De novo activated transcription of newborn coding sequences is inheritable in the plant genome

Takayuki Hata[1,2], Naoto Takada[1], Chihiro Hayakawa[1], Mei Kazama[1], Tomohiro Uchikoba[3], Makoto Tachikawa[1], Mitsuhiro Matsuo[2], Soichirou Satoh[1,3], and Junichi Obokata[2*]

[1] Graduate School of Life and Environmental Sciences, Kyoto Prefectural University, Kyoto-shi, Kyoto, Japan

[2] Faculty of Agriculture, Setsunan University, Hirakata-shi, Osaka, Japan

[3] Faculty of Life and Environmental Sciences, Kyoto Prefectural University, Kyoto-shi, Kyoto, Japan

* Corresponding author

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

Takayuki Hata and Naoto Takada should be considered joint first author.

Running Head

Genetic behaviors of *de novo* transcription

Keywords

de novo transcriptional activation, artificial evolutionary experiment, promoter-trap screening, gene evolution, epigenome, chromatin remodeling, H2A.Z, *Arabidopsis thaliana*

ABSTRACT

The manner in which newborn genes become transcriptionally activated and fixed in the plant genome is poorly understood. To examine such processes of gene evolution, we performed an artificial evolutionary experiment in *Arabidopsis thaliana.* As a model of gene-birth events, we introduced a promoterless coding sequence of the firefly luciferase (*LUC*) gene and established 386 T2-generation transgenic lines. Among them, we determined the individual *LUC* insertion loci in 76 lines and found that one-third of them were transcribed *de novo* even in the intergenic or inherently unexpressed regions. In the transcribed lines, transcription-related epigenetic marks were detected across the newly activated transcribed regions. These results agreed with our previous findings in *A. thaliana* cultured cells under a similar experimental scheme. The comparison of the results of the T2-plant and cultured cell experiments revealed that the *de novo*-activated transcription caused by local chromatin remodelling was inheritable. During one-generation inheritance, it seems likely that the transcription activities of the *LUC* inserts trapped by the endogenous genes/transcripts became stronger, while those of *de novo* transcription in the intergenic/untranscribed regions became weaker. These findings may offer a clue for the elucidation of the mechanism via which newborn genes become transcriptionally activated and fixed in the plant genome.

INTRODUCTION

Genomes exhibit a steady state of the dynamic activity between the gain and loss of genes. Comparative functional genomics among closely related species revealed how genomes acquired such genetic novelty during their evolution, duplication–diversification, transposition, gene transfer or *de novo* origination (Kaessmann, 2010; Cardoso-Moreira and Long, 2012; McLysaght and Guerzoni, 2015; Van Oss and Carvunis, 2019). However, the studies of gene evolution reported to date have focused on the manner in which new protein-coding sequences arise; thus, the mechanism of acquisition of promoters by such newly originated coding sequences remains unknown.

A gene promoter is the region in which transcription is initiated and is a central component of gene regulation (Haberle and Stark, 2018; Andersson and Sandelin, 2020). In eukaryotes, promoter-specific sequences and epigenetic marks are well characterized (Haberle and Stark, 2018; Andersson and Sandelin, 2020). Importantly, to become functional, new genes must acquire a promoter during their evolution. To examine the mechanism via which new genes acquire their promoters in the plant genome, we previously carried out an artificial evolutionary experiment (Satoh and Hata *et al.*, 2020). To mimic a gene-birth event, we introduced exogenous promoterless coding sequences into *Arabidopsis thaliana* T87 cultured cells and analysed their transcriptional fates at the genome-wide level (Satoh and Hata *et al.*, 2020). Interestingly, we found that promoterless coding sequences became transcriptionally activated via two distinct mechanisms: (1) the so-called promoter trapping, in which integrated sequences capture the endogenous promoter activities of pre-existing genes/transcripts; or (2) *de novo* transcriptional activation, which occurs ubiquitously across the entire genome, and does so stochastically in about 30% of the integration events independent of the chromosomal locus (Satoh and Hata *et al.*, 2020). We speculated that the insertion of exogenous coding sequences might activate local chromatin remodelling to shape the promoter-like epigenetic landscape, resulting in such *de novo* transcriptional activation. Although the evidence of this type of epigenetic rewiring was not confirmed in the studies of cultured cells, we observed that promoter-like histone species were newly localized in the vicinity of the 5′ end of the inserted promoterless coding sequences under a similar experimental system in transgenic plants (Kudo, Matsuo, and Satoh *et al.*, 2020).

By providing a homogeneous and simple experimental system, cultured cells allowed us

to study the molecular mechanisms via which newly originated coding sequences acquire transcriptional competence (Satoh and Hata *et al.*, 2020; Hata *et al.*, 2020). However, as cultured cells experience only vegetative propagation, they are not suitable for the assessment of whether the *de novo* transcriptional activation is an inheritable phenomenon. Therefore, a genetic approach is necessary to understand the whole process of gene/promoter evolution, from the origination to the subsequent selection/adaptation steps.

In this study, we aimed to establish a model system to elucidate the mechanism via which newborn coding sequences acquire their promoters and become fixed as functional genes in the plant genome. We carried out a large-scale promoter-trap screening in the T2 generation of *A. thaliana* plants under an experimental scheme similar to that used in our previous study of cultured cells (Satoh and Hata *et al.*, 2020). By comparing the results obtained in plants with those of cultured cells, we concluded that *de novo* transcriptional activation together with chromatin remodelling is an inheritable phenomenon in the plant genome. After one generation, a sign of adaptation had already appeared: the transcriptional activities of introduced coding sequences trapped by endogenous genes/transcripts became much stronger, while those of the intergenic/untranscribed regions became much weaker. These findings may contribute to the elucidation of how newborn genes become transcriptionally activated and fixed in the plant genome at their early evolutionary stages.

RESULTS

Establishment of transgenic lines for large-scale promoter-trap screening in *A. thaliana*

To investigate the mechanism of promoter birth and their genetic behaviours beyond one generation, we performed a promoter-trap screening using *A. thaliana* plants under conditions that were essentially the same as those used in a previous study of cultured cells (Satoh and Hata *et al.*, 2020). Based on *Agrobacterium*-mediated transformation (Clough and Bent, 1998), we introduced the promoterless coding sequence of a firefly luciferase (*LUC*) gene into *A. thaliana* (Figure 1). Each *LUC* gene was tagged by distinct short random sequences called “barcodes” (Figure 1), which were used as identification codes for individual transgenic lines in the subsequent *in silico* analysis. In this study, to analyse the transgenic lines without the selection bias caused by *LUC* gene function, we screened the T1 seeds against the kanamycin

(Km) resistance of the co-transformed expression cassette, rather than the strength of the LUC luminescence (Figure 1). Finally, we established a T2 generation of 386 individual transgenic lines (termed T2-plants hereafter).

Genetic behaviours of *de novo*-activated transcription in *A. thaliana*

To identify the insertion loci and corresponding transcription levels of the individual *LUC* genes, we performed a massively parallel reporter assay based on the TRIP method (Akhtar *et al.*, 2013; Satoh and Hata *et al.*, 2020).

First, three seeds per individual T2-plant were grown using the non-selective condition and seedlings were harvested as a mixed sample (Figure 1). Because the T2 generation is not homozygous, theoretically, one-fourth of T2 seeds were expected to be wild type (WT). However, as we grew three seedlings per line, no less than 98% of T2-plants (380/386) were expected to be recovered. In the TRIP method, individual transgenic lines are identified via *in silico* analysis based on the tagged barcode sequence of the reporter construct, as a molecular identifier (Akhtar *et al.*, 2013). Specifically, we extracted DNA and RNA from the mixed samples and prepared the next-generation sequencing (NGS) library. For the determination of the insertion locus of each promoterless *LUC* gene, we performed inverse PCR followed by NGS, to read out the *LUC*–genome junction and barcode sequence. The transcription level of each *LUC* gene was determined utilizing NGS by counting the molecular abundance of each barcode in the RNA sample. Finally, each *LUC* gene insertion locus and transcription level was assigned according to its barcode sequences. Note that T-DNAs are often inserted tandemly or with a large deletion on the reporter gene (De Buck *et al.*, 2009). We carefully omitted such lines from further analysis because we could not determine their insertion loci uniquely. Based on this scheme, we determined individual insertion loci and corresponding transcription levels in 76 T2-plants (Figures 1 and 2a). To confirm the results of the *in silico* analysis, we verified individual barcode sequences and insertion loci in randomly chosen T2-plants using Sanger sequencing and locus-specific PCR (Figure S1). As shown in Figure 2a, promoterless *LUC* genes were inserted throughout the *A. thaliana* genome with low frequency in pericentromeric regions, which agreed with the reported preference of *Agrobacterium* T-DNA integration (Kim, Veena and Gelvin, 2007; Satoh and Hata *et al.*, 2020). One-third of the 76 *LUC* genes (n = 27) were transcribed (Figure 2b). To examine further the manner in which these promoterless *LUC* genes became transcribed, we classified them according to their insertion types: an endogenous genic region with the sense

(Genic Sense) and antisense (Genic AS) orientation, and the remaining intergenic regions (Intergenic). Based on this classification, the Genic Sense, Genic AS, and Intergenic types accounted for 26.3%, 21.1%, and 52.6% of the transcribed *LUC* genes, respectively (Figure 2c). Because the genic and intergenic regions of the *A. thaliana* genome have almost the same length (Berardini *et al.*, 2015), these results suggest that our established T2-plants exhibited no insertion-locus preference.

Previously, based on the cultured cell experiment, we found that exogenously inserted promoterless genes became transcriptionally activated in two distinct types: promoter trapping and *de novo* transcriptional activation (Satoh and Hata *et al.*, 2020; Hata *et al.*, 2020). To examine whether similar transcriptional activation mechanisms occurred in our T2-plants, the abundance of the transcribed fraction was compared between the corresponding insertion types of T2-plants and cultured cells (Figure 2d and e). As shown in Figure 2d, ~30% of the promoterless *LUC* genes were transcribed similarly in T2-plants and cultured cells (Figure 2d). Their transcription levels ranged from the 10^1 to the 10^7 orders, with a peak at 10^4 (Figure 2e). Regarding the three insertion types, the abundances of transcribed *LUC* genes were almost the same in T2-plants and cultured cells, with the exception of the Genic Sense type, in which the transcribed fraction was much greater in the T2-plants (Figure 2d). In both T2-plants and cultured cells, the Genic Sense type showed the highest transcriptional activity among the three insertion types, with 10^5 as a peak (Figure 2e).

