Cryptic promoter activation occurs by at least two different mechanisms in the *Arabidopsis* genome

Hisayuki Kudo[1], Mitsuhiro Matsuo[2], Soichirou Satoh[2], Rei Hachisu[1], Masayuki Nakamura[1], Yoshiharu Y Yamamoto[3], Takayuki Hata[2], Hiroshi Kimura[4], Minami Matsui[5] and Junichi Obokata[2] *

[1] Center for G Research, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8602 Japan

[2] Graduate School of Life and Environmental Sciences, Kyoto Prefectural University, 1-5 Hangi-cho, Shimogamo, Sakyo-ku, Kyoto 606-8522, Japan

[3] Faculty of Applied Biological Sciences, Gifu University, 1-1 Yanagito, Gihu-shi, Gifu 501-1193, Japan

[4] Graduate School of Bioscience and Biotechnology, Tokyo Institute of Technology, Yokohama City, Kanagawa 226-8501, Japan

[5] RIKEN Plant Science Center, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan

Corresponding author* Junichi Obokata

Tel: +81-72-896-5402; Email: junichi.obokata@setsunan.ac.jp

"The authors wish it to be known that, in their opinion, the first 3 authors should be regarded as joint First Authors".

Present Address: [Junichi Obokata and Mitsuhiro Matsuo] Faculty of Agriculture, Setsunan University, 45-1, Nagao-Touge-cho, Hirakata, Osaka, 573-0101, Japan

Running Head: Cryptic promoter activation mechanisms

Keywords: cryptic promoters, chromatin remodeling, promoter trapping, transcription initiation, H3K4me3, H2A.Z, *Arabidopsis thaliana,* transcription start sites, endosymbiotic gene transfer

ABSTRACT

In gene-trap screening of plant genomes, promoterless reporter constructs are often expressed without trapping of annotated gene promoters. The molecular basis of this phenomenon, which has been interpreted as the trapping of cryptic promoters, is poorly understood. In this study, using *Arabidopsis* gene-trap lines in which a firefly luciferase (*LUC*) open reading frame (ORF) was expressed from intergenic regions, we found that cryptic promoter activation occurs by at least two different mechanisms: one is the capturing of pre-existing promoter-like chromatin marked by H3K4me3 and H2A.Z, and the other is the entirely new formation of promoter chromatin near the 5′ end of the inserted *LUC* ORF. To discriminate between these, we denoted the former mechanism as “cryptic promoter capturing”, and the latter one as “promoter *de novo* origination”. The latter finding raises a question as to how inserted *LUC* ORF sequence is involved in this phenomenon. To examine this, we performed a model experiment with chimeric *LUC* genes in transgenic plants. Using *Arabidopsis psaH1* promoter–*LUC* constructs, we found that the functional core promoter region, where transcription start sites (TSS) occur, cannot simply be determined by the upstream nor core promoter sequences; rather, its positioning proximal to the inserted *LUC* ORF sequence was more critical. This result suggests that the insertion of the *LUC* ORF sequence alters the local distribution of the TSS in the plant genome. The possible impact of the two types of cryptic promoter activation mechanisms on plant genome evolution and endosymbiotic gene transfer is discussed. (248 words)

INTRODUCITON

Gene-trap screening is a useful tool in functional genomics because it reveals the function and spatio-temporal expression profile of the genes captured by promoterless reporter constructs (Springer, 2000; Stanford *et al*., 2001). However, gene-trap screening of plant genomes often causes unexpected expression of the constructs inserted in the intergenic genomic regions or in the reverse orientation in the coding regions, without trapping the annotated genes (Fobert *et al*., 1994; Topping *et al*., 1994; Ökrész *et al*., 1998; Mollier *et al*., 2000; Plesch *et al*., 2000; Yamamoto *et al*., 2003; Sivanandan *et al*., 2005; Stangeland *et al*., 2005). This type of enigmatic expression has been interpreted as the trapping of cryptic promoters.

Although its molecular basis is poorly understood, a cryptic promoter is presumed to be a kind of promoter whose function is not detectable unless reporter constructs are inserted just downstream of it (Fobert *et al*., 1994; Topping *et al*., 1994; Ökrész *et al*., 1998; Mollier *et a*l., 2000; Plesch *et al*., 2000; Yamamoto *et al*., 2003; Sivanandan *et al*., 2005; Stangeland *et al*., 2005). Unidentified non-coding RNA (ncRNA) genes could be a source of cryptic promoters in gene-trap screening. An alternative possibility is the new occurrence of chromatin remodelling to form a chromatin structure that exhibits promoter function, just upstream of the inserted reporter genes. Although the origin and evolution of protein-coding sequences are well documented (Long *et al*., 2003; Kaessmann, 2010; Tautz and Domazet-Lošo, 2011; Carvunis *et al*., 2012), the mechanism via which newly emerging protein-coding sequences in the eukaryotic genome acquire transcriptional competence is less well understood.

We have been interested in the endosymbiotic evolutionary process of chloroplasts. During this process, most genes of the endosymbiotic organelle move to the host nucleus and become integrated into the nuclear gene network (Martin *et al*., 1998; Timmis *et al*., 2004; Matsuo *et al*., 2005). For this functional gene transfer, the translocated genes should acquire eukaryotic promoters at their genomic integration loci. Several cases of acquisition of eukaryotic promoters by organelle-derived coding sequences by trapping pre-existing nuclear genes have been reported (Kadowaki *et al*., 1996; Kubo *et al*., 1999; Stegemann and Bock, 2006; Wang *et al*., 2014). This promoter-acquisition mechanism is easy to understand, but one promoter-acquisition event will result in one disruption of pre-existing genes. Therefore an alternative mechanism might be necessary to explain how thousands of organelle-derived coding sequences have become transcriptionally active in the nucleus. To address this question, we became interested in the similarity of the promoter-acquisition event between gene-trap screening and organelle–nucleus functional gene transfer. In this respect, cryptic promoter activation is a thought-provoking phenomenon.

Recent nucleosome and transcription studies have revealed that core promoter regions in the eukaryotic genome have a specific chromatin structure: the transcriptional pre-initiation complex (PIC) and transcription start sites (TSS) occur in the nucleosome-free region (NFR), which is flanked by nucleosomes containing modified histones and histone variants (Guenther *et al*., 2007; Li *et al*., 2007; Cairns, 2009; Jiang and Pugh, 2009; Zhang *et al*., 2009; Deal and Henikoff, 2011; Haberle and Stark, 2018; Andersson and Sandelin, 2020), and the downstream coding region possesses a more closed chromatin structure that represses aberrant transcription initiation within the gene body (Hennig *et al*., 2012; Hennig and Fischer, 2013; Neri *et al*., 2017). In this study, we scrutinized the cryptic promoters found by gene-trap screening of the *Arabidopsis* genome (Yamamoto *et al*., 2003) and characterized them according to their chromatin-remodelling state. We found that “cryptic promoter activation” can be caused by at least two different mechanisms: one is the capturing of pre-existing promoter-like chromatin, with its inherent transcripts being hardly detectable, and the other is the entirely new formation of promoter chromatin near the 5′ end of the inserted *LUC* ORF. The latter case raises a question as to whether the inserted *LUC* ORF sequence takes part in the formation and/or localization process of the new core promoter region emerging near its 5′ end. To examine this, we performed a model promoter experiment in transgenic plants, indicating that the inserted *LUC* ORF sequence is involved in at least localization process of PIC and TSS to its 5’ proximal region. These findings for the transgenic *LUC* ORF provide new insights into a possible mechanism by which newly emerged protein-coding sequences in the plant genome acquire transcriptional competence.

RESULTS

Expression of the firefly luciferase (LUC) trap vector in intergenic regions

Gene-trap screening of the *A. thaliana* genome revealed that, depending on the vector design, 23% to 67% of *LUC* ORF-activation events did not depend on the capture of annotated gene promoters (Yamamoto *et al*., 2003). To understand how a promoterless *LUC* gene (Figure 1a) can become transcriptionally active after insertion into intergenic regions, we analysed 59 *LUC*-expressing intergenic insertion lines. To simplify the analysis, eight lines were prescreened using RNA–gel blot hybridization for the discernible *LUC* transcript. Rapid amplification of cDNA 5′ ends (5′ RACE) analysis of the eight selected lines revealed that their *LUC* transcripts had the TSS in the flanking genomic regions, and not within the T-DNA inserts. From these lines, we selected YB41 and YB84 for further detailed analysis (Figure 1b) because their *LUC* transcript 5′ UTRs contained *A. thaliana* genomic sequences that were sufficiently long to examine their transcript levels, before and after the T-DNA insertion, by reverse transcription PCR (RT–PCR). The exact TSS of YB41 and YB84 were identified using the biotinylated cap-trapper method (Carninci and Hayashizaki, 1999), which showed that they were distributed 30–170 bp upstream of the T-DNA insertion sites (Figure 1c), predominantly at pyrimidine (Py) and purine (Pu) junctions (Figure S1). In WT plants, we could not detect any TSS in the corresponding genomic regions.

We then used real-time RT–PCR to measure the transcript levels of these genomic loci before and after the T-DNA insertions normalized by the endogenous UBC (ubiquitin gene, AT5G25760) mRNA level as a reference (Czechowski *et al*., 2005) (Figure 1d). The RNA level of the YB41 locus in transgenic plants was only 1.5% of the endogenous UBC mRNA level, with a faint signal also detected in WT plants (Figure 1d). Because no TSS was detected upstream of the YB41 locus in WT plants, this faint signal might reflect the read-through products from the upstream neighbouring gene (Figure 1b). The RNA level of the YB84 locus in transgenic plants was as high as 13% of the UBC mRNA level, whereas no RT–PCR signal was detected from WT plants (Figure 1d). These results strongly suggest that the integration of the *LUC*-trap vector at the YB41 and YB84 genomic loci (Figure 1b) caused the new occurrence of transcripts.

Chromatin signature discrimination between "cryptic promoter capturing" and "promoter de novo origination"

We examined what occurred at the YB41 and YB84 integration sites at the chromatin level. To do this, we first prepared a custom-made DNA tiling array covering the –480 to +300 base regions of the YB41 and YB84 T-DNA insertion sites (Figure S2). We performed chromatin immunoprecipitation followed by microarray analysis (ChIP-on-chip) to examine the localization of nucleosomes containing a trimethylated form of histone H3 (H3K4me3) and a variant of histone H2A (H2A.Z), both of which are localized at core promoter regions in plants, yeast and animals (Guenther *et al*., 2007; Li *et al*., 2007; Cairns, 2009; Jiang and Pugh, 2009; Zhang *et al*., 2009; Deal and Henikoff, 2011; Weber and Henikoff, 2014; Hyun et al., 2017; Haberle and Stark, 2018; Giaimo et al., 2019; Andersson and Sandelin, 2020). We also analysed the binding site of the TATA-binding protein (TBP) as a representative component of the transcriptional pre-initiation complex (PIC).

