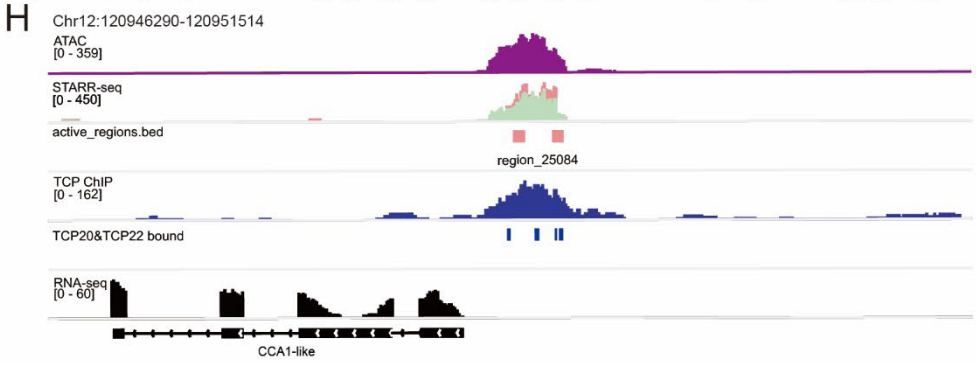

B
ATAC peaks
active regions
H3K27ac peaks
H3K4me3 peaks
Percentage(%)
Promoter (<=1kb)
Promoter (1−2kb)
3' UTR
1st Exon
Other Exon
1st Intron
Other Intron
Downstream (<=300)
Distal Intergenic
5' UTR


E
enrichment score
−log10(P adjust)
TF family
GATA
C2H2
ERF
LBD
NAC
WRKY
bZIP
MYB
bHLH
TCP
Other


D
H3K27ac
H3K4me3
H3K27ac_H3K4me3_input
H3K9me2
H3K9me2_input
ChIP-seq signal
Distal
Proximal
Distal active regions
Proximal active regions
gene distance (bp)


F
ATAC-seq cut-counts
ATAC-STARR-seq cut-counts
TCP_ChIP-seq
enhancer activity
bound
unbound
gene distance (bp)


G
Response to red or far red light
Circadian process
Rhythmic process
Blue light response
GeneRatio
p.adjust
Count


H
Chr12:120946290-120951514
ATAC
[0 - 359]
STARR-seq
[0 - 450]
active_regions.bed
region_25084
TCP ChIP
[0 - 162]
TCP20&TCP22 bound
RNA-seq
[0 - 60]
CCA1-like