To highlight the differences between the cultured cells and T2-plants, we divided the transcribed *LUC* lines into two fractions: that with a lower transcription level (10^1–10^4) and that with a higher transcription level (10^5–10^7). As shown in Figure 2f, the relative abundances of the higher and lower fractions in T2-plants exhibited a greater bipolarization than they did in cultured cells; *LUC* transcription became much stronger in the Genic Sense and AS types, while it became much weaker in the Intergenic type (Figure 2f). As the Genic Sense and AS types were transcribed presumably by trapping the transcriptional activities of endogenous genes (Hata *et al.*, 2020), these features suggested that gene-trapping events are more prone to occur in plants vs. cultured cells. Conversely, a type of transcriptional repression might have occurred on the Intergenic type in the T2 generation.

Taken together, these results suggest that the transcriptional behaviours of the promoterless *LUC* genes are remarkably similar between the T2-plants and the vegetatively

grown cultured cells (Figure 2d and e). Therefore, it is likely that *de novo* transcriptional activation events are not specific to the vegetatively growing cultured cells; rather, they seem to be an inheritable phenomenon through a plant's generation.

Comparison of *LUC* transcription with inherent transcriptional status

Next, we investigated whether the transcription of *LUC* genes in the T2-plants was caused by the trapping of endogenous transcripts in the WT genome. First, we prepared a dataset of the transcribed region in the WT genome using the publicly available RNA-seq data of *A. thaliana*, which were obtained using growth conditions similar to those used here. The dataset covered 58.5% of the *A. thaliana* genome and 70.4% (19,308/27,416) of the annotated protein-coding genes. We classified the *LUC* inserts of T2-plants and T87 cells (Satoh and Hata *et al.*, 2020) as follows: a *LUC* gene was (i) transcribed or (ii) untranscribed in the inherently transcribed region, or a *LUC* gene was (iii) transcribed or (iv) untranscribed in the inherently untranscribed region. The insert classification breakdown in T2-plants and T87 cells (Satoh and Hata *et al.*, 2020) was as follows: (i) 14.5% and 6.9%, (ii) 7.9% and 7.4%, (iii) 21.1% and 23.1% and (iv) 56.6% and 62.5%, respectively (Figure 3a).

To clarify the differences between the T2-plants and cultured cells, next we normalized each fraction of the (i)~(iv) types to the subtotal of (iii) and (iv) types (total *LUC* genes inserted in the inherently untranscribed regions) which was set as 100% (Figure 3b). As shown in Figure 3b, the percentage of type (i) in the T2-plants was approximately 2-fold that of the cultured cells, whereas the remaining types were almost similar between T2-plants and cultured cells (Figure 3b). This similarity indicated the generality of the *de novo* transcriptional activation that occurs in the plant genome. Conversely, the over-representation of type (i) in T2-plants indicated that the insertion preference of *LUC* genes into inherently transcribed loci was stronger in T2-plants compared with cultured cells (Figure 3a and b). As most of the transcribed regions in the WT genome overlapped with the annotated protein-coding genes, this result agreed with the larger frequency of transcribed inserts in genic regions observed in T2-plants (Figure 2d). Although the cause of this preference is unknown, this feature suggests that the plants are more prone to exhibit transcriptional activation of exogenous genes via the trapping of endogenous transcripts compared with cultured cells.

Generally, in promoter-trapping experiments, the expressed reporter genes are expected to reflect the activities of trapped endogenous promoters (Springer, 2000). However, we

previously found that the transcription levels of promoterless *LUC* genes did not reflect those of their inherent endogenous transcripts in the experiment that used cultured cells (Satoh and Hata *et al.*, 2020). To confirm whether this feature was specific to the vegetatively growing cultured cells, we compared the transcription levels between T2-plants and their corresponding regions in the WT genome. We found that there was no correlation between them (Figure 3c). Thus, the observation that the trapping type of newly activated transcription events did not retain their inherent transcriptional status, at least in our experimental conditions, appeared to be a general feature of the plants and cultured cells. As insertions of the fragments were likely to disrupt the transcriptional activities of given loci, this result suggests two possibilities: (1) the original transcriptional activities had not yet been recovered in the vegetative propagation or within one generation; or (2) the transcriptional activities were overwritten by the *de novo*-activated transcription.

Epigenetic rewiring occurred across the newly activated transcribed regions

Eukaryotic transcription is regulated by the control of the localization of transcription-related epigenetic marks (Haberle and Stark, 2018; Andersson and Sandelin, 2020). Therefore, next we focused on the epigenetic status around *LUC* inserts to examine whether the transcribed T2-plants were regulated by such epigenetic marks. First, we screened T2-plants according to the following criteria: existence of transcription evidence in the TRIP experiment, and an unlikeliness to be affected by the pre-existing promoters or transcription units. Based on these criteria, we finally selected two lines: T2:161 and T2:205 (Figure 4a and b). The T2:161 line was classified as a Genic AS type in which the *LUC* insert was found in the opposite strand of an endogenous gene (AT3G23750) (Figure 4a). In the T2:205 line, the *LUC* insert was located in an intergenic region, in which an endogenous gene (AT5G01110) was detected downstream of the *LUC* insert on the opposite strand (Figure 4b). Inserted promoterless *LUC* genes were transcribed in both lines (Figure 4c, and Figure S2), whereas inherent transcripts were not observed, at least in WT RNA-seq data. For these two lines, we scanned the localization of epigenetic marks around the *LUC* insertion loci and compared them with those obtained from the WT genome. In this study, we analysed three transcription-related epigenetic marks: methylated cytosine (mC), lysine 36 tri-methylation of histone H3 (H3K36me3), and the histone variant H2A.Z. In the WT genome, enrichments of mC and H3K36me3 were observed within the gene bodies of AT3G23750 and AT5G01110, respectively (Figure 3d and e), which agreed with the general properties of these epigenetic marks (Jones, 2012; Wagner and Carpenter, 2012).

However, in the T2-plants, these two epigenetic marks were not found within the *LUC* gene bodies (Figure 3d and e). Although weak signals were observed 200 bp upstream from the *LUC* insert in the T2:161 line (Figure 3d), they reflected the epigenetic marks of the WT allele in the T2-plant, because these plants were not homozygous. Conversely, the localization patterns of the H2A.Z variant were clearly different from those of the other two epigenetic marks (Figure 3d and e). Both lines showed significant enrichments of H2A.Z throughout the *LUC* gene bodies, while there were almost no H2A.Z signals in the corresponding regions in the WT genome (Figure 3d and e). Although H2A.Z is a marker histone for the promoter region, it also appears in the gene bodies of genes with low expression (Lashgari *et al.*, 2017; Gómez-Zambrano, Merini and Calonje, 2019; Lei and Frederic, 2020). In addition, mC and H3K36me3 were reportedly deposited within a gene body in a transcription-coupled manner (Teissandier and Bourc'his, 2017), which would be undetectable in the low-expressed genes (Cermakova *et al.*, 2019). Thus, these distribution patterns of epigenetic marks in the T2-plants were plausible because the transcriptional strength of these two lines was low compared with that of the constitutive genes (Figure 4c, and Figure S2).

In the T2:161 line, H2A.Z was newly localized 200 bp upstream from the *LUC* insert (Figure 4d), which suggests that epigenetic rewiring occurred even outside of the *LUC* insert. We hypothesized that H2A.Z is localized throughout the transcribed region of the *LUC* insert. To confirm this hypothesis, next we analysed the transcription start site (TSS) of *LUC* inserts. However, it was challenging to determine the TSSs of T2-plants using general methods (Maruyama and Sugano, 1994; Carninci *et al.*, 1996) because of the low transcription levels of these plants. The template-switching method has the advantage of yielding full-length cDNAs from low-input RNA (Salimullah *et al.*, 2011). In this study, we applied inverse PCR to this template-switching method to specifically amplify the full-length cDNAs of *LUC* genes. Based on this method, we analysed TSS distribution in T2-plants. Unfortunately, the transcription level of the T2:205 line was too low to obtain any TSS signals. Conversely, in the T2:161 line, a TSS was found ~1.1 kb upstream of the *LUC* insertion locus (Figure 5a, and Figure S3). Sanger sequencing revealed that this transcript was spliced (Figure 5a, and Figure S3). We reanalysed the distribution profiles of H3K36me3 and H2A.Z around the determined TSS (Figure 5b). There was no significant enrichment of H3K36me3 around the *LUC*-TSS, as the enrichment levels were almost the same among the transgenic plants and the WT genome (Figure 5b, upper panel). In contrast, we observed that H2A.Z was newly localized starting from the *LUC*-TSS,

whereas H2A.Z was not observed in the corresponding locus in the WT genome (Figure 5b, lower panel).

Overall, the epigenetic and TSS analyses revealed that exogenously inserted promoterless genes acquired a brand-new epigenetic status, and that such epigenetic rewiring occurred throughout the newly activated transcription unit. In addition, this epigenetic rewiring might also have been responsible for the transcriptional behaviour of the trapping type of *LUC* transcription (Figure 3c): *de novo*-activated transcription events caused by the epigenetic rewiring might overwrite their inherent transcriptional status.

DISCUSSION

In this study, based on the large-scale promoter-trap screening of *A. thaliana* plants, we demonstrated the genetic behaviour of the newly activated transcription of exogenous genes. A comparison with the results of a previous study using cultured cells (Satoh and Hata *et al.*, 2020) showed that *de novo* transcriptional activation is an inheritable phenomenon of the plant genome (Figures 1–3). We also demonstrated that epigenetic rewiring occurred across all transcribed regions of the inserted coding sequences (Figures 4 and 5), which probably regulated *de novo*-activated transcription by overwriting the inherent transcriptional status.