Figure 2a shows the chromatin configuration of the YB41 locus before and after the T-DNA insertion. No signals for H2A.Z, H3K4me3, TBP or TSS were detected over this genomic locus in WT plants, in which this locus is covered totally by nucleosomes containing the canonical histone H4. After T-DNA insertion, all these signals appeared and exhibited a chromatin configuration that was characteristic of the pol II TSS, as depicted in Figure 2c. The TSS occurred in the nucleosome-free region (NFR) flanked by two nucleosomes, both containing H2A.Z and H3K4me3, and TBP overlapped the 5′-flanking nucleosome (namely, –1 nucleosome relative to the NFR). The 3′-flanking nucleosome (+1 nucleosome) was located around the 5′ end of the *LUC* ORF in the T-DNA region. Because the YB41 transgenic plants used here were not homozygous, a histone H4 distribution profile similar to that of WT plants was still found in the ChIP-on-chip profile (–450 to +0). Taken together, these results indicate that a chromatin configuration that was capable of transcription initiation was newly formed at the YB41 locus after the insertion of the *LUC*-trap vector.

The case of the YB84 locus was quite different. As shown in Figure 2b, a chromatin configuration similar to that shown in Figure 2c was already present in the WT plants. Therefore, the *LUC*-trap vector should have captured pre-existing promoter-like chromatin at this genomic locus. We did not detect any transcripts or TSS at this WT locus, which suggests that its inherent transcripts are hardly detectable because of, for example, poor stability, extremely low abundance, rapid processing and/or pausing of RNA polymerase. Even in this case, the +1 nucleosome of the transgenic plants was localized at the 5′ end of the *LUC* ORF, similarly to what was observed for YB41 (Figure 2a and 2b).

The analyses of YB41 and YB84 described above revealed that the so-called phenomenon of “cryptic promoter activation” was caused by at least two different mechanisms. To discriminate between these, we denoted phenomena such as the YB41 case as “promoter *de novo* origination” and those of YB84 as “cryptic promoter capturing”.

The inserted LUC ORF appears to be involved in the localization of TSS to its 5’ proximal region.

The finding of promoter *de novo* origination raises a question regarding the nature of its underlying mechanism. It is probable that its occurrence should depend on the individual genomic insertion sites, with as yet unidentified properties. However, it is also intriguing how the *LUC* ORF sequence is involved in this process; is it only a passively transcribed sequence, or does it have any influence on the origination and/or positioning of the TSS? To explore the latter possibility, we designed an experiment based on the following rationale (Figure 3a). In a typical protein-coding gene of plants, the core promoter region, where the PIC is formed and TSS occur, is located just upstream of the coding region and downstream of the regulatory promoter region. We attempted to separate these relationships by triplicating the core promoter segments according to three possible scenarios, as follows. (1) If the core promoter segment is an integral part of the whole promoter region, the 5′-most segment should be used preferentially (Figure 3a, i). (2) If the core promoter is an autonomous functional unit, the TSS should occur at each of the segments (Figure 3a, ii). (3) If the functional core promoter region occurs in association with the proximal coding region sequence, the 3′-most segment should be used preferentially (Figure 3a, iii).

To execute this experimental design with the *LUC* ORF, we triplicated the core promoter segment within the context of the promoter of the *A. thaliana psaH1* photosynthesis gene*.* In the endogenous *psaH1*, the TSS were distributed from –52 to –12 relative to the ATG initiation codon, with the peak TSS located at –51 (Figure 3b and Figure S3). The core promoter region of *psaH1* does not have a characteristic core promoter motif, but the TSS occur at the alteration sites of Py and Pu (Yamamoto *et al*., 2009) (Figure S3). First, we prepared a psaH1::LUC construct in which the *psaH1* gene region (–2000 to +12) was translationally fused with the *LUC* ORF (Figure 3b, details provided in Figure S4). In the derived construct HC$_{111}$x3::LUC, the 111 bp core promoter segment (–111 to –1) of psaH1::LUC was triplicated as direct repeats (A, B and C in Figure 3b and 3c, details given in Figure S5). The TSS selection at these artificial genes was identified using the biotinylated cap-trapper method with

RNA from transgenic plants (Figure 3b and Figures S3–S5). The TSS distribution profile of psaH1::LUC was similar to that of the WT *psaH1* gene, with a peak TSS located at –51. Surprisingly, the TSS of HC_{111}x3::LUC occurred preferentially in segment C (85%), with only a small fraction observed in segments A (4%) and B (11%), even though these three segments have identical DNA sequences (Figure 3b and Figure S5). The TSS distribution profile within segment C was very similar to that observed for the core promoter region of psaH1::LUC (Figure 3b).

To confirm that the TSS occurred preferentially in segment C, we performed a ChIP–PCR analysis to examine the localization of the PIC using antibodies against pol II and TBP. Because segments A, B and C have identical DNA sequences, we discriminated between them using PCR primer sets designed at their junctions (Figure 3c). Fine-tuning of the PCR conditions allowed us to amplify a single-unit-sized DNA fragment for each primer set (Figure 3d, see methodological details in Table S2). Under this condition, the ChIP signal was quantified by real-time PCR analysis and showed that pol II and TBP were localized preferentially in segment C (Figure 3e), in accordance with the TSS distribution.

In this experiment, we did not determine the locations of nucleosomes on the chimeric promoters, because the 111 bp direct repeat sequence hinders nucleosome mapping by ChIP-on-chip or ChIP-seq analysis. However, the results obtained demonstrated that the functional core promoter region, where the PIC and TSS occur, cannot simply be determined by the sequences of its own or whole promoter region; rather, its positioning proximal to the *LUC* ORF is more critical, at least in this case.

DISCUSSION

This study revealed that the enigmatic phenomenon of "cryptic promoter activation" was caused by at least two different mechanisms. One was the capturing of the cryptic promoter, which was the pre-existing promoter-like chromatin whose inherent transcript was hardly detectable (YB84 in Figure 1), and the other was the entirely new formation of promoter chromatin, which we denoted as promoter *de novo* origination (YB41 in Figure 1). We also analysed the third insertion line, YAB111, at the chromatin level; however, during this analysis, we became aware that this line trapped a microRNA gene, miR398 (Sunkar *et al*., 2012). We also found that the *LUC* insert of YB84 line located in the vicinity of a transposable element (AT5TE27670). Because some transposable elements have promoter activities (Feschotte, 2008), it may be responsible for the transcriptional activation of YB84 line, although we did not observe the evidence of active transcription of this element. These imply that some portion of the cases that are currently thought of as "cryptic promoter capturing" may be reclassified as the capturing of some ncRNA genes or the other non-coding elements. In respect to this, the density of genetic elements other than protein-coding genes in the plant genome deserves further attention. Now we are proceeding with large-scale study focusing on the relative frequency of the cryptic promoter capturing and promoter *de novo* origination (Satoh and Hata *et al.*, 2020).

Our finding of "promoter *de novo* origination" in YB41 raises the question of how it occurs. Because a promoter-specific chromatin structure was not found in the YB41 locus of wild-type plants, it seems likely that the insertion of the *LUC*-trap vector sequence triggered chromatin remodeling to form the core promoter chromatin. The underlying mechanism could be investigated from two angles: the properties inherent to individual genomic insertion sites, and the possible roles of the inserted sequences. Regarding the first aspect, more examples are needed to analyse the common properties of the chromosomal integration sites.

Relevant to the second aspect, the triplicate core promoter experiment (Figure 3) revealed three intriguing properties of the transcription initiation of *LUC* chimeric genes. First, the transcription initiation region could not be determined only by the promoter sequences. Second, the sequences located downstream of the TSS, in this case the *LUC* ORF sequence, appeared to be involved in the determination of the transcription initiation region. Third, once the transcription initiation region was fixed, fine TSS distribution within the region was determined by the region's sequence (Figure 3b and S5). Taking these hierarchical properties of the TSS determination of the *LUC* chimeric gene into consideration, it is very likely that the inserted *LUC* ORF sequence is involved in the positioning process of the newly emerged transcription initiation region proximal to its 5′ end. In this regard, it is quite intriguing that the +1 nucleosomes (3′-flanking nucleosome of the NFR) of YB41, YB84 and YAB111 are all located at the 5′ end of the *LUC* ORF, thus within the inserted T-DNA region. This suggests that the inserted *LUC* ORF sequence provides a suitable site for fixing the +1 nucleosome. The position and remodelling of the +1 nucleosome is important for the nucleosome landscape of the promoter and coding regions (Mavrich *et al*., 2008; Jiang and Pugh, 2009; Möbius and Gerland, 2009; Valen and Sandelin, 2011; Lenhard *et al*., 2012, Klemm *et al*., 2019). The NFR is generally a one-nucleosome-wide region located between the –1 and +1 nucleosomes, and H3K4me3 of these nucleosomes is thought to interact with TAF of TFII-D to localize the PIC (Lenhard et al., 2012; Lauberth *et al*., 2013). Therefore, the mechanism via which the +1 nucleosome is localized at the 5′ end of the inserted *LUC* ORF deserves further attention. In yeast, histone chaperons FACT and Spt6 contribute to the promoter-specific deposition of H2A.Z by selectively preventing the accumulation of H2A.Z within the gene bodies (Jeronimo *et al*., 2015). Similar mechanism by which plant genome localizes H2A.Z in the 5' proximal region of the gene bodies may operate promoter-specific chromatin remodeling (Verbsky and Richards, 2001; Choi *et al*., 2007; Sura *et al*., 2017; Potok *et al*., 2019). This possibility requires further examination.

Although the molecular mechanism underlying the "cryptic promoter capturing" and "promoter *de novo* origination" remains to be analysed, the discovery of these transcriptional activation mechanisms provides important clues to elucidate how newly emerged protein-coding sequences in the plant genome acquire transcriptional competence. For example, in the promoter-acquisition process of endosymbiotic gene transfer from the organelle to the nucleus, these new activation mechanisms will leave negligible traces of the activation process on the transcriptionally activated genes, and yield little damage to the pre-existing nuclear gene network compared with the conventional model of foreign-gene trapping of pre-existing nuclear gene promoters (Kadowaki *et al*., 1996; Kubo *et al*., 1999; Stegemann and Bock, 2006; Wang *et al*., 2014). We speculate that the expression level of the coding sequences that are activated by these mechanisms might generally be as low as the basal transcription level; however, from the evolutionary viewpoint and timescale, once proto-genes (Carvunis *et al*, 2012) or young genes obtain basal transcription activity, they may evolve and acquire better promoter context and elements via subsequent natural selection. We expect that this speculated mechanism will provide a new explanation for the promoter-acquisition process of the following cases: evolution of the protein-coding sequences emerging in response to, for example, stochastic changes in the genome sequences or exon shuffling (Long *et al*, 2003; Kaessmann, 2010; Tautz and Domazet-Lošo, 2011; Carvunis *et al*., 2012,), horizontal gene transfer (Keeling and Palmer 2008; Syvanen, 2012; Soucy *et al*., 2015; Husnik and McCutcheon, 2018) and organelle-to-nucleus DNA flux (Timmis *et al*., 2004; Matsuo *et al*, 2005; Bock, 2017).

Based on the cryptic promoters found in gene-trap screenings of the plant genome, this study led us to find two activation mechanisms of cryptic promoters, and to a speculation regarding the potential impact of these phenomena on the plant genome evolution. Although this study was intensive regarding both time and effort, its final output included only a few gene examples. To extend this study regarding both the number of examples and the depth of the molecular analysis, we are improving its general experimental design to achieve a high-throughput analysis, which will be described elsewhere (Satoh and Hata *et al*., 2020).

MEXPERIMENTAL PROCEDURES

Gene-trap plant lines

LUC-expressing intergenic insertion lines were screened from the gene-trap lines of *Arabidopsis thaliana* (Yamamoto *et al*., 2003).

Northern hybridization analysis

Northern hybridization was performed as described previously (Matsuo and Obokata, 2002). The hybridization probe was prepared from the *LUC* gene using the primer pairs LUCpr1 and LUCpr2 (Table S1).

5′ RACE analysis

5′ RACE was performed using the 5′-full RACE Core Set (TKR6122, TaKaRa) according to the manufacturer's instructions. The primers that hybridized within the luciferase-coding sequence were L-RT, L-S1, L-A1, L-S2 and L-A2 (Table S1).

Determination of the TSS

The TSS were identified in the WT (col-0) and transgenic (YB41, YB84, psaH1::LUC and HC_{111}x3::LUC) seedlings of *A. thaliana* grown on MS agar plates for 10 days using the biotinylated cap-trapper method (Carninci and Hayashizaki, 1999). Total RNAs were extracted from the aerial parts of the seedlings, and cDNAs were synthesized using the following primers: for the transgenic plants, an equimolar mixture of polyT primer (5′-NT_{20}-3′) and the *LUC*-gene-specific reverse primer, LUCR2 (Table S1), was used; for the WT plants, an equimolar mixture of the polyT primer (5′-NT_{20}-3′) and YB41wt or YB84wt primers (Table S1), which hybridize downstream of the YB41 and YB84 junction sites, respectively, was used. The resultant cDNAs were purified using the biotinylated cap-trapper method, to give full-length cDNAs, which were subsequently ligated with a synthetic linker, GN5 (Table S1), at their 3′ ends, and amplified by nested PCR with primers corresponding to the linker and coding-region sequences. The PCR products obtained were cloned into the T-Vector (pMD20, TaKaRa), and the plasmid DNAs obtained were sequenced using an ABI PRISM 3100 Genetic Analyzer (Applied Biosystems).

Real-time RT–PCR analysis of the YB41 and YB84 trap lines

First-strand cDNAs were synthesized from 1 μg of total RNA using ReverTra Ace (TOYOBO) and an oligo dT primer (18-mer). Real-time PCR experiments were performed with the Thunderbird® qPCR Mix (TOYOBO) and the Eco Real-Time PCR System (Illumina). The primers and thermal cycling conditions used for real-time PCR analysis are summarized in Table S2.

ChIP-on-chip analysis

WT (col-0) and transgenic (YB41 and YB84) *A. thaliana* seedlings grown on MS agar plates under continuous white light for 10 days were subjected to cross-linking and chromatin isolation as described by Saleh *et al*. (Saleh *et al*., 2008), with modifications. The isolated nuclei were suspended in MN digestion buffer (500 U/mL of micrococcal nuclease, 3 mM $CaCl_2$, 5 mM $MgCl_2$, 60 mM Kill, 15 mM NaCl, 0.25 M sucrose and 50 mM HEPES; pH 7.5) and incubated at 37 °C for 8 min, and the digestion was stopped by the addition of a 0.25 volume of nuclear lysis buffer (150 mM NaCl, 10 mM EDTA, 1% SDS, 0.1% sodium deoxycholate, 1% Triton X-100, 50 mM HEPES; pH 7.5). After the addition of 10 volumes of ChIP dilution buffer (50 mM HEPES, pH 7.5, 150 mM NaCl, 0.0875% sodium deoxycholate, 0.875% Triton X-100, 1 mg/mL pepstatin A and 1 mg/mL aprotinin), the mixture was centrifuged at 14,000 × *g* for 10 min and the supernatants were subjected to chromatin immunoprecipitation according to the method of Kimura *et al*. (Kimura *et al*., 2008), with slight modifications. Antibodies used in this study were described below. The immunoprecipitated DNAs (IP DNAs) obtained were blunted with T4 DNA polymerase, ligated to the annealed products of linker1 and linker2 (Table S1) and PCR amplified with linker1. The amplified IP DNAs were labelled with the BioPrime Array CGH Labeling System (Invitrogen) according to the manufacturer's instructions. One microgram each of IP DNA and control Input DNA were labelled with cy5 and cy3, respectively, mixed and precipitated with ethanol. The pellet was dissolved in 10 μL of hybridization buffer (10% formamide, 0.02 g dextran sulfate, 3× SSC, 20 μg yeast tRNA, 4% SDS, 20 μg human cot-1 DNA (Invitrogen)), dropped onto a custom-made DNA chip (NGK Insulators Ltd.) (Figure S1), covered with a coverglass and incubated at 42 °C for 24 h in a hybridization chamber. The DNA chip was designed to contain a 60-mer tiling array covering –480 to +300 relative to the genome–T-DNA (YB41 and YB84) junctions and their corresponding WT genomic regions. After incubation, the DNA chip was washed once with 2× SSC with 0.1% SDS at 30 °C for 5 min, once with 2× SSC with 50% formamide (pH 7.0) at 30 °C for 15 min, once with PN buffer (0.1 M NaH_2PO_4/Na_2HPO_4, pH 8.0, 0.1% NP-40) at 30 °C for 30 min, and finally with 2× SSC at room temperature for 5 min. The hybridization signals were analysed on a GenePix 4000B scanner using the GenePix Pro 4.0 software (both from Axon Instruments).

Chimeric promoter constructs and plant transformation

The promoter region (–2000 to +12 relative to the ATG initiation codon) of *psaH1* (AT3G16140.1) of *A. thaliana* was translationally fused with the firefly *LUC* gene and cloned into pPZP221 to obtain the psaH1::LUC construct (sequence details are given in Figure S4). $HC_{111}x3$::LUC was generated by triplicating the 111 bp segment (–111 to –1) of psaH1::LUC (sequence details are given in Figure S5). *Agrobacterium*-mediated transformation of *A. thaliana* col-0 was performed as described previously (Yamamoto *et al*., 2003).

ChIP–PCR analysis of the triplicate core promoter construct

Cross-linking treatment of *A. thaliana* seedlings was performed as described above. Chromatin was isolated according to the method of Gendrel *et al*.

(Gendrel *et al.*, 2005) and was suspended in nuclear lysis buffer (50 mM Tris-HCl (pH 8.0), 10 mM EDTA, 1% SDS, 1 mM PMSF, 2 μg/mL pepstatin A and 2 μg/mL aprotinin). After the addition of nine volumes of ChIP dilution buffer (50 mM Tris-HCl (pH 8.0), 167 mM NaCl, 1.1% Triton X-100, 0.11% sodium deoxycholate, 1 mM PMSF, 1 μg/mL pepstatin A and 1 μg/mL aprotinin), chromatins were fragmented to 50–500 bp, with a peak at 200 bp, by sonication using a UD-201 ultrasonic disruptor (Tomy Seiko). Chromatin immunoprecipitation was performed essentially as described above, using antibodies against pol II and TBP. The IP DNAs obtained were subjected to real-time PCR analysis, as summarized in Table S2.

Antibodies

The anti-*A. thaliana* TBP rabbit polyclonal antibodies were prepared using the synthetic peptides TBP1a (N-PVDLSKHPSGIVPTL-C), TBP1b (N-GFPAKFKDFKIQNIV-C) and TBP1c (N-ENIYPVLSEFRKIQQ-C). These three peptides were injected into different rabbits. The anti-*A. thaliana* H2A.Z rabbit polyclonal antibodies were prepared according to the method of Deal *et al.* (Deal *et al.*, 2007): equal amounts of synthetic peptides representing the N-termini of HTA9 (At1g52740) and HTA11 (At3g54560) were mixed and injected into rabbits. The anti-H3K4me3 mouse polyclonal antibody was described previously (Kimura *et al.*, 2008). A mouse monoclonal antibody (8WG16) against RNA polymerase II CTD repeats SPTSPS was purchased from Abcam. Normal mouse IgG (sc-2025) and rabbit IgG (sc-2027) were purchased from Santa Cruz Biotechnology, Inc.

ACKNOWLEDGEMENTS

We thank M. Sugiura, T. Yoshitsugu, T. Wakasugi, M. Tachikawa and A. Takabayashi for discussions and encouragement; S. Shirakawa, K. Hasegawa, Y. Yasui and M, Ohtani for assistance with experiments; and S. Kushnir for critical reading of this manuscript. This work was supported in part by grants from the MEXT Japan and JSPS (grant numbers 22370002, 23125512, 23657040, 23117006 and 25650131 to JO), from Sekisui Chemical Co. Ltd., and the Strategic Research Funds at Kyoto Prefectural University.

SUPPORTING INFORMATION

Additional supporting information is found in the online version of this article.

Figure S1. Detailed sequence and TSS distributions of the YB41 and YB84 loci.

Figure S2. Custom-made DNA tiling array.

Figure S3. TSS distribution profile of psaH1 of *Arabidopsis thaliana*.

Figure S4. TSS distribution profile of PsaH1::LUC.

Figure S5. TSS distribution profile of HC111x3::LUC.

Table S1. Miscellaneous primers.

Table S2. Primers and thermal cycling conditions of real-time PCR analysis.

REFERENCES

Andersson, R. and Sandelin, A. (2020) Determinants of enhancer and promoter activities of regulatory elements. *Nat Rev Genet,* **21**, 71–87. doi:10.1038/s41576-019-0173-8

Bock, R. (2017) Witnessing Genome Evolution: Experimental Reconstruction of Endosymbiotic and Horizontal Gene Transfer. *Annu Rev Genet,* **51**, 1–22. https://doi.org/10.1146/annurev-genet-120215-035329

Cairns, B. R. (2009) The logic of chromatin architecture and remodelling at promoters. *Nature,* **461**, 193–198. https://doi.org/10.1038/nature08450.

Carninci, P. and Hayashizaki, Y. (1999) High-efficiency full-length cDNA cloning. *Methods Enzymol,* **303**, 19–44. https://doi.org/10.1016/s0076-6879(99)03004-9

Carvunis, A. R., Rolland, T., Wapinski, I., Calderwood, M. A., Yildirim, M. A., Simonis, N., Charloteaux, B., Hidalgo, C. A., Barbette, J., Santhanam, B., Brar, G. A., Weissman, J. S., Regev, A., Thierry-Mieg, N., Cusick, M. E. and Vidal, M. (2012) Proto-genes and de novo gene birth. *Nature,* **487**, 370–374. https://doi.org/10.1038/nature11184

Choi, K., Park, C., Lee, J., Oh, M., Noh, B. and Lee, I. (2007) Arabidopsis homologs of components of the SWR1 complex regulate flowering and plant development. *Development,* **134**, 1931–1941. https://doi.org/10.1242/dev.001891

Czechowski, T., Stitt, M., Altmann, T., Udvardi, M. K. and Scheible, W. R. (2005) Genome-wide identification and testing of superior reference genes for transcript normalization in Arabidopsis. *Plant Physiol,* **139**, 5–17. https://doi.org/10.1104/pp.105.063743

Deal, R. B. and Henikoff, S. (2011) Histone variants and modifications in plant gene regulation. *Curr Opin Plant Biol,* **14**, 116-122. https://doi.org/10.1016/j.pbi.2010.11.005

Deal, R. B., Topp, C. N., McKinney, E. C. and Meagher, R. B. (2007) Repression of flowering in Arabidopsis requires activation of

FLOWERING LOCUS C expression by the histone variant H2A.Z. *Plant Cell,* **19**, 74–83. https://doi.org/10.1105/tpc.106.048447

Feschotte, C. (2008) Transposable elements and the evolution of regulatory networks. *Nat Rev Genet,* **9**, 397–405. https://doi.org/10.1038/nrg2337

Fobert, P. R., Labbé, H., Cosmopoulos, J., Gottlob-McHugh, S., Ouellet, T., Hattori, J., Sunohara, G., Iyer, V. N. and Miki, B. L. (1994) T-DNA tagging of a seed coat-specific cryptic promoter in tobacco. *Plant J,* **6**, 567–577. https://doi.org/10.1046/j.1365-313x.1994.6040567.x

Gendrel, A. V., Lippman, Z., Martienssen, R. and Colot, V. (2005) Profiling histone modification patterns in plants using genomic tiling microarrays. *Nat Methods,* **2**, 213–218. https://doi.org/10.1038/nmeth0305-213

Giaimo, B. D., Ferrante, F., Herchenröther, A., Hake, S. B. and Borggrefe, T. (2019) The histone variant H2A.Z in gene regulation. *Epigenetics Chromatin,* **12**, 37. https://doi.org/10.1186/s13072-019-0274-9.