In the T2:161 line, the TSS was located on the 3′ end of an endogenous gene (AT3G23750), where no detectable transcripts existed in the WT genome (Figure 5a). It is plausible to propose that this was caused by activating (rather than trapping) a cryptic antisense transcript of the given locus (Hata *et al.*, 2020). Conversely, we speculated that the T2:205 line may be transcribed from a *de novo*-activated TSS located in the proximal intergenic region, although we could not identify this TSS in this study. This speculation was based on a previous finding from the cultured cell experiment: *de novo* TSS occurs about 100 bp upstream of the inserted coding sequences in the intergenic region (Hata *et al.*, 2020). The localization pattern of H2A.Z in the T2:205 line agreed with this prediction, as the H2A.Z signal clearly dropped to almost zero at 200 bp upstream of the *LUC* insert (Figure 4e).

Generally, in promoter-trap screening, transgenic lines are screened based on the expression of the inserted promoterless reporter genes (Springer, 2000). In contrast, we did not carry out the screening of T2-plants according to the expression of *LUC* genes; rather, we

selected them according to the activity of the co-transformed Km-resistance gene (Figure 1). This selection method enabled the isolation of lines without the selection bias that was caused by the transcription levels of the *LUC* genes. However, we found differences between the results of plants and cultured cells, despite the similar experimental conditions used in the two experiments. For instance, compared with the cultured cells, plants were more prone to be transcriptionally activated by the trapping of endogenous gene/transcripts (Figures 2d and 3b), and the transcriptional strength of such activated transcription tended to be bipolarized to lower and higher transcription levels according to the insertion type (Figure 2f). How can these features of T2-plants be explained? Although transgenic cultured cells were regarded as the T1 generation, we used the T2 generation of transgenic plants in this study. Plants require a greater number of genes than do cultured cells during this one-cycle generation, because plants experience germination, development, differentiation and sexual reproduction, while the cultured cells are only in the state of vegetative propagation in a constant culture condition. Gene-insertion events might cause lethal effects on a certain population of transgenic plants by disrupting various genes that are essential for their growth over the life cycle (Meinke, 2020). Therefore, although we assumed that the T2-plant lines were established under a non-selective condition for *LUC* activity, the population might be distorted through a generation. Km-based selection might also affect the T2-plant population; T-DNA insertion sometimes fails to confer Km resistance and causes embryonic lethality (Errampalli *et al.*, 1991; Francis and Spiker, 2005). In addition, under the selective condition, T-DNAs tended to be inserted in open-chromatin and hypomethylated regions (Shilo *et al.*, 2017). Thus, Km-based selection might enrich transgenic lines in which inserts were located in the transcriptionally permissive regions where the Km-resistance genes could function sufficiently. Overall, the transcriptional fates of promoterless *LUC* inserts were likely to be affected by the experienced life stages and selective conditions during the establishment of transgenic plants. Hence, to grasp the extent to which inserted promoterless coding genes actually become transcribed in plants, alternative experimental strategies are needed; for example, selection-free transformation or the use of a binary vector system to introduce reporter and selection marker genes independently (Komari *et al.*, 1996).

To date, studies of the evolutionary processes via which genetic novelty emerges were mainly led by comparative genomics (Carvunis *et al.*, 2012; Zhao *et al.*, 2014; Li *et al.*, 2016; Li, Lenhard and Luscombe, 2018; Zhang *et al.*, 2019). However, because such genomics approaches are established based solely on the evolutionary winners, the resultant scenario

lacks perspective from the great majority of evolutionary losers. The resolution depends on the divergence time, ranging from millions to billions of years. Conversely, our artificial evolutionary approach sheds light even on evolutionary losers within a much shorter timescale. Specifically, as the *LUC* genes used in this study are not profitable for plants, most of them would presumably become silenced or pseudogenized, while a few of them might occasionally be retained. How many generations and populations are needed to reach such endings? Our approach based on the use of plants will reveal the types of genetic/epigenetic variations that become winners/losers, thus enabling the tracing of the fates of newly activated transcripts in the population over the generations. In contrast, the cultured cells will be a useful model to investigate the molecular mechanisms underlying promoter birth, thus providing a homogeneous and simple experimental system.

It is intriguing to utilize a stress-tolerance/inducible gene as a promoterless reporter gene in our artificial evolutionary experiment. This would be a useful model to investigate the manner in which newborn genes adapt and evolve against exposed stress or selective environments. It is also interesting to try such experiments among different developmental phases and tissues. For example, the promoterless genes might be more prone to be transcribed in the pollen, where new genes often arise because of the transcriptionally permissive status caused by the accessible chromatin configuration (Wu *et al.*, 2014). Such an approach allows the investigation of gene evolution in multicellular organisms, thus providing insights into how newborn genes become integrated into pre-existing spatio-temporal genetic networks.

In conclusion, our artificial evolutionary experiment provided insight into the initial genetic behaviour of newly activated transcription in the plant genome. We showed that the *de novo*-activated transcription caused by the local chromatin remodelling was inheritable. To evaluate the contribution of this phenomenon to the plant genome evolution, examination of the genetic behaviour of the *de novo* transcribed genes over an increasing number of generations with/without selective pressures will provide further clues regarding this phenomenon.

EXPERIMENTAL PROCEDURES

Plant materials and transformation

Arabidopsis thaliana (ecotype; Col-0) plants were grown at 23°C with continuous illumination

(20–50 μmol m^{-2} s^{-1}). Ti-plasmid libraries containing a promoterless *LUC*-coding sequence, a 12 bp random sequence ("barcode"), a nos-terminator and an expression cassette of a kanamycin (Km)-resistance gene between the left (LB) and right (RB) borders of the T-DNA were constructed according to a published method (Satoh and Hata *et al.*, 2020). *Agrobacterium*-mediated transformation of *A. thaliana* was performed according to the floral-dip method (Clough and Bent, 1998). Transformed seeds were selected on Murashige and Skoog (MS) medium [1× strength of MS plant salt mixture (Nihon Pharmaceutical), 1% sucrose, 0.05% MES, 0.8% agar, pH 5.7] supplemented with 25 μg ml^{-1} of Km. The screened 386 individual Km-resistant T1 seedlings were grown at 23°C with continuous illumination (20–50 μmol m^{-2} s^{-1}). The seeds of individual T2-generation plants were harvested and subjected to further experiments.

Sequence library preparation and data analysis

Three seeds of individual T2-plants were stratified at 4°C in the dark for 2 days, then grown on MS medium [half-strength MS medium including vitamins (Duchefa Biochemie), 1% sucrose, 0.8% agar, pH 5.7] at 23°C with continuous illumination (40–60 μmol m^{-2} s^{-1}) for 10 days. All seedlings were harvested and ground under liquid nitrogen to a fine powder, for thorough mixing. DNA and RNA were extracted using a DNeasy Plant Mini Kit (QIAGEN) and RNeasy Plant Mini Kit (QIAGEN), respectively. Next-generation sequencing (NGS) libraries for determining insertion loci and transcription levels of promoterless *LUC* genes were prepared and sequenced according to a published method (Satoh and Hata *et al.*, 2020). All primers used in this study are listed in Table S1.

For the determination of insertion loci and their barcode-labelled sequences, NGS reads were first processed, before mapping to the genome according to a published method (Hata *et al.*, 2020), with the following modifications. Specifically, NGS reads were aligned to the T-DNA vector sequence using Blastn (version: 2.4.0+) (Camacho *et al.*, 2009), to obtain individual flanking sequences from the *LUC* insert and barcode. The obtained flanking sequences were mapped on the TAIR10 version of the *A. thaliana* genome using bowtie (Langmead *et al.*, 2009) allowing three mismatches. Precise locus–barcode pairs were determined according to the following criteria: (1) at least two read counts; (2) the read count of the most frequent locus–barcode pair accounted for ≥60% of them, including their PCR/sequencing artefacts; and (3) exclusion from subsequent analysis of two or more distinct *LUC* inserts with the same barcode

sequences. The transcription level of each T2-plant was analysed according to a published method (Satoh and Hata *et al.*, 2020). Subsequently, individual *LUC* loci and transcription levels were integrated based on their barcode sequences. The insertion loci of T2-plants were classified according to the TAIR10 version of the genomic annotation of *A. thaliana* under the following classification: genomic regions where annotated protein-coding genes were defined as 'Genic' regions, whereas the remainder of the genome was classified as 'Intergenic'. The insertion strand of *LUC* genes was considered.

Validation of *LUC* insertion loci and barcode sequences

Randomly chosen T2-plants were stratified at 4°C in the dark for 2 days, then grown on MS medium supplemented with 25 μg ml^{-1} of Km at 23°C with continuous illumination (20–30 μmol m^{-2} s^{-1}) for 10 days. Km-resistant seedlings were harvested and subjected to DNA extraction. Four types of PCR were performed to amplify the barcode region, the RB–genome junction, the LB–genome junction and the T-DNA insert, respectively. The PCR products obtained were then analysed by agarose gel electrophoresis and Sanger sequencing, for validation of the insertion locus and barcode sequence, respectively.

Comparison with WT transcriptome data

RNA-seq data of WT *A. thaliana* (col-0) plants were retrieved from the *NCBI Short-Read Archive* under accessions SRR6388204, SRR6388205 and SRR770510. The sequencing reads were subjected to adapter trimming and quality trimming, followed by mapping to the *A. thaliana* genome (TAIR10) using STAR (v2.5.3) (Dobin *et al.*, 2013) with the following parameters: *STAR –alignIntronMax 6000 –outSAMstrandField intronMotif –two passMode Basic*. Transcribed regions and their transcription levels (in fragments per kilobase of exon per million reads mapped (FPKM)) were analysed using StringTie (v2.1.4) (Pertea *et al.*, 2015). Subsequently, the transcription level of each T2-plant was compared with the FPKM of the inherent transcribed region in the WT genome. In the case of inherent transcripts with multiple isoforms, each FPKM was summed up.