Guenther, M. G., Levine, S. S., Boyer, L. A., Jaenisch, R. and Young, R. A. (2007) A chromatin landmark and transcription initiation at most promoters in human cells. *Cell,* **130**, 77–88. https://doi.org/10.1016/j.cell.2007.05.042

Haberle, V. and Stark, A. (2018) Eukaryotic core promoters and the functional basis of transcription initiation. *Nat Rev Mol Cell Biol,* **19**, 621–637. https://doi.org/10.1038/s41580-018-0028-8

Hennig, B. P., Bendrin, K., Zhou, Y. and Fischer, T. (2012) Chd1 chromatin remodelers maintain nucleosome organization and repress cryptic transcription. *EMBO Rep,* **13**, 997–1003. https://doi.org/10.1038/embor.2012.146

Hennig, B. P. and Fischer, T. (2013) The great repression: chromatin and cryptic transcription. *Transcription,* **4**, 97–101. https://doi.org/10.4161/trns.24884

Husnik, F. and McCutcheon, J. P. (2018) Functional horizontal gene transfer from bacteria to eukaryotes. *Nat Rev Microbiol,* **16**, 67–79. https://doi.org/10.1038/nrmicro.2017.137

Hyun, K., Jeon, J., Park, K. and Kim, J. (2017) Writing, erasing and reading histone lysine methylations. *Exp Mol Med,* **49**, e324. https://doi.org/10.1038/emm.2017.11

Jeronimo, C., Watanabe, S., Kaplan, C. D., Peterson, C. L. and Robert, F. (2015) The Histone Chaperones FACT and Spt6 Restrict H2A.Z from Intragenic Locations. *Mol Cell,* **58**, 1113–1123. https://doi.org/10.1016/j.molcel.2015.03.030

Jiang, C. and Pugh, B. F. (2009) Nucleosome positioning and gene regulation: advances through genomics. *Nat Rev Genet,* **10**, 161–172. https://doi.org/10.1038/nrg2522

Kadowaki, K., Kubo, N., Ozawa, K. and Hirai, A. (1996) Targeting presequence acquisition after mitochondrial gene transfer to the nucleus occurs by duplication of existing targeting signals. *EMBO J,* **15**, 6652–6661. https://doi.org/10.1002/j.1460-2075.1996.tb01055.x

Kaessmann, H. (2010) Origins, evolution, and phenotypic impact of new genes. *Genome Res,* **20**, 1313–1326. https://doi.org/10.1101/gr.101386.109

Keeling, P. J. and Palmer, J. D. (2008) Horizontal gene transfer in eukaryotic evolution. *Nat Rev Genet,* **9**, 605–618. https://doi.org/10.1038/nrg2386

Kimura, H., Hayashi-Takanaka, Y., Goto, Y., Takizawa, N. and Nozaki, N. (2008) The organization of histone H3 modifications as revealed by a panel of specific monoclonal antibodies. *Cell Struct Funct,* **33**, 61–73. https://doi.org/10.1247/csf.07035

Klemm, S. L., Shipony, Z. and Greenleaf, W. J. (2019) Chromatin accessibility and the regulatory epigenome. *Nat Rev Genet,* **20**, 207–220. https://doi.org/10.1038/s41576-018-0089-8

Kubo, N., Harada, K., Hirai, A. and Kadowaki, K. (1999) A single nuclear transcript encoding mitochondrial RPS14 and SDHB of rice is processed by alternative splicing: common use of the same mitochondrial targeting signal for different proteins. *Proc Natl Acad Sci U S A,* **96**, 9207–9211. https://doi.org/10.1073/pnas.96.16.9207

Lauberth, S. M., Nakayama, T., Wu, X., Ferris, A. L., Tang, Z., Hughes, S. H. and Roeder, R. G. (2013) H3K4me3 interactions with TAF3 regulate preinitiation complex assembly and selective gene activation. *Cell,* **152**, 1021–1036. https://doi.org/10.1016/j.cell.2013.01.052

Lenhard, B., Sandelin, A. and Carninci, P. (2012) Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat Rev Genet,* **13**, 233–245. https://doi.org/10.1038/nrg3163

Li, B., Carey, M. and Workman, J. L. (2007) The role of chromatin during transcription. *Cell,* **128**, 707–719. https://doi.org/10.1016/j.cell.2007.01.015

Long, M., Betrán, E., Thornton, K. and Wang, W. (2003) The origin of new genes: glimpses from the young and old. *Nat Rev Genet,* **4**, 865–875. https://doi.org/10.1038/nrg1204

Martin, W., Stoebe, B., Goremykin, V., Hapsmann, S., Hasegawa, M. and Kowallik, K. V. (1998) Gene transfer to the nucleus and the evolution of chloroplasts. *Nature,* **393**, 162–165. https://doi.org/10.1038/30234

Matsuo, M., Ito, Y., Yamauchi, R. and Obokata, J. (2005) The rice nuclear genome continuously integrates, shuffles, and eliminates the chloroplast genome to cause chloroplast-nuclear DNA flux. *Plant Cell,* **17**, 665–675. https://doi.org/10.1105/tpc.104.027706

Matsuo, M. and Obokata, J. (2002) Dual roles of photosynthetic electron transport in photosystem I biogenesis: light induction of mRNAs and chromatic regulation at post-mRNA level. *Plant Cell Physiol,* **43**, 1189–1197. doi:10.1093/pcp/pcf146

Mavrich, T. N., Ioshikhes, I. P., Venters, B. J., Jiang, C., Tomsho, L. P., Qi, J., Schuster, S. C., Albert, I. and Pugh, B. F. (2008) A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res,* **18**, 1073–1083. https://doi.org/10.1101/gr.078261.108

Mollier, P., Hoffmann, B., Orsel, M. and Pelletier, G. (2000) Tagging of a cryptic promoter that confers root-specific gus expression in Arabidopsis thaliana. *Plant Cell Rep,* **19**, 1076–1083. https://doi.org/10.1007/s002990000241

Möbius, W. and Gerland, U. (2010) Quantitative test of the barrier nucleosome model for statistical positioning of nucleosomes up- and downstream of transcription start sites. *PLoS Comput Biol,* **6**. e-1000891. https://doi.org/10.1371/journal.pcbi.1000891

Neri, F., Rapelli, S., Krepelova, A., Incarnato, D., Parlato, C., Basile, G., Maldotti, M., Anselmi, F. and Oliviero, S. (2017) Intragenic DNA methylation prevents spurious transcription initiation. *Nature,* **543**, 72–77. https://doi.org/10.1038/nature21373

Ökrész, L., Máthé, C., Horváth, É., Schell, J., Koncz, C. and Szabados, L. (1998) T-DNA trapping of a cryptic promoter identifies an ortholog of highly conserved SNZ growth arrest response genes in Arabidopsis. *Plant Sci.* **138**, 217–228. https://doi.org/10.1016/S0168-9452(98)00163-0

Plesch, G., Kamann, E. and Mueller-Roeber, B. (2000) Cloning of regulatory sequences mediating guard-cell-specific gene expression. *Gene,* **249**, 83–89. https://doi.org/10.1016/s0378-1119(00)00150-5

Potok, M. E., Wang, Y., Xu, L., Zhong, Z., Liu, W., Feng, S., Naranbaatar, B., Rayatpisheh, S., Wang, Z., Wohlschlegel, J. A., Ausin, I. and Jacobsen, S. E. (2019) Arabidopsis SWR1-associated protein methyl-CpG-binding domain 9 is required for histone H2A.Z deposition. *Nat Commun,* **10**, 3352. https://doi.org/10.1038/s41467-019-11291-w

Saleh, A., Alvarez-Venegas, R. and Avramova, Z. (2008) An efficient chromatin immunoprecipitation (ChIP) protocol for studying histone modifications in Arabidopsis plants. *Nat Protoc,* **3**, 1018–1025. https://doi.org/10.1038/nprot.2008.66

Satoh, S., Hata, T., Takada, N., Tachikawa, M., Matsuo, M., Kushnir, S., and Obokata, J. Plant genome response to incoming coding sequences: stochastic transcriptional activation independent of chromatin configuration. (submitted in The Plant Journal simulataneously with this manuscript)

Sivanandan, C., Sujatha, T. P., Prasad, A. M., Resminath, R., Thakare, D. R., Bhat, S. R. and Srinivasan, R. (2005) T-DNA tagging and characterization of a cryptic root-specific promoter in Arabidopsis. *Biochim Biophys Acta,* **1731**, 202–208. https://doi.org/10.1016/j.bbaexp.2005.10.006

Soucy, S. M., Huang, J. and Gogarten, J. P. (2015) Horizontal gene transfer: building the web of life. *Nat Rev Genet,* **16**, 472–482. https://doi.org/10.1038/nrg3962

Springer, P. S. (2000) Gene traps: tools for plant development and genomics. *Plant Cell,* **12**, 1007-20. *Plant Cell,* ***12***, 1007–1020. https://doi.org/10.1105/tpc.12.7.1007

Stanford, W. L., Cohn, J. B. and Cordes, S. P. (2001) Gene-trap mutagenesis: past, present and beyond. *Nat Rev Genet,* **2**, 756–768. https://doi.org/10.1038/35093548

Stangeland, B., Nestestog, R., Grini, P. E., Skrbo, N., Berg, A., Salehian, Z., Mandal, A. and Aalen, R. B. (2005) Molecular analysis of Arabidopsis endosperm and embryo promoter trap lines: reporter-gene expression can result from T-DNA insertions in antisense orientation, in introns and in intergenic regions, in addition to sense insertion at the 5' end of genes. *J Exp Bot,* **56**, 2495–2505. https://doi.org/10.1093/jxb/eri242

Stegemann, S. and Bock, R. (2006) Experimental reconstruction of functional gene transfer from the tobacco plastid genome to the nucleus. *Plant Cell,* **18**, 2869–2878. https://doi.org/10.1105/tpc.106.046466

Sunkar, R., Li, Y. F. and Jagadeeswaran, G. (2012) Functions of microRNAs in plant stress responses. *Trends Plant Sci,* **17**, 196–203. https://doi.org/10.1016/j.tplants.2012.01.010

Sura, W., Kabza, M., Karlowski, W. M., Bieluszewski, T., Kus-Slowinska, M., Pawełoszek, Ł., Sadowski, J. and Ziolkowski, P. A. (2017) Dual Role of the Histone Variant H2A.Z in Transcriptional Regulation of Stress-Response Genes. *Plant Cell,* **29**, 791–807. https://doi.org/10.1105/tpc.16.00573