Chromatin immunoprecipitation (ChIP) and MBD immunoprecipitation (MBDIP) analysis

The T2:161 and T2:205 lines were stratified at 4°C in the dark for 3 days, then grown on MS medium [half-strength MS medium including vitamins (Duchefa Biochemie), 1% sucrose, 0.8% agar, pH 5.7] supplemented with 15 μg ml^{-1} of Km at 23°C with continuous illumination (20–30

μmol m^{-2} s^{-1}) for 8 days. Km-resistant seedlings were harvested and subjected to ChIP and MBDIP analysis. For control experiments, transgenic *A. thaliana* harbouring an expression cassette of the Km-resistance gene without the *LUC* reporter gene (termed WT in Figures 4d and e, and 5b) were prepared and grown under the same condition as that used for T2-plants. ChIP and MBDIP were performed according to a published method (Kudo, Matsuo, and Satoh *et al.*, 2020; Satoh and Hata *et al.*, 2020; and Hata *et al.*, 2020), with the following modifications. For the ChIP assay, ~10 ng of solubilized chromatin (median, 200 bp) and antibodies (2.4 μg of an anti-H2A.Z antibody (Kudo, Matsuo, and Satoh *et al.*, 2020) and 2.0 μg of an anti-H3K36me3 antibody (Abcam: ab9050), respectively) were used for each experiment. For the MBDIP assay, the methylated DNA fraction (mC) was collected from 1.0 μg of sheared DNA (median, 200 bp) using an EpiXplore Methylated DNA Enrichment Kit (Clontech) according to the manufacturer's instructions. Successful enrichment of ChIPed DNA and mC was validated by quantitative PCR (qPCR) in the control sites (Table S1) according to Deal *et al*. (Deal *et al.*, 2007) for H2A.Z, to Yang *et al.* (Yang, Howard and Dean, 2014) for H3K36me3 and to Erdmann et al. (Erdmann *et al.*, 2014) for mC. In both T2-plants and WT, relative enrichments of H2A.Z, H3K36me3 and mC around the *LUC* insertion loci were calculated based on the enrichment of the control sites, which was set as 100%, respectively.

Expression and TSS analysis

The T2:161 and T2:205 lines were grown and harvested under the same condition as that used for the ChIP experiments. Total RNA was isolated using an RNeasy Plant Mini Kit followed by DNase I treatment. For expression analysis, cDNA was synthesized from 5.0 μg of the total RNA using an oligo dT_{20} primer and Super Script III Reverse Transcriptase (Thermo Fisher Scientific). The transcription level of the *LUC* gene was normalized to that of the ubiquitin gene (*UBQ10*: AT4G05320).

LUC-TSS was analysed according to a published method (Plessy *et al.*, 2010; Salimullah *et al.*, 2011), with the following modifications. Specifically, polyadenylated RNA was extracted using a Dynabeads mRNA Purification Kit (Invitrogen) according to the manufacturer's protocol. Polyadenylated RNA (1.0 μg) was used for reverse-transcription and template-switching reactions. During these reactions, SgfI sites were added at both ends of the full-length cDNA by the primer used for reverse transcription and the template-switching oligo. The full-length cDNAs obtained were then digested completely by SgfI. Subsequently, digested cDNAs were

circularized and subjected to inverse PCR to specifically amplify *LUC*-containing cDNAs. The resulting nested PCR products were analysed by Sanger sequencing.

ACKNOWLEDGEMENTS

We thank Moyuru Shirasu for his help in maintaining the transgenic *A. thaliana* plants. This work was supported by Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Number: 26660008.

SUPPORTING INFORMATION

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

Figure S1. Validation of *LUC* insertion loci.

Figure S2. Expression analysis of T2-plants.

Figure S3. Sequence alignment of TSS of T2:161 line.

Table S1. Primer list.

REFERENCES

Akhtar, W., de Jong, J., Pindyurin, A. V., Pagie, L., Meuleman, W., de Ridder, J., Berns, A., Wessels, L. F., van Lohuizen, M. and van Steensel, B. (2013) Chromatin position effects assayed by thousands of reporters integrated in parallel. *Cell,* **154**, 914–927. https://doi.org/10.1016/j.cell.2013.07.018

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

Berardini, T. Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E. and Huala, E. (2015) The Arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome. *Genesis,* **53**, 474–485. https://doi.org/10.1002/dvg.22877

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and

Madden, T. L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics,* **10**, 421. https://doi.org/10.1186/1471-2105-10-421

Cardoso-Moreira, M. and Long, M. (2012) The origin and evolution of new genes. *Methods Mol Biol,* **856**, 161–86. https://doi.org/10.1007/978-1-61779-585-5_7

Carninci, P., Kvam, C., Kitamura, A., Ohsumi, T., Okazaki, Y., Itoh, M., Kamiya, M., Shibata, K., Sasaki, N., Izawa, M., Muramatsu, M., Hayashizaki, Y. and Schneider, C. (1996) High-efficiency full-length cDNA cloning by biotinylated CAP trapper. *Genomics,* **37**, 327–336. https://doi.org/10.1006/geno.1996.0567

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

Cermakova, K., Smith, E., Veverka, V. and Hodges, H. (2019) Dynamics of transcription-dependent H3K36me3 marking by the SETD2:ISW1:SPT6 ternary complex. *bioRxiv*. PPR78974 https://doi.org/10.1101/636084

Clough, S. J. and Bent, A. F. (1998) Floral dip: a simplified method for Agrobacterium-mediated transformation of Arabidopsis thaliana. *Plant J,* **16**, 735–743. https://doi.org/10.1046/j.1365-313x.1998.00343.x

De Buck, S., Podevin, N., Nolf, J., Jacobs, A. and Depicker, A. (2009) The T-DNA integration pattern in Arabidopsis transformants is highly determined by the transformed target cell. *Plant J,* **60**, 134–145. https://doi.org/10.1111/j.1365-313X.2009.03942.x

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

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T. R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics,* **29**, 15–21. https://doi.org/10.1093/bioinformatics/bts635

Erdmann, R. M., Souza, A. L., Clish, C. B. and Gehring, M. (2014) 5-hydroxymethylcytosine is not present in appreciable quantities in Arabidopsis DNA. *G3 (Bethesda),* **5**, 1–8. https://doi.org/10.1534/g3.114.014670

Errampalli, D., Patton, D., Castle, L., Mickelson, L., Hansen, K., Schnall, J., Feldmann, K. and Meinke, D. (1991) Embryonic Lethals and T-DNA Insertional Mutagenesis in Arabidopsis. *Plant Cell,* **3**, 149–157. https://doi.org/10.1105/tpc.3.2.149

Francis, K. E. and Spiker, S. (2005) Identification of Arabidopsis thaliana transformants without selection reveals a high occurrence of silenced T-DNA integrations. *Plant J,* **41**, 464–477. https://doi.org/10.1111/j.1365-313X.2004.02312.x

Gómez-Zambrano, Á., Merini, W. and Calonje, M. (2019) The repressive role of Arabidopsis H2A.Z in transcriptional regulation depends on AtBMI1 activity. *Nat Commun,* **10**, 2828. https://doi.org/10.1038/s41467-019-10773-1

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

Hata, T., Satoh, S., Takada, N., Matsuo, M., Obokata, J. (2020) Kozak-Sequence plays a Negative Regulator for *de novo* Transcription Initiation of Newborn Coding Sequences in the Plant Genome. (submitted in Biorxiv simulataneously with this manuscript)

Jones, P. A. (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet,* **13**, 484–492. https://doi.org/10.1038/nrg3230

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

Kim, S. I., Veena and Gelvin, S. B. (2007) Genome-wide analysis of Agrobacterium T-DNA integration sites in the Arabidopsis genome generated under non-selective conditions. *Plant J,* **51**, 779–791. https://doi.org/10.1111/j.1365-313X.2007.03183.x

Komari, T., Hiei, Y., Saito, Y., Murai, N. and Kumashiro, T. (1996) Vectors carrying two separate T-DNAs for co-transformation of higher plants mediated by Agrobacterium tumefaciens and segregation of transformants free from selection markers. *Plant J,* **10**, 165–174. https://doi.org/10.1046/j.1365-313x.1996.10010165.x

Kudo, H., Matsuo, M., Satoh, S., Hachisu, R., Nakamura, M., Yamamoto, Y., Y., Hata, T., Kimura. H., Matsui, M., and Obokata, J. (2020) Cryptic promoter activation occurs by at least two different mechanisms in the *Arabidopsis* genome. (submitted in Biorxiv simulataneously with this manuscript)

Langmead, B., Trapnell, C., Pop, M. and Salzberg, S. L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol,* **10**, R25. https://doi.org/10.1186/gb-2009-10-3-r25

Lashgari, A., Millau, J. F., Jacques, P. and Gaudreau, L. (2017) Global inhibition of transcription causes an increase in histone H2A.Z incorporation within gene bodies. *Nucleic Acids Res,* **45**, 12715–12722. https://doi.org/10.1093/nar/gkx879

Lei, B. and Frederic, B. (2020) H2A Variants in Arabidopsis: Versatile Regulators of Genome Activity. *Plant Communications,* **1**, 100015. https://doi.org/https://doi.org/10.1016/j.xplc.2019.100015

Li, C., Lenhard, B. and Luscombe, N. M. (2018) Integrated analysis sheds light on evolutionary trajectories of young transcription start sites in the human genome. *Genome Res,* **28**, 676–688. https://doi.org/10.1101/gr.231449.117

Li, Z. W., Chen, X., Wu, Q., Hagmann, J., Han, T. S., Zou, Y. P., Ge, S. and Guo, Y. L. (2016) On the Origin of De Novo Genes in Arabidopsis thaliana Populations. *Genome Biol Evol,* **8**, 2190–2202. https://doi.org/10.1093/gbe/evw164

Maruyama, K. and Sugano, S. (1994) Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. *Gene,* **138**, 171–174. https://doi.org/10.1016/0378-1119(94)90802-8

McLysaght, A. and Guerzoni, D. (2015) New genes from non-coding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation. *Philos Trans R Soc Lond B Biol Sci,* **370**, 20140332. https://doi.org/10.1098/rstb.2014.0332

Meinke, D. W. (2020) Genome-wide identification of EMBRYO-DEFECTIVE (EMB) genes required for growth and development in Arabidopsis. *New Phytol,* **226**, 306–325. https://doi.org/10.1111/nph.16071

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T. and Salzberg, S. L. (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol,* **33**, 290–295. https://doi.org/10.1038/nbt.3122

Plessy, C., Bertin, N., Takahashi, H., Simone, R., Salimullah, M., Lassmann, T., Vitezic, M., Severin, J., Olivarius, S., Lazarevic, D., Hornig, N., Orlando, V., Bell, I., Gao, H., Dumais, J., Kapranov, P., Wang, H., Davis, C. A., Gingeras, T. R., Kawai, J., Daub, C. O., Hayashizaki, Y., Gustincich, S. and Carninci, P. (2010) Linking promoters to functional transcripts in small samples with nanoCAGE and CAGEscan. *Nat Methods,* **7**, 528–534. https://doi.org/10.1038/nmeth.1470

Salimullah, M., Sakai, M., Mizuho, S., Plessy, C. and Carninci, P. (2011) NanoCAGE: a high-resolution technique to discover and interrogate cell transcriptomes. *Cold Spring Harb Protoc,* **2011**, pdb.prot5559. https://doi.org/10.1101/pdb.prot5559

Satoh, S., Hata, T., Takada, N., Tachikawa, M., Matsuo, M., Kushnir, S., and Obokata, J. (2020) Plant Genome Response to the Incoming Coding Sequences: Stochastic Transcription Activation independently of the Chromatin Configuration. (submitted in Biorxiv simulataneously

with this manuscript)

Shilo, S., Tripathi, P., Melamed-Bessudo, C., Tzfadia, O., Muth, T. R. and Levy, A. A. (2017) T-DNA-genome junctions form early after infection and are influenced by the chromatin state of the host genome. *PLoS Genet,* **13**, e1006875. https://doi.org/10.1371/journal.pgen.1006875

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

Teissandier, A. and Bourc'his, D. (2017) Gene body DNA methylation conspires with H3K36me3 to preclude aberrant transcription. *EMBO J,* **36**, 1471–1473. https://doi.org/10.15252/embj.201796812

Van Oss, S. B. and Carvunis, A. R. (2019) De novo gene birth. *PLoS Genet,* **15**, e1008160. https://doi.org/10.1371/journal.pgen.1008160

Wagner, E. J. and Carpenter, P. B. (2012) Understanding the language of Lys36 methylation at histone H3. *Nat Rev Mol Cell Biol,* **13**, 115–126. https://doi.org/10.1038/nrm3274

Wu, D. D., Wang, X., Li, Y., Zeng, L., Irwin, D. M. and Zhang, Y. P. (2014) "Out of pollen" hypothesis for origin of new genes in flowering plants: study from Arabidopsis thaliana. *Genome Biol Evol,* **6**, 2822–2829. https://doi.org/10.1093/gbe/evu206

Yang, H., Howard, M. and Dean, C. (2014) Antagonistic roles for H3K36me3 and H3K27me3 in the cold-induced epigenetic switch at Arabidopsis FLC. *Curr Biol,* **24**, 1793–1797. https://doi.org/10.1016/j.cub.2014.06.047

Zhang, L., Ren, Y., Yang, T., Li, G., Chen, J., Gschwend, A. R., Yu, Y., Hou, G., Zi, J., Zhou, R., Wen, B., Zhang, J., Chougule, K., Wang, M., Copetti, D., Peng, Z., Zhang, C., Zhang, Y., Ouyang, Y., Wing, R. A., Liu, S. and Long, M. (2019) Rapid evolution of protein diversity by de novo origination in Oryza. *Nat Ecol Evol,* **3**, 679–690. https://doi.org/10.1038/s41559-019-0822-5

Zhao, L., Saelao, P., Jones, C. D. and Begun, D. J. (2014) Origin and spread of de novo genes in Drosophila melanogaster populations. *Science,* **343**, 769–772. https://doi.org/10.1126/science.1248286

FIGURE LEGENDS

Figure 1. Experimental design of the promoter-trap experiment in *A. thaliana* plants. Schematic illustration of the TRIP experiment performed in the T2 generation of *A. thaliana* transgenic lines. T-DNA including a barcode, a promoterless *LUC* gene and an expression cassette with a Km-resistance gene was introduced into *A. thaliana* via *Agrobacterium*-mediated

transformation. T2 seeds were harvested from Km-resistant T1 lines. Three seeds per T2 transgenic line were grown under the non-selective condition and subjected to subsequent locus and transcription-level analysis based on the TRIP method. NptII, neomycin phosphotransferase II; *nos*p, nopaline synthase promoter; *nos*t, nopaline synthase terminator.

Figure 2. An artificial evolutionary experiment revealed the genetic behaviours of the activated transcription of coding sequences inserted in *A. thaliana* plants. (**a**) The insertion loci and transcription levels of determined T2-plants (n = 76) were mapped on the *A. thaliana* chromosomes. The coloured bars indicate individual insertion sites and corresponding transcription levels based on their percentiles (High: 100–67, Mid: 66–34, and Low: 33–1). (**b**) Classification of T2-plants according to their transcription. (**c**) Number of T2-plants according to their insertion types: Genic sense, AS (antisense) or intergenic. The definition of each type is provided in the Experimental Procedures. (**d**) Fraction of transcribed *LUC* genes among T2-plants (n = 76) and T87 cultured cells (n = 4,443) (Satoh and Hata *et al.*, 2020) against each insertion type, as in (c). (**e**) Fraction of *LUC* genes in T2-plants (upper panel) and T87 cells (lower panel) (Satoh and Hata *et al.*, 2020) against their transcription levels, as normalized using the total number of each insertion type as 100%. ND, untranscribed *LUC* genes. (**f**) The abundance of transcribed *LUC* genes in each insertion type was classified according to their transcription levels; lower (10^1–10^4) and higher (10^5–10^7), as in (e). Each frequency was normalized to the number of transcribed *LUC* genes in each insertion type, which was set as 100%.

Figure 3. Transcription status of *LUC* insertion loci in T2-plants and WT plants. (**a**) Breakdown of *LUC* insertion loci in T2-plants (upper panel) and T87 cells (lower panel) (Satoh and Hata *et al.*, 2020) according to their transcription status compared with that of the inherent locus from WT transcriptome data. (**b**) Relative fractions of the breakdown classifications shown in (a) normalized to the percentage of untranscribed types in WT as 100%. (**c**) The transcription levels of T2-plants and corresponding inherent transcripts in WT plants were compared. *R*, Spearman's correlation test.

Figure 4. Localization analysis of epigenetic marks in selected T2-plants. (**a** and **b**) Locus details of the (a) T2:161 and (b) T2:205 lines. The genomic loci of *LUC* inserts are represented as the individual position of the RB–genome junction. **(c)** Transcription levels of the T2:161 and T2:205 lines relative to the endogenous *UBQ10* gene (AT4G05320). (**d** and **e**) Localization

patterns of three epigenetic marks (mC: upper panel; H3K36me3: middle panel; and H2A.Z: lower panel) around individual *LUC* insertion loci of the (d) T2:161 and (e) T2:205 lines, respectively. Individual localization signals were normalized to the enrichment of the control locus of each epigenetic mark (see Experimental Procedures) as 100%. The red bars indicate the analysed positions, which were normalized to the genomic position of the start codon of *LUC* inserts as zero. Error bar, ±SD of two biological replicates.

Figure 5. Localization of epigenetic marks around the transcription start site of the T2:161 line. (**a**) Transcription start site of the T2:161 line, as determined using a template-switching-based method. (**b**) Localization analysis of H2A.Z and H3K36me3 in T2:161 and WT plants, as in Figure 4d.

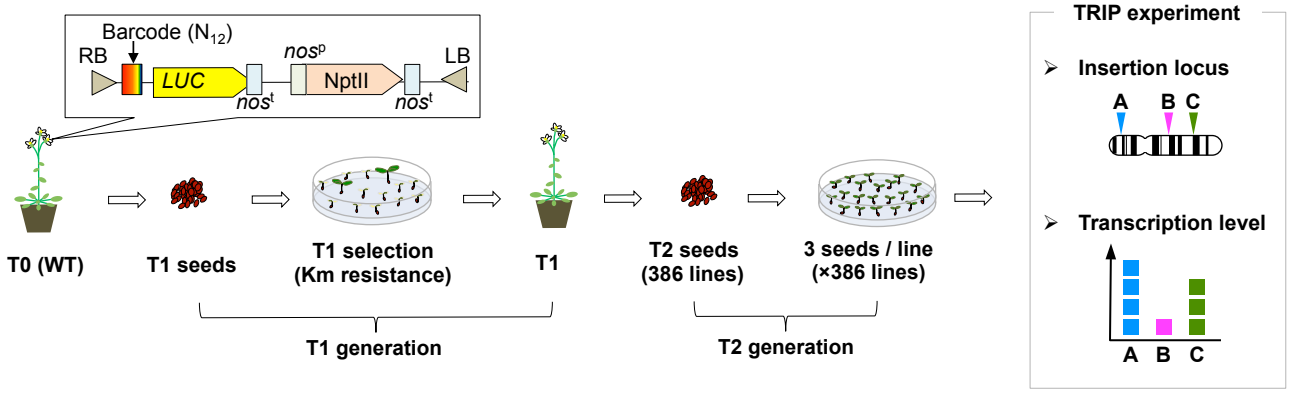

Figure 1. Experimental design of the promoter-trap experiment in *A. thaliana* plants. Schematic illustration of the TRIP experiment performed in the T2 generation of *A. thaliana* transgenic lines. T-DNA including a barcode, a promoterless *LUC* gene and an expression cassette with a Km-resistance gene was introduced into *A. thaliana* via *Agrobacterium*-mediated transformation. T2 seeds were harvested from Km-resistant T1 lines. Three seeds per T2 transgenic line were grown under the non-selective condition and subjected to subsequent locus and transcription-level analysis based on the TRIP method. NptII, neomycin phosphotransferase II; *nos*p, nopaline synthase promoter; *nos*t, nopaline synthase terminator.