Syvanen, M. (2012) Evolutionary implications of horizontal gene transfer. *Annu Rev Genet,* **46**, 341–358. https://doi.org/10.1146/annurev-genet-110711-155529

Tautz, D. and Domazet-Lošo, T. (2011) The evolutionary origin of orphan genes. *Nat Rev Genet,* **12**, 692–702. https://doi.org/10.1038/nrg3053

Timmis, J. N., Ayliffe, M. A., Huang, C. Y. and Martin, W. (2004) Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nat Rev Genet,* **5**, 123–135. https://doi.org/10.1038/nrg1271

Topping, J. F., Agyeman, F., Henricot, B. and Lindsey, K. (1994) Identification of molecular markers of embryogenesis in Arabidopsis thaliana by promoter trapping. *Plant J,* **5**, 895–903. https://doi.org/10.1046/j.1365-313x.1994.5060895.x

Valen, E. and Sandelin, A. (2011) Genomic and chromatin signals underlying transcription start-site selection. *Trends Genet,* **27**, 475–485. https://doi.org/10.1016/j.tig.2011.08.001

Verbsky, M. L. and Richards, E. J. (2001) Chromatin remodeling in plants. *Curr Opin Plant Biol,* **4**, 494–500. https://doi.org/10.1016/s1369-5266(00)00206-5

Wang, D., Qu, Z., Adelson, D. L., Zhu, J. K. and Timmis, J. N. (2014) Transcription of nuclear organellar DNA in a model plant system. *Genome Biol Evol,* **6**, 1327–1334. https://doi.org/10.1093/gbe/evu111

Weber, C. M. and Henikoff, S. (2014) Histone variants: dynamic punctuation in transcription. *Genes Dev,* **28**, 672–682. https://doi.org/10.1101/gad.238873.114

Yamamoto, Y. Y., Tsuhara, Y., Gohda, K., Suzuki, K. and Matsui, M. (2003) Gene trapping of the Arabidopsis genome with a firefly luciferase reporter. *Plant J,* **35**, 273–283. https://doi.org/10.1046/j.1365-313x.2003.01797.x

Yamamoto, Y. Y., Yoshitsugu, T., Sakurai, T., Seki, M., Shinozaki, K. and Obokata, J. (2009) Heterogeneity of Arabidopsis core promoters revealed by high-density TSS analysis. *Plant J,* **60**, 350–362. https://doi.org/10.1111/j.1365-313X.2009.03958.x

Zhang, X., Bernatavichute, Y. V., Cokus, S., Pellegrini, M. and Jacobsen, S. E. (2009) Genome-wide analysis of mono-, di- and trimethylation of histone H3 lysine 4 in Arabidopsis thaliana. *Genome Biol,* **10**, R62. https://doi.org/10.1186/gb-2009-10-6-r62

FIGURE LEGENDS

Figure 1. Two cryptic promoters found in the genome of *Arabidopsis thaliana* by gene trap screening. (**a**) Schematic structure of a luciferase (*LUC*)-based trap vector (Yamamoto *et al*., 2003) containing RB (right border), LB (left border), SD (splicing donor), SA (splicing acceptor), pNOS (NOS promoter) and NPTII (kanamycin-resistance marker). (**b**) Genomic maps of the YB41 and YB84 insertion sites. (**c**) TSS distributions of the inserted *LUC* genes identified using the cap-trapper method. Genomic regions used for RT-PCR in (d) are also indicated. (**d**) Transcript levels of YB41 and YB84 integration sites determined by real-time RT-PCR analysis and normalized to the intrinsic ubiquitin (AT5G25760) mRNA level (Czechowski *et al*., 2005). Data are means ± s.d.(n=3)

Figure 2. Chromatin states of the YB41 and YB84 loci before and after the insertion of a gene trap vector. (**a**) ChIP-on-chip analysis of the YB41 integration site before (right) and after (left) the T-DNA insertion. Tiling array (grey blocks) covers -480 to +300 base region relative to the genome-T-DNA junction. (**b**) ChIP-on-chip analysis of the YB84 integration site as in (a). (**c**) Schematic illustration of the chromatin structure found at the YB41 and YB84 core promoters.

Figure 3. Triplicate core-promoter experiments to investigate the localization mechanism of PIC/TSS in the plant genome. (**a**) Experimental design and hypothesis. (**b**) Schematic model of the WT *psaH1* gene and chimeric promoter constructs, and TSS distributions on them. Vertical bars indicate TSS tag numbers detected at the respective genomic sites, and the heights of the peak TSSs of three genes are normalized to the same size. The total numbers of TSS tags determined for each gene are shown in parentheses as n. Triangles represent the 111 bp core promoter segment. (**c**) Strategy for the ChIP-PCR analysis of the triplicate segments. Primer sets a, b and c each can amplify both single-unit (solid line) and double-unit (dotted line) fragments of the sizes indicated. (**d**) Fine tuning of PCR conditions allowed us to amplify only single-unit fragments using primer sets a, b and c (experimental details are described in Table S2). PCR products of the chimeric core promoter segments were amplified only in the transgenic plants (Tr) and not in the wild-type plants (WT). A control primer set for an intrinsic gene (*psaL*) amplified PCR products from both WT and Tr. (**e**) ChIP-PCR analysis of the triplicate core-promoter segments A, B and C, using antibodies against pol II and TBP. ChIP signal intensities of each segment to indicate the recovery of input DNA were normalized to that of segment a, which was set at 1.0. Dotted lines indicate background levels of the control ChIP signal using mouse IgG (pol II) or rabbit IgG (TBP). Data are means ± s.d. (n=3)

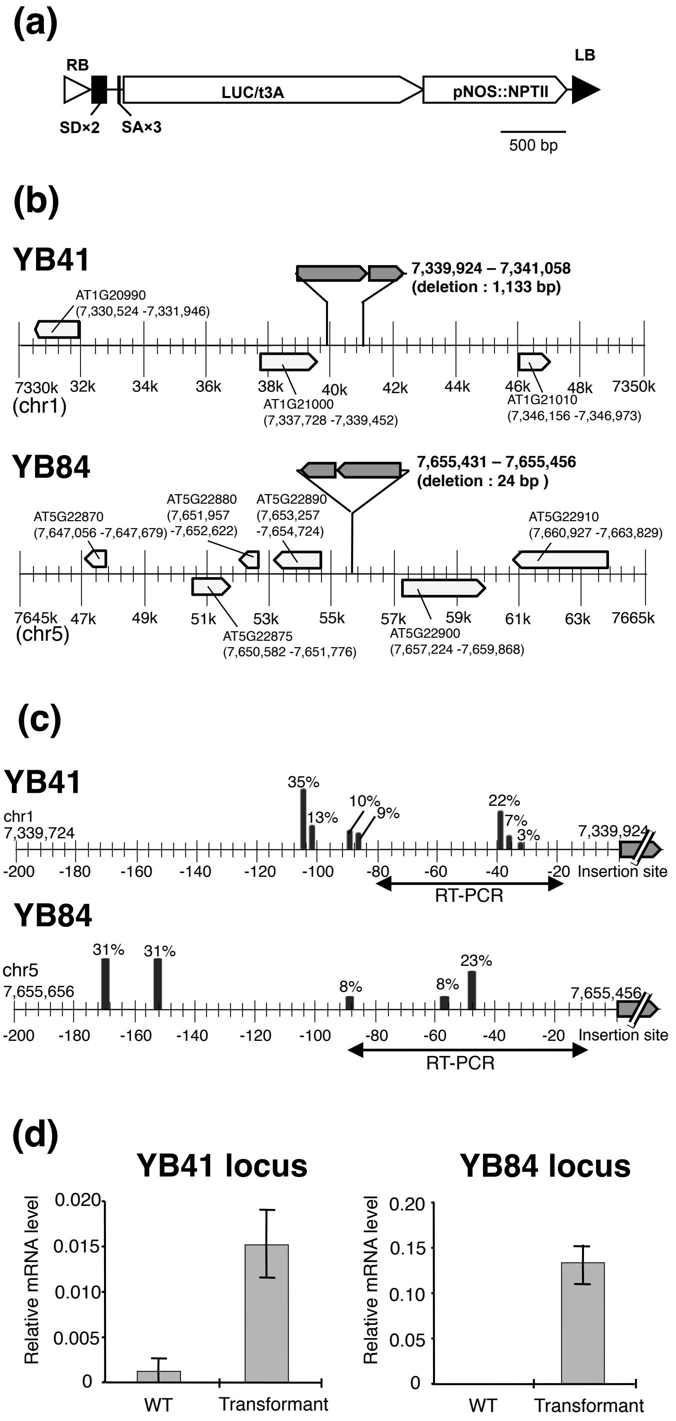

(a)
RB
LB
LUC/t3A
pNOS::NPTII
SD×2
SA×3
500 bp
(b)
YB41
7,339,924 – 7,341,058
(deletion : 1,133 bp)
AT1G20990
(7,330,524 -7,331,946)
7330k
(chr1)
32k
34k
36k
38k
40k
42k
44k
46k
48k
7350k
AT1G21000
(7,337,728 -7,339,452)
AT1G21010
(7,346,156 -7,346,973)
YB84
7,655,431 – 7,655,456
(deletion : 24 bp)
AT5G22870
(7,647,056 -7,647,679)
AT5G22880
(7,651,957
-7,652,622)
AT5G22890
(7,653,257
-7,654,724)
AT5G22910
(7,660,927 -7,663,829)
7645k
(chr5)
47k
49k
51k
53k
55k
57k
59k
61k
63k
7665k
AT5G22875
(7,650,582 -7,651,776)
AT5G22900
(7,657,224 -7,659,868)
(c)
YB41
chr1
7,339,724
35%
13%
10%
9%
22%
7%
3%
7,339,924
-200
-180
-160
-140
-120
-100
-80
-60
-40
-20
Insertion site
RT-PCR
YB84
chr5
7,655,656
31%
31%
8%
8%
23%
7,655,456
-200
-180
-160
-140
-120
-100
-80
-60
-40
-20
Insertion site
RT-PCR
(d)
YB41 locus
Relative mRNA level
0.020
0.015
0.010
0.005
0
WT
Transformant
YB84 locus
Relative mRNA level
0.20
0.15
0.10
0.05
0
WT
Transformant