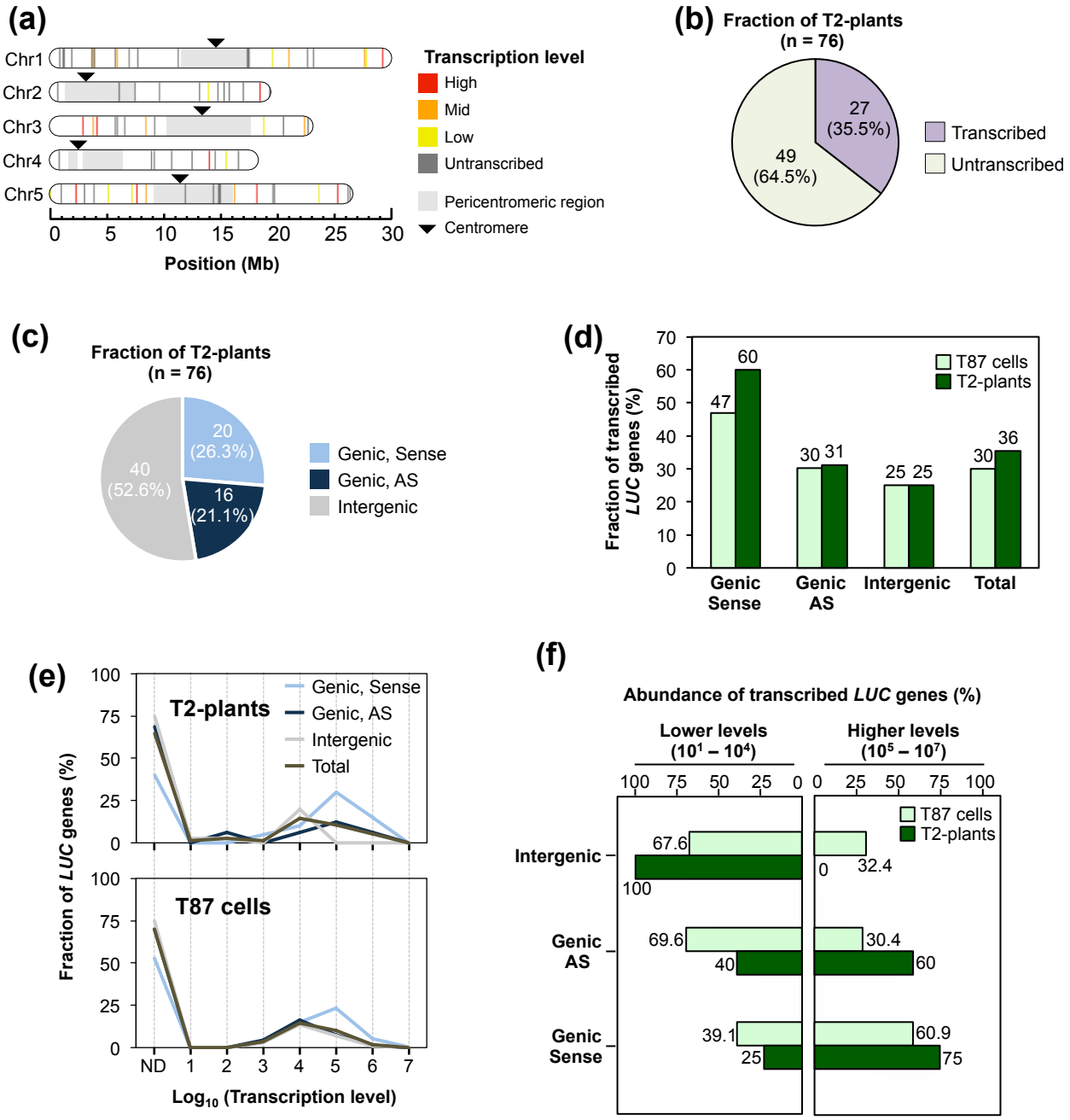

Figure 2. An artificial evolutionary experiment revealed the genetic behaviours of the activated transcription of coding sequences inserted in *A. thaliana* plants. (**a**) The insertion loci and transcription levels of determined T2-plants (n = 76) were mapped on the *A. thaliana* chromosomes. The coloured bars indicate individual insertion sites and corresponding transcription levels based on their percentiles (High: 100–67, Mid: 66–34, and Low: 33–1). (**b**) Classification of T2-plants according to their transcription. (**c**) Number of T2-plants according to their insertion types: Genic sense, AS (antisense) or intergenic. The definition of each type is provided in the Experimental Procedures. (**d**) Fraction of transcribed *LUC* genes among T2-plants (n = 76) and T87 cultured cells (n = 4,443) (Satoh and Hata *et al.*, 2020) against each insertion type, as in (c). (**e**) Fraction of *LUC* genes in T2-plants (upper panel) and T87 cells (lower panel) (Satoh and Hata *et al.*, 2020) against their transcription levels, as normalized using the total number of each insertion type as 100%. ND, untranscribed *LUC* genes. (**f**) The abundance of transcribed *LUC* genes in each insertion type was classified according to their transcription levels; lower (10^1–10^4) and higher (10^5–10^7), as in (e). Each frequency was normalized to the number of transcribed *LUC* genes in each insertion type, which was set as 100%.

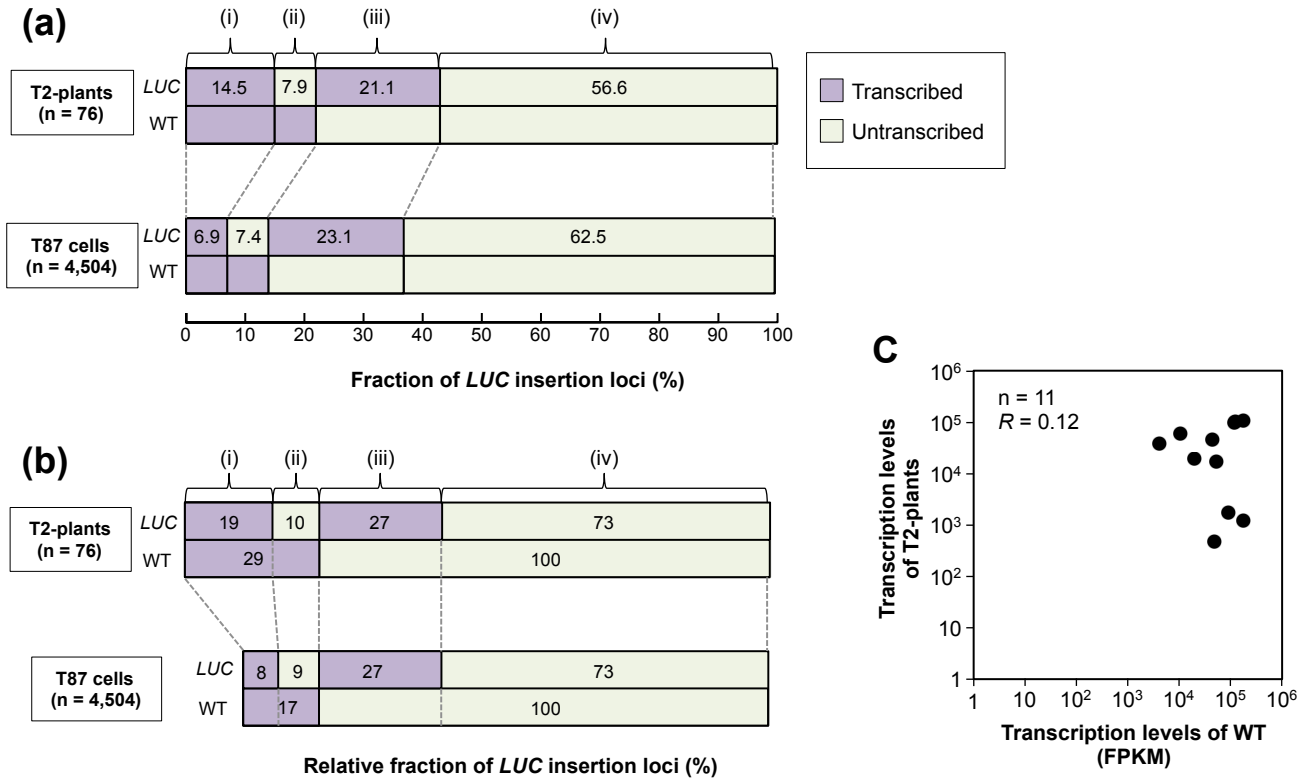

Figure 3. Transcription status of *LUC* insertion loci in T2-plants and WT plants. (**a**) Breakdown of *LUC* insertion loci in T2-plants (upper panel) and T87 cells (lower panel) (Satoh and Hata *et al.*, 2020) according to their transcription status compared with that of the inherent locus from WT transcriptome data. (**b**) Relative fractions of the breakdown classifications shown in (a) normalized to the percentage of untranscribed types in WT as 100%. (**c**) The transcription levels of T2-plants and corresponding inherent transcripts in WT plants were compared. *R*, Spearman's correlation test.