Figure 1

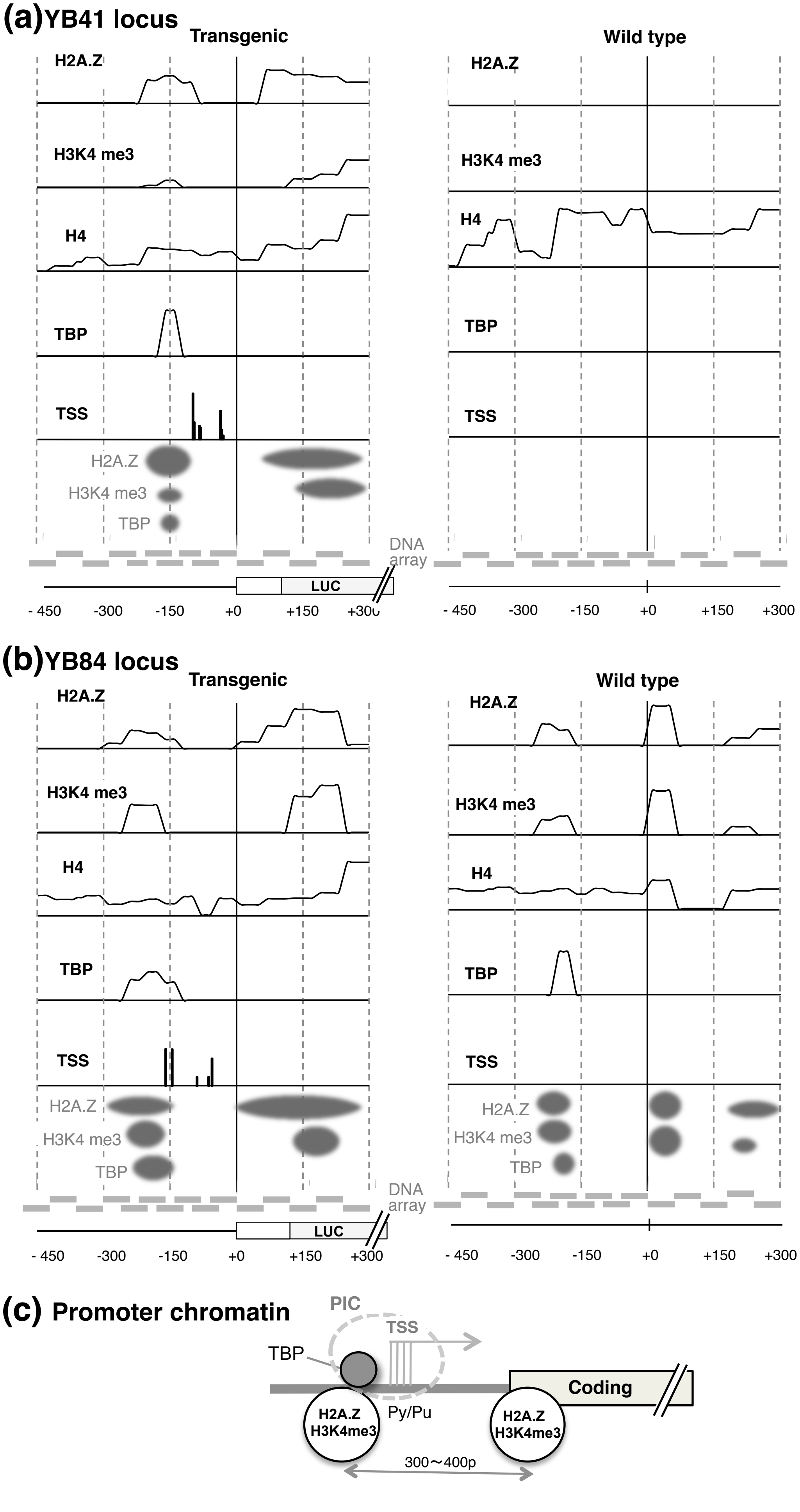

(a) YB41 locus
Transgenic
Wild type
H2A.Z
H3K4 me3
H4
TBP
TSS
H2A.Z
H3K4 me3
TBP
DNA array
LUC
- 450
-300
-150
+0
+150
+300
(b) YB84 locus
Transgenic
Wild type
H2A.Z
H3K4 me3
H4
TBP
TSS
H2A.Z
H3K4 me3
TBP
DNA array
LUC
- 450
-300
-150
+0
+150
+300
(c) Promoter chromatin
PIC
TSS
TBP
Coding
H2A.Z H3K4me3
Py/Pu
H2A.Z H3K4me3
300〜400p

Figure 2

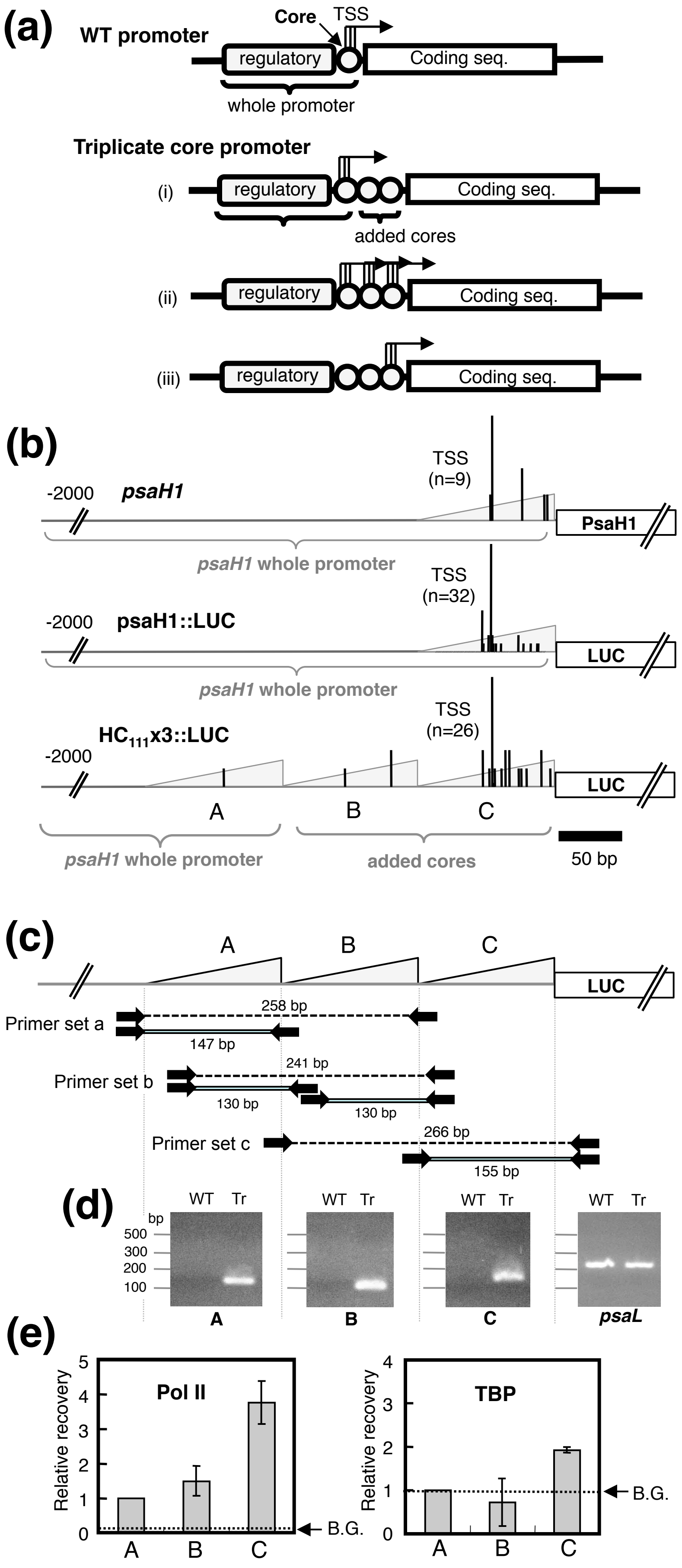

(a)
WT promoter
Core
TSS
regulatory
Coding seq.
whole promoter
Triplicate core promoter
(i)
regulatory
Coding seq.
added cores
(ii)
regulatory
Coding seq.
(iii)
regulatory
Coding seq.
(b)
-2000
psaH1
TSS
(n=9)
PsaH1
psaH1 whole promoter
-2000
psaH1::LUC
TSS
(n=32)
LUC
psaH1 whole promoter
$HC_{111}x3$::LUC
-2000
TSS
(n=26)
LUC
A
B
C
psaH1 whole promoter
added cores
50 bp
(c)
A
B
C
LUC
Primer set a
258 bp
147 bp
Primer set b
241 bp
130 bp
130 bp
Primer set c
266 bp
155 bp
(d)
bp
500
300
200
100
WT
Tr
A
B
C
psaL
(e)
Pol II
Relative recovery
0
1
2
3
4
5
A
B
C
B.G.
TBP
Relative recovery
0
1
2
3
4
A
B
C
B.G.

Figure 3

Table S1. Miscellaneous primers.

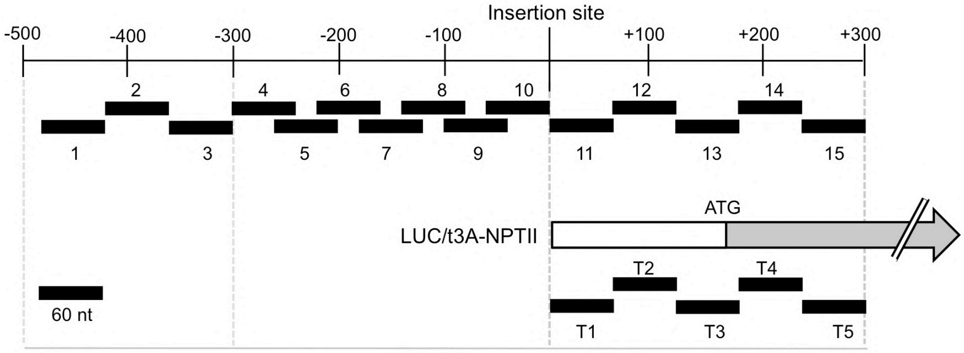
Primer Name	usage	Sequence
LUCpr1	Preparation of LUC hybridization probe	5’-GATTGACAAATACGATTTATCTAATTTACA-3’
LUCpr2	Preparation of LUC hybridization probe	5’-AAATGTTCGTCTTCGTCCCAGTAAGCTATG-3’
L-RT primier	5’ RACE analysis of LUC trap vector	5’-CCCTGGTAATCCGTT-3’
L-S1 primer	5’ RACE analysis of LUC trap vector	5’-TTGGGCGCGTTATTTATCGG-3’
L-A1 primer	5’ RACE analysis of LUC trap vector	5’-AACCAGGGCGTATCTCTTCA-3’
L-S2 primer	5’ RACE analysis of LUC trap vector	5’-CAACAGTATGGGCATTTCGC-3’
L-A2 primer	5’ RACE analysis of LUC trap vector	5’-TCCAGCGGTTCCATCTTCCA-3’
LUCR2	Determination of TSS	5'-CCATCCTCTAGAGGATAGAATGGC-3'
YB41wt	Determination of TSS	5'-TGAAAATATGTGATTACCGCCTTA-3'
YB84wt	Determination of TSS	5'-GGGACACACGTTAGGTTACATTCCA-3'
GN5 linker	Determination of TSS	5'-AGAGAGAGGCTCGAGCTCTATTTAGGTGACACTATAGAACCAGNNNNN-3'
linker1	Probe preparation for ChIP-on-chip	5'-GCGGTGACCCGGGAGATCTGAATTC-3'
linker2	Probe preparation for ChIP-on-chip	5'-GAATTCAGATC-3'

Table S2. Primers and thermal cycling conditions of real-time PCR analysis.