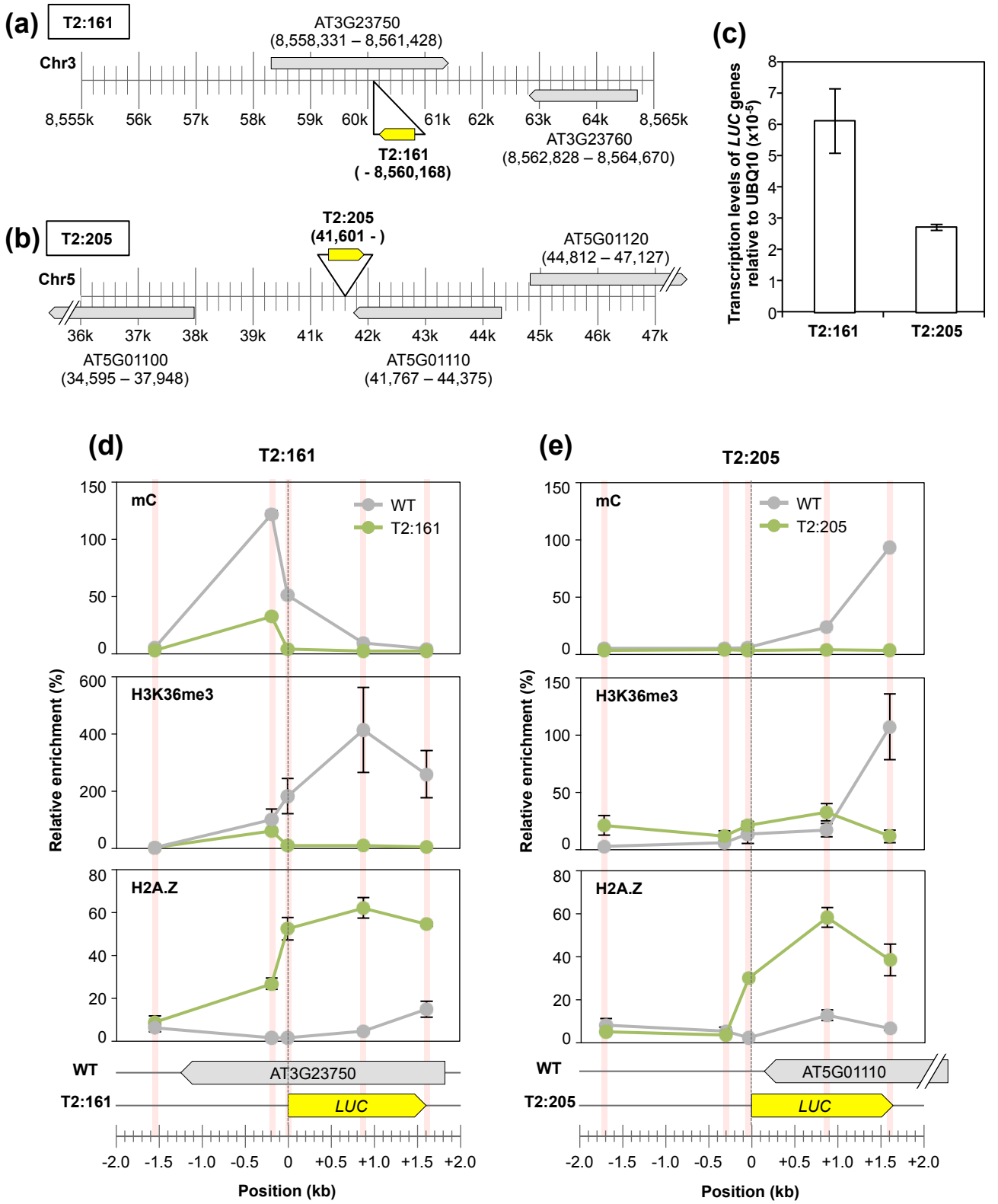

Figure 4. Localization analysis of epigenetic marks in selected T2-plants. (**a** and **b**) Locus details of the (a) T2:161 and (b) T2:205 lines. The genomic loci of *LUC* inserts are represented as the individual position of the RB–genome junction. **(c)** Transcription levels of the T2:161 and T2:205 lines relative to the endogenous *UBQ10* gene (AT4G05320). (**d** and **e**) Localization patterns of three epigenetic marks (mC: upper panel; H3K36me3: middle panel; and H2A.Z: lower panel) around individual *LUC* insertion loci of the (d) T2:161 and (e) T2:205 lines, respectively. Individual localization signals were normalized to the enrichment of the control locus of each epigenetic mark (see Experimental Procedures) as 100%. The red bars indicate the analysed positions, which were normalized to the genomic position of the start codon of *LUC* inserts as zero. Error bar, ±SD of two biological replicates.

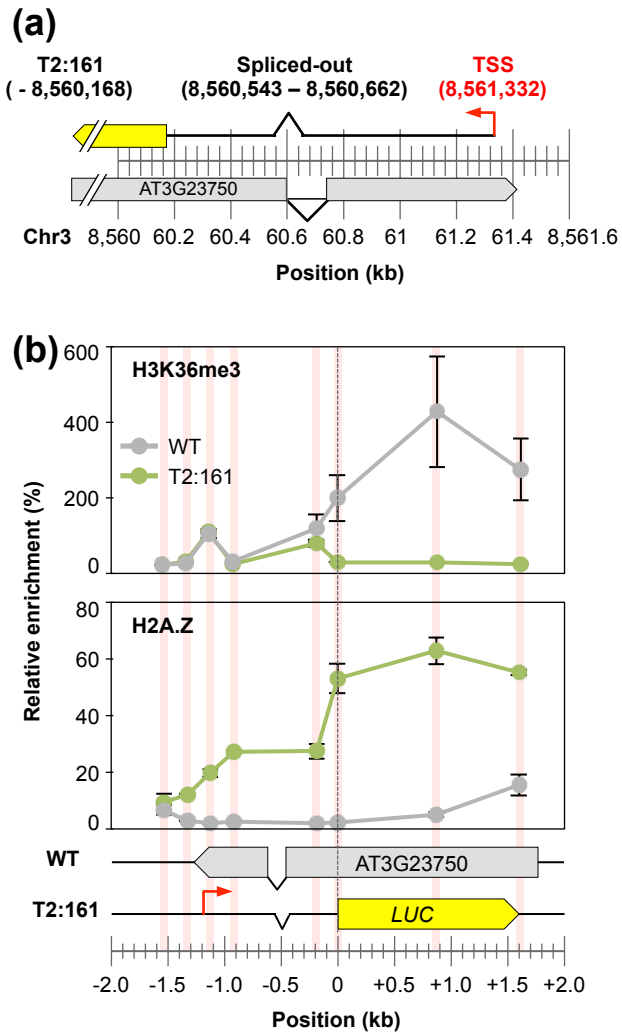

Figure 5. Localization of epigenetic marks around the transcription start site of the T2:161 line. (**a**) Transcription start site of the T2:161 line, as determined using a template-switching-based method. (**b**) Localization analysis of H2A.Z and H3K36me3 in T2:161 and WT plants, as in Figure 4d.

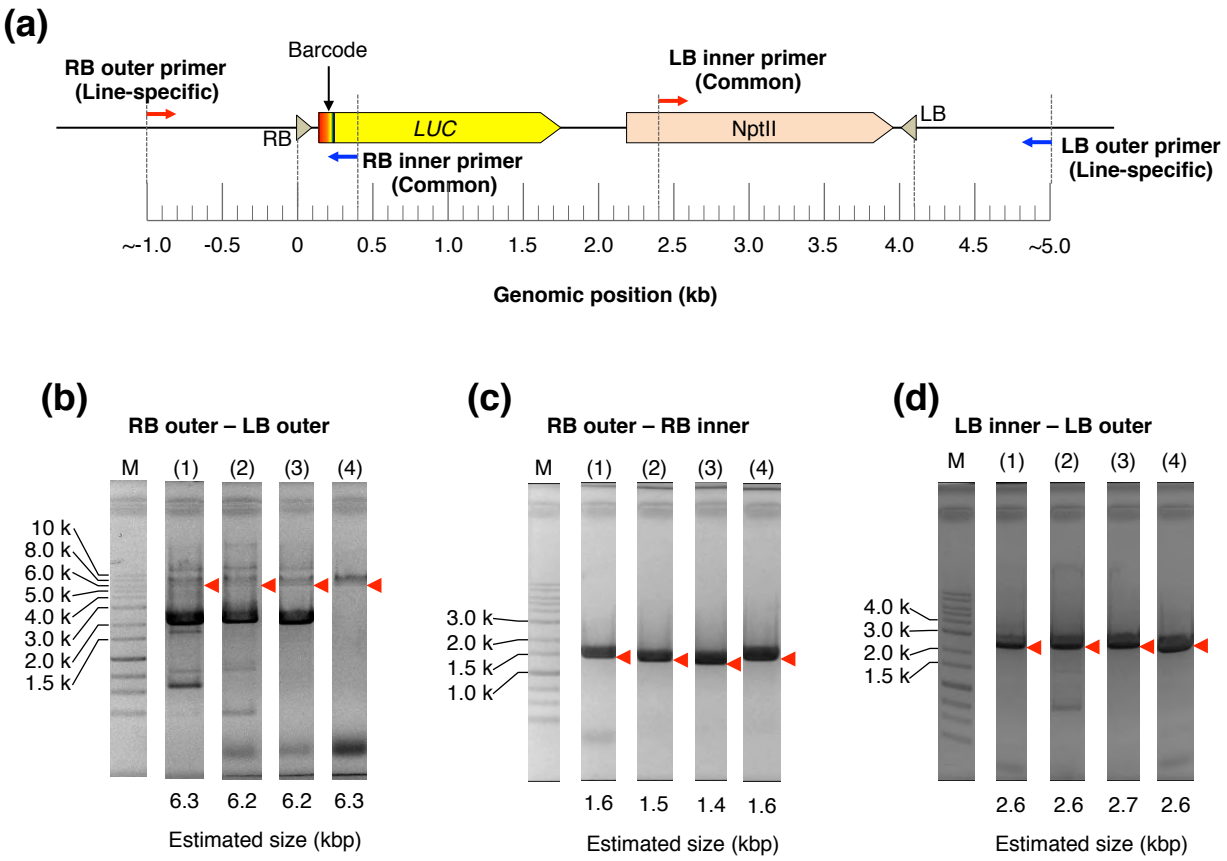

Figure S1. Validation of *LUC* insertion loci. (**a**) Schematic illustration of PCR-based validation of *LUC* insertion locus. In each selected transgenic line, RB outer and LB outer primers were designed about ±1.0 kb from RB and LB, respectively. RB inner and LB inner primers were in common with each line. (**b** and **d**) PCR products of randomly selected four lines were analyzed by the agarose gel electrophoresis. Primer sets used were (**b**) RB outer and LB outer, (**c**) RB outer and RB inner, and (**d**) LB inner and LB outer, respectively. The estimated sizes of PCR products in each line were calculated according to the determined locus by TRIP experiments. The bands corresponding to the expected sizes were indicated by red triangles. M: Molecular size marker.

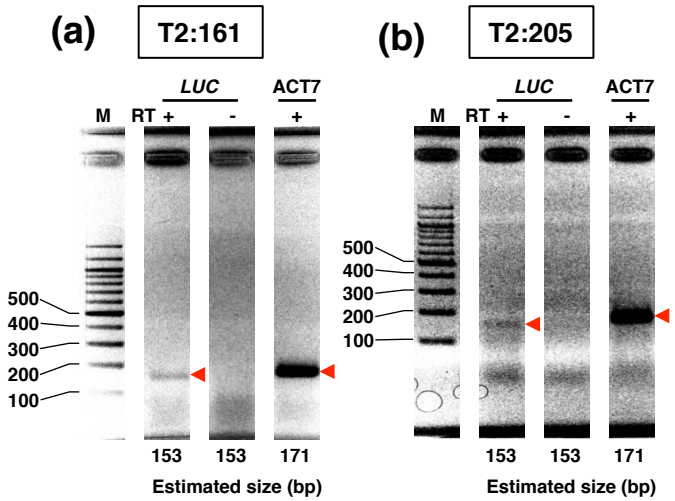

Figure S2. Expression analysis of T2-plants. (**a** and **b**) Expressions of (**a**) T2:161 line and (**b**) T2:205 line were validated by RT-PCR followed by gel electrophoresis. The bands corresponding to the expected sizes were indicated by red triangles. M: Molecular size marker, RT: Reverse transcription, and ACT7: AT5G09810.