Primer Name	Sequence	PCR condition	Target
YB41-Fw	5'-TAAACTCTCTACTAATCCACG-3'	95˚C 1min, (95˚C 15sec, 50˚C 30sec, 72˚C 30sec) x40	YB41 locus
YB41-Rv	5'-AATTTTATAAGCAATAACAG-3'	95˚C 1min, (95˚C 15sec, 50˚C 30sec, 72˚C 30sec) x40	YB41 locus
YB84-Fw	5'-TATACATGACAAATGCATTTAG-3'	95˚C 1min, (95˚C 15sec, 50˚C 30sec, 72˚C 30sec) x40	YB84 locus
YB84-Rv	5'-CGCTAGAAAAGTTCAATAGAC-3'	95˚C 1min, (95˚C 15sec, 50˚C 30sec, 72˚C 30sec) x40	YB84 locus
UBQ10-Fw*[1]	5'-GGCCTTGTATAATCCCTGATGAATAAG-3'	95˚C 1min, (95˚C 15sec, 60˚C 1min) x40	*UBQ10*
UBQ10-Rv*[1]	5'-AAAGAGATAACAGGAACGGAAACATAGT-3'	95˚C 1min, (95˚C 15sec, 60˚C 1min) x40	*UBQ10*
coreA-Fw	5'-AATCCCCGGGAGGAAATCAGT-3'	94˚C 1min, (94˚C 5sec, 60˚C 15sec) x50	*psaH* triplicate core promoter
coreA-Rv*[2]	5'-AGAGTACACGTGGCAGAGCTTCG-3'	94˚C 1min, (94˚C 5sec, 60˚C 15sec) x50	*psaH* triplicate core promoter
coreB-Fw*[2]	5'-CTTGGAAAACTCGACTCATTGT-3'	94˚C 1min, (94˚C 5sec, 60˚C 15sec) x50	*psaH* triplicate core promoter
coreB-Rv*[2]	5'-CGCTCTCGAGTTTTGATCTTCT-3'	94˚C 1min, (94˚C 5sec, 60˚C 15sec) x50	*psaH* triplicate core promoter
coreC-Fw*[2]	5'-CACTAAGTGATCGAAGCTCTGC-3'	94˚C 1min, (94˚C 5sec, 60˚C 15sec) x50	*psaH* triplicate core promoter
coreC-Rv	5'-CCATTTGGATCCCAAGAGA-3'	94˚C 1min, (94˚C 5sec, 60˚C 15sec) x50	*psaH* triplicate core promoter
psaL-Fw	5'-AGCTGGCCCATTAAGGAAC-3'	95˚C 1min, (80˚C 5sec, 60˚C 30sec, 72˚C 30sec) x10, (94˚C 15sec, 60˚C 30sec, 72˚C 30sec) x35	*psaL* exon2
psaL-Rv	5'-AGTGAACTTAGCCCATCCATC-3'	95˚C 1min, (80˚C 5sec, 60˚C 30sec, 72˚C 30sec) x10, (94˚C 15sec, 60˚C 30sec, 72˚C 30sec) x35	*psaL* exon2

*[1] According to Czechowski *et al*. (Czechowski *et al*.,2005)

*[2] Primers named core are used for specific amplification of the unit segment A, B or C from the triplicate core promoter ABC (Fig. 3c). Red letters indicate mutations that prevent inadequately hybridized PCR products from extending the strand and exponentially amplifying the multimeric products. Using this technical innovation and carefully choosing the PCR conditions that favor the preferential amplification of single-unit-size fragments rather than oligo- or multi-unit fragments (described in this table) allowed us to selectively amplify the single-unit fragment, A, B and C as shown in Fig. 3d.



```
YB41        1： 5’-TGACTAGTGACTACTATATTTAGTCTATATCACCATTTTGTGACATTCATCTAAAACAAA-3’
            2： 5’-TGTTAGTGTAGTCTAATCCTTTCATGAACTGTTTTATATGGTAATGTGTAACTGAAGTCA-3’
            3： 5’-TGTCGTCATGGTTACATGGTTTGTATAAAAGTCGTAAGAGTGACACTGAGACATATTTTT-3’
            4： 5’-CAATACTATTGATTGATTTGTTGTAGTTTGACGTAATCAAGAAGAATATTATAATAGTAT-3’
            5： 5’-GAAGAATATTATAATAGTATACGCTAAAGAATATTATTATGCGGTTTAAAAAGTGAGCAA-3’
            6： 5’-GCGGTTTAAAAAGTGAGCAAGTGGATAGAGGAAAATAAGAGCCGTGTGGAATCCGAAAGG-3’
            7： 5’-GCCGTGTGGAATCCGAAAGGCTGCGATTCTCTCTCACCGTTAAGAGACGGGTGTTCATTC-3’
            8： 5’-TAAGAGACGGGTGTTCATTCCTTTTTCTGATCGCTCAAAATTATTCAACATCATATTTTA-3’
            9： 5’-TTATTCAACATCATATTTTAAACTCTCTACTAATCCACGGAATCGAATAATTGGCAAACT-3’
           10： 5’-AATCGAATAATTGGCAAACTGTTATTGCTTATAAAATTTCATTAGTTATAATAATGTTTT-3’
           11： 5’-AATTATGCCATTTCGTAAACACAAAAATCGTTTTCTGTAGCTTTTTAGGTATCCCACAAA-3’
           12： 5’-TTTACGTTTTCTTCTTTTCACTAGGTTTTCACACTTAAAGCATCACAACTTTTTACTTTT-3’
           13： 5’-ACCAAAAAAGTTCCGAGCTTGGTTCATAATTTCTCGGAGTCGAAAACTAGAATTTCAAAT-3’
           14： 5’-TATCTTGTTTAGGATCACTACAAATTTTATTTTACGAAAATGTTAACATCAAAAATAATT-3’
           15： 5’-TATATAAGGCGGTAATCACATATTTTCAGTATATATAGTTTTTTCACTGTAGTTTGGTTT-3’

YB84        1： 5’-AAAGAAAAAAAAAATATGTGTGCATGTAAGCAGATAAGAGAACTGAATGTAGTCGCAACTT-3’
            2： 5’-TTTTTTTAAAGGATTGGATTTGACATAAGATATTATATAGATGTGAGCTTGACAGTTACA-3’
            3： 5’-GAACGTATTGTATATAGAACAATACAAGAAAAGGCCATGAAAGAACACTAACACTTGACT-3’
            4： 5’-TAGGTTAATCATTCTTTAATTACATCGGGATTGACCTTCTTCGTTTATTTAACTCATAAG-3’
            5： 5’-TCGTTTATTTAACTCATAAGATGTAGCTGACAAAGTCTAGTTAAGGTTTCATCCGGGGGA-3’
            6： 5’-TTAAGGTTTCATCCGGGGGAGGATTTATGTGGGGATGGTCGCTTTTGCTATCTCGAGTTG-3’
            7： 5’-GCTTTTGCTATCTCGAGTTGACTTCCAAGTTAATCCTAACCACATCACATACTTCTATTA-3’
            8： 5’-CACATCACATACTTCTATTATTCCATTTTCAACCTAATGTAAATAATTCATATACATGAC-3’
            9： 5’-AAATAATTCATATACATGACAAATGCATTTAGAGAGAAGAAAAAACATCATTCATGACGA-3’
           10： 5’-AAAAACATCATTCATGACGACAAAAACTTAGTCTATTGAACTTTTCTAGCGTTTTCTCTT-3’
           11： 5’-GCCAGCATCCGGGTTCATCAAAATCTGTATATTGCATATTATTGTCATTATTGACAAAAA-3’
           12： 5’-GACATTATTTACAACTTCTATGTATCGTATATTTATAAACGTCTAGATGTCTTTTTAATT-3’
           13： 5’-TAAAACATCTTTGTAAATCTTTAATTAAATATCTGATTTTAAATAAAATCAAACATAAAC-3’
           14： 5’-ATAAACAAATATAGTTGTGGAAGTCCTTTGGCCCAATGGTTGATTAAGGGTTCATCAACC-3’
           15： 5’-CCCGCTAATTTTTTTTTTTCTTTCTAAAACAGATGAATGTAACCTAACGTGTGTCCCGTCT-3’

YY323 gene trap vector T1： 5’-AAGCTTCAAGGTCTCTCTTCAAGGTGAGTTTTTTTCTGTTCACTCTCTTAGATGCCAAAA-3’
                       T2： 5’-CTTGAGTTATTGCTTAATGTTTCAATTGTTGTGGACTCTGTGTATGTGTAGGTTATATGC-3’
                       T3： 5’-AGGTTATATGCAGGTTATATGGGAGGTGGAGGGATCCAAACAATGGCTATGGCTGAAGAC-3’
                       T4： 5’-GCCAAAAACATAAAGAAAGGCCCGGCGCCATTCTATCCTCTAGAGGATGGAACCGCTGGA-3’
                       T5： 5’-GAGCAACTGCATAAGGCTATGAAGAGATACGCCCTGGTTCCTGGAACAATTGCTTTTACA-3’
```

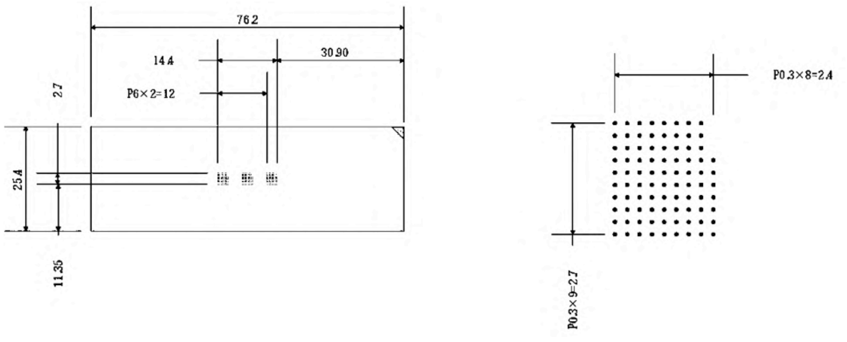
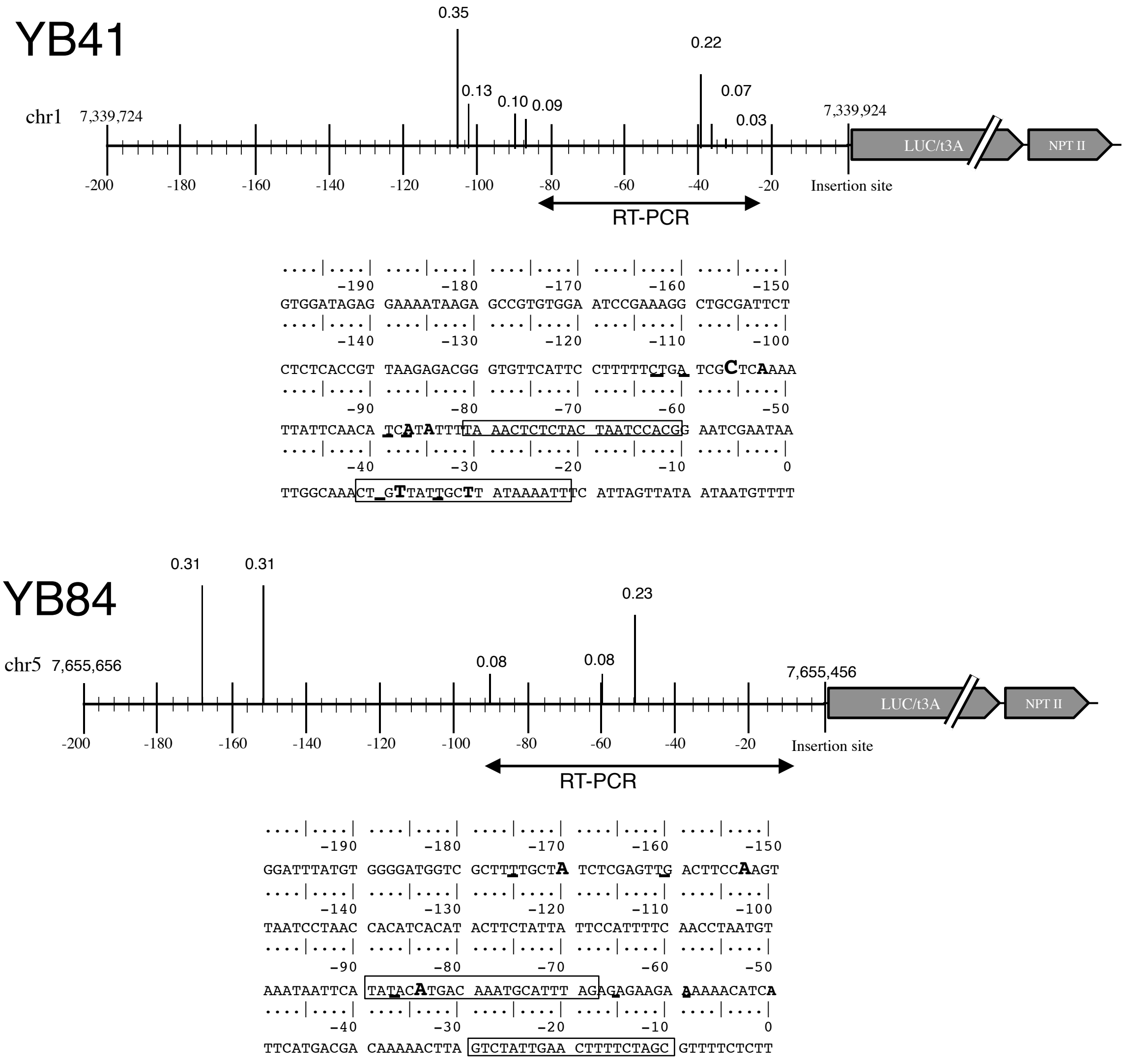


Figure S1. Custom-made DNA tiling array. The top panel shows the layout of DNA fragments 1–15, and T1–T5, on the schematic gene map. Nucleotide sequences of the synthetic 60-mer DNA fragments are shown in the middle panel. The bottom panel shows the spotting layout on the slide glass.



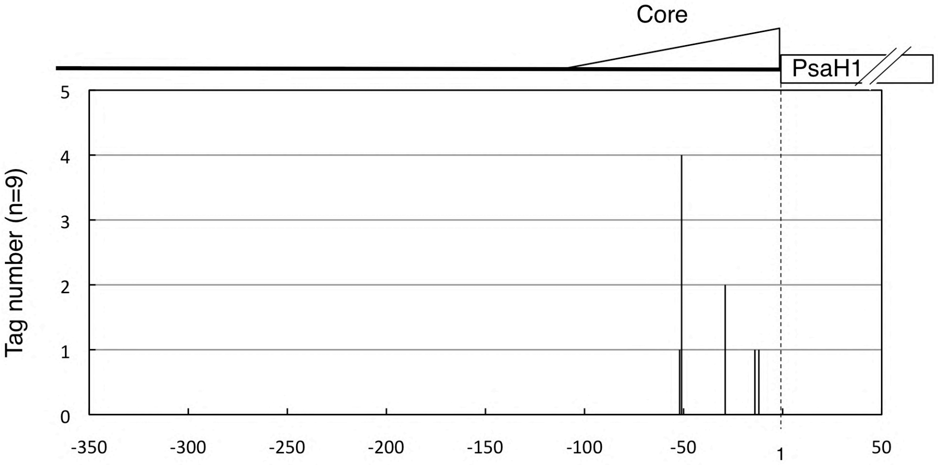
```
....|....| ....|....| ....|....| ....|....| ....|....|
       -190       -180       -170       -160       -150
GTGGATAGAG GAAAATAAGA GCCGTGTGGA ATCCGAAAGG CTGCGATTCT
....|....| ....|....| ....|....| ....|....| ....|....|
       -140       -130       -120       -110       -100

CTCTCACCGT TAAGAGACGG GTGTTCATTC CTTTTTCTGA TCGCTCAAAA
....|....| ....|....| ....|....| ....|....| ....|....|
        -90        -80        -70        -60        -50
TTATTCAACA TCATATTTTA AACTCTCTAC TAATCCACGG AATCGAATAA
....|....| ....|....| ....|....| ....|....| ....|....|
        -40        -30        -20        -10          0
TTGGCAAACT_GTTATTGCTT ATAAAATTTC ATTAGTTATA ATAATGTTTT
```

```
....|....| ....|....| ....|....| ....|....| ....|....|
       -190       -180       -170       -160       -150
GGATTTATGT GGGGATGGTC GCTTTTGCTA TCTCGAGTTG ACTTCCAAGT
....|....| ....|....| ....|....| ....|....| ....|....|
       -140       -130       -120       -110       -100
TAATCCTAAC CACATCACAT ACTTCTATTA TTCCATTTTC AACCTAATGT
....|....| ....|....| ....|....| ....|....| ....|....|
        -90        -80        -70        -60        -50
AAATAATTCA TATACATGAC AAATGCATTT AGAGAGAAGA AAAAACATCA
....|....| ....|....| ....|....| ....|....| ....|....|
        -40        -30        -20        -10          0
TTCATGACGA CAAAAACTTA GTCTATTGAA CTTTTCTAGC GTTTTCTCTT
```

Figure S2. Detailed sequence and TSS distributions of the YB41 and YB84 loci. Vertical bars on the sequence map show the distribution profile of TSSs, with the total sum of TSSs as 1.0. Bold letters in the nucleotide sequences indicate TSSs. Boxes indicate the sites of PCR primers used for RT-PCR.

intrinsic *psaH1* gene

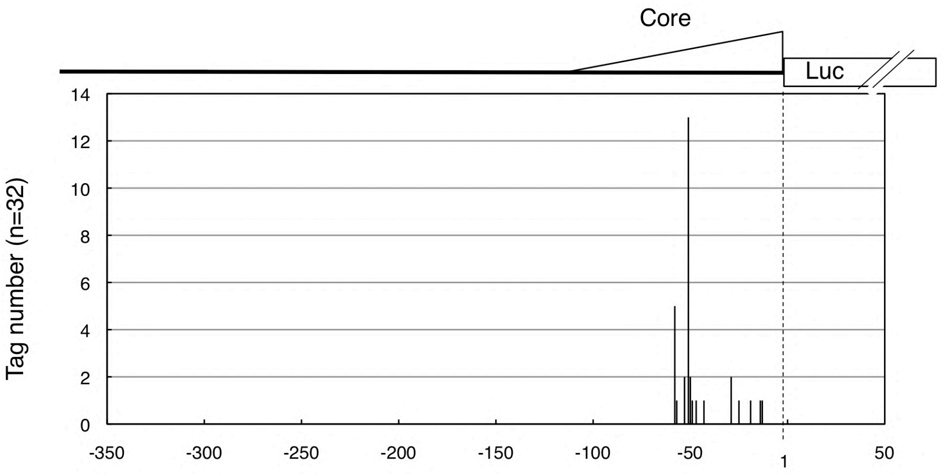


```
                                                                                ●●                        ●               ●  ●
-111   T GCCACGTGTC AGAAGAAGAT CAAAACTCGA CTCATTGTGA CCAAAGGCTA TTATTTCCTA CACAAAATTC CCACTCACCA CACACAACAA AAGAATAGTG ATCGAAGCTC   -1
                                                       (Core )

+1     ATGGCGTCTC TTGCAACCGT CGCCGCTGTG                                                                                          +30
                  PsaH1
```

Figure S3. TSS distribution profile of *psaH1* of *Arabidopsis thaliana*.
ATG in red indicates the translation start site (+1) of *psaH1*.

psaH1::LUC

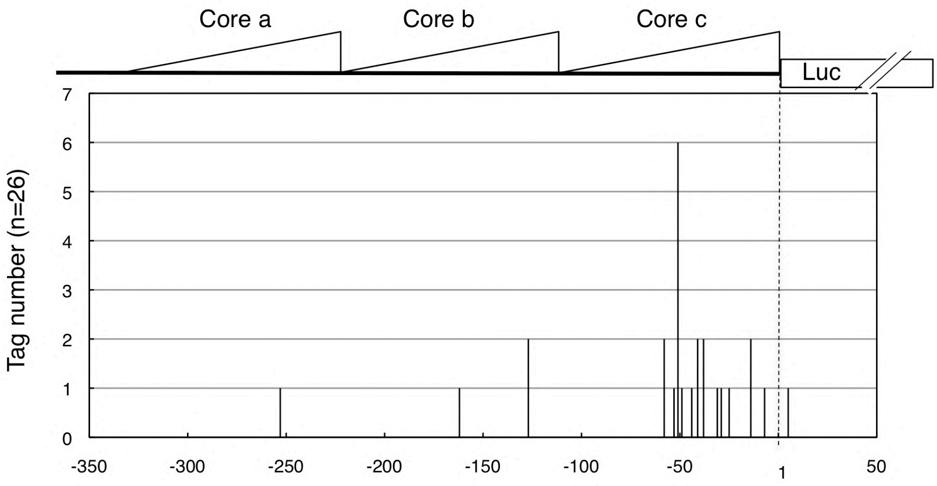


```
                                                                 ●●   ● ⬤ ●● ●   ●               ●   ●      ●    ●●
-111  T GCCACGTGTC AGAAGAAGAT CAAAACTCGA CTCATTGTGA CCAAAAGCTA TTATTTCCTA CACAAAATTC CCACTCACCA CACACAACAA AAGAATAGTG ATCGAAGCTC    -1
                                                                (Core )

 +1   ATGGCGTCTC TTGGGATCCA AATGGAAGAC                                                                                              +30
                Luc
```

Figure S4. TSS distribution profile of PsaH1::LUC. Whole promoter region plus translation initiation context of *psaH1* (–2000 to +12 relative to the ATG initiation codon) was translationally fused with the firefly luciferase (LUC) gene via a *Bam*HI linker. ATG initiation codons (red) of *psaH1* and LUC are in frame, but are separated by 18 bp.

$HC_{111}x3$::LUC



```
                                                                                   ●
-333  T GCCACGTGTC AGAAGAAGAT CAAAACTCGA CTCATTGTGA CCAAAAGCTA TTATTTCCTA CACAAAATTC CCACTCACCA CACACAACAA AAGAATAGTG ATCGAAGCTC  -223
                                                               (Core a)

                                                                        ●                                      ●
-222  T GCCACGTGTC AGAAGAAGAT CAAAACTCGA CTCATTGTGA CCAAAAGCTA TTATTTCCTA CACAAAATTC CCACTCACCA CACACAACAA AAGAATAGTG ATCGAAGCTC  -112
                                                               (Core b)

                                                                 ●    ● ⬤  ●       ●   ●      ●  ●   ●           ●       ●
-111  T GCCACGTGTC AGAAGAAGAT CAAAACTCGA CTCATTGTGA CCAAAAGCTA TTATTTCCTA CACAAAATTC CCACTCACCA CACACAACAA AAGAATAGTG ATCGAAGCTC  -1
                                                               (Core c)

           ●
+1    ATGGCGTCTC TTGGGATCCA AATGGAAGAC                                                                                           +30
      Luc
```

Figure S5. TSS distribution profile of $HC_{111}x3$::LUC. TSSs occur preferentially in segment c.