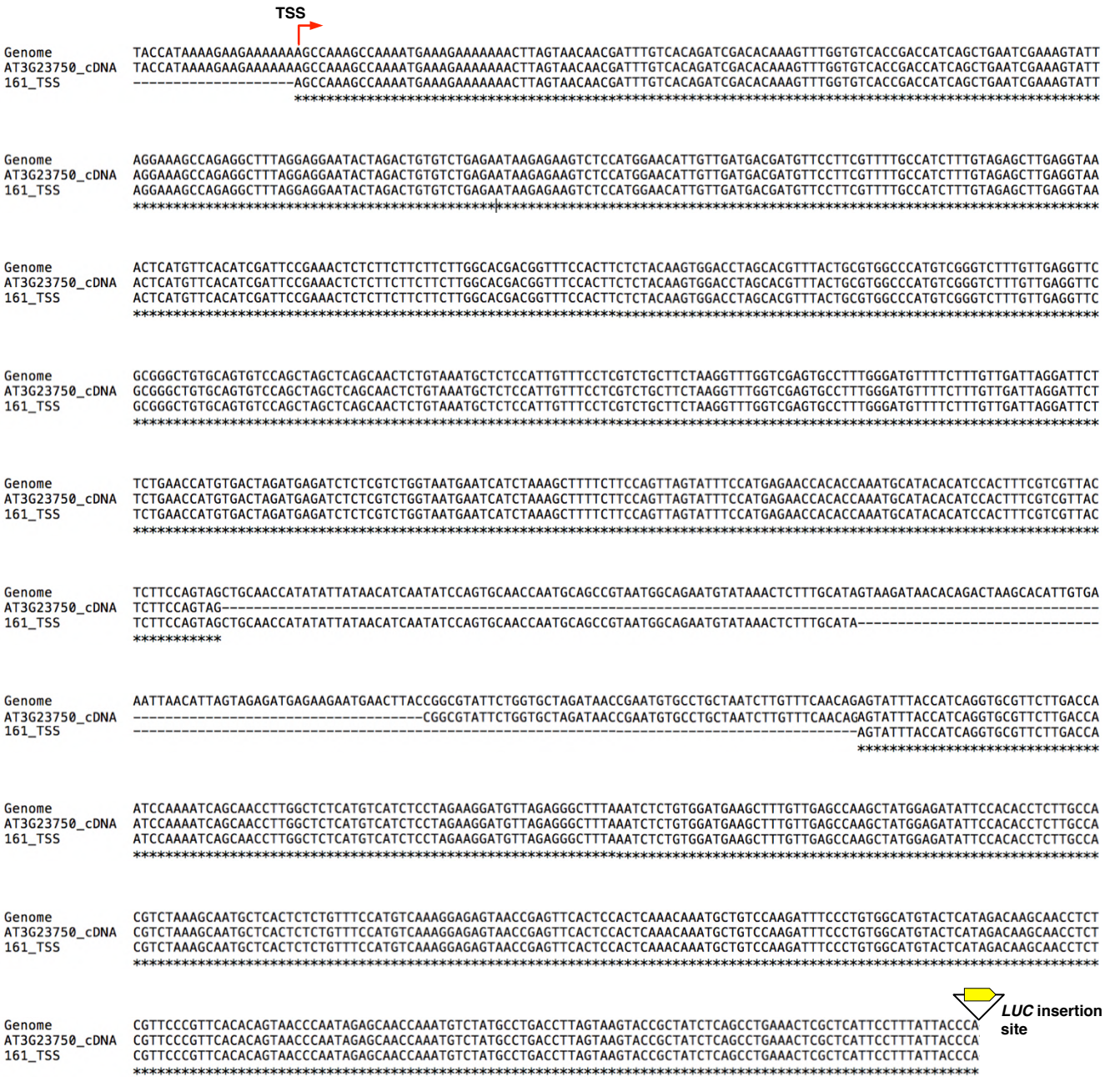

Figure S3. Sequence alignment of TSS of T2:161 line. Multiple alignments among genome, cDNA sequence of AT3G23750, and TSS flanking sequence of T2:161 line were showed. The red arrow indicated the TSS of the T2:161 line. Dash lines in the alignments indicated spliced regions. The alignment was calculated by using Mafft (v7.221) (Katoh and Standley, 2013).

Table S1. Primer list

T-DNA library construction

Name	Sequence (5' -> 3')	Descriptions
TRIP_LUC_EcoRI_r	TTAGGTAACCCAGTAGATCCAGAGG	These primers were used to introduce barcode into the T-DNA. Barcode was indicated by n.
TRIP_ITLB_barcodeF	AAAGTCGACGTTATCAGCTTACAGnnnnnnnnnnnnATGGAAGACGCCAAAAACAT	

Sequencing library preparation for the locus determination

Name	Sequence (5' -> 3')	Descriptions
TRIP_LUC_iPCR_F1.1	GTTGGGCGCGTTATTTATCGGAGTT	Primer set for the inverse PCR to specifically amplify *LUC*-including DNAs
TRIP_LUC_iPCR_R1	GTTTTCACTGCATACGACGATTCTG	
TRIP_iPCRAmpSeq_F2.1	gtctcgtgggctcggagatgtgtataagagacagCACATCTCATCTACCTCCCGGTTT	Primter set for the TAILed-PCR following the inverse PCR in order to add adapter sequence for next-generation sequencing. Adapter sequences were lowercased.
TRIP_iPCRAmpSeq_R2.1	tcgtcggcagcgtcagatgtgtataagagacagCTCTAGAGGATAGAATGGCGCCG	

Sequencing library preparation for the transcription level analysis

Name	Sequence (5' -> 3')	Descriptions
TRIP_AmpSeq_F_New2	tcgtcggcagcgtcagatgtgtataagagacagTCAAGGCCTCGACGTTATCAGC	Primer set for amplification of barcode region of cDNA/DNA with adding adapter sequence for next-generation sequencing. Adapter sequences were lowercased.
TRIP_AmpSeq_R	gtctcgtgggctcggagatgtgtataagagacagTCTAGAGGATAGAATGGCGCCGG	

Validation of *LUC* insertion loci

Name	Sequence (5' -> 3')	Descriptions
LUC_F_50	TAGAGGATGGAACCGCTGGAGA	A primer for the amplification of Barcode sequence
RB_inner	TCATAGCTTCTGCCAACCGAACG	A primer for the amplification of RB-genome junction and Barcode sequence
LB_inner	ATGACTGGGCACAACAGACAATC	A primer for the amplification of LB-genome junction
85_RB_outer	TGCAATCGTATCGGATTGGTTTCG	A primer for the amplification of RB-genome junction and T-DNA insert
85_LB_outer	ATGGGACGTTCTTACTGGCTTGTG	A primer for the amplification of LB-genome junction and T-DNA insert
161_RB_outer	CCGACCATCAGCTGAATCGAAAGT	A primer for the amplification of RB-genome junction and T-DNA insert
161_LB_outer	CGGAATAGTACCTCCGACGCTTCT	A primer for the amplification of LB-genome junction and T-DNA insert
201_RB_outer	AGCACAGCTCCACTCATAATTCCG	A primer for the amplification of RB-genome junction and T-DNA insert
201_LB_outer	TTTGACACCTCCACGTACACAAGC	A primer for the amplification of LB-genome junction and T-DNA insert
205_RB_outer	CGAACTCACTGATTTGATACCTGACCT	A primer for the amplification of RB-genome junction and T-DNA insert
205_LB_outer	AACGCGTTGTGCAGTAAAGGC	A primer for the amplification of LB-genome junction and T-DNA insert

Expression analysis of T2 plants

Name	Sequence (5' -> 3')	Descriptions
LUC_F_50	TAGAGGATGGAACCGCTGGAGA	RT-qPCR primer set for the *LUC* genes.
LUC_RB-0	TCATAGCTTCTGCCAACCGAACG	
ACTIN_2317_F	CTTTAGGATGCTTGTGATGATG	RT-qPCR primer set for the ACT7 (AT5G09810).
ACTIN_2463_R	CACCCGATACTTAAATAATTGTCTC	
UBQ10_F	GGCCTTGTATAATCCCTGATGAATAAG	RT-qPCR primer set for the UBQ10 (AT4G05320).
UBQ10_R	AAAGAGATAACAGGAACGGAAACATAGT	

TSS analysis of T2 plants

Name	Sequence (5' -> 3')	Descriptions
SgfI_Rd1Sp_T15V	AAAgcgatcgcTGTCTCTTATACACATCTGACGCTGCCGACGATTTTTTTTTTTTTTTV	Oligo dT primer for the reverse transcription. SgfI site and adapter sequence were added to the 5' end of cDNAs. SgfI site was lowercased.
SgfI_SMART_TSoligo	AAGCAGTGGTATCAACGCAGAGTgcgatcgc(rG)(rG)(rG)	Template-switching oligo. SgfI site and adapter sequence were added to the 3' end of cDNAs. Riboguanosine was indicated by (rG). SgfI site was lowercased.
TSanchor	AAGCAGTGGTATCAACGCAGAGT	Primer for the synthesis of the 2nd strand of the cDNA.
TRIP_LUC_iPCR_F1.1	GTTGGGCGCGTTATTTATCGGAGTT	Primer set for the inverse PCR to specifically amplify *LUC*-including cDNAs.
TRIP_LUC_iPCR_R1	GTTTTCACTGCATACGACGATTCTG	
RT_Rd1SpAnchor	GGCAGCGTCAGATGTGTATAAGA	Primer set for the nested PCR to specifically amplify *LUC*-including cDNAs.
LUC_RB-0	TCATAGCTTCTGCCAACCGAACG	

ChIP-PCR (T2:205 line)

Name	Sequence (5' -> 3')	Descriptions
205_LUC-1706_F	CCAAGTGAGTGAATGAGTGT	Primer set for ChIP-PCR in the T2:205 line. Primers aligned to the flanking region of *LUC* insert.
205_LUC-1706_R	CGTCCCGTATTTAGTTCGCA	
205_LUC-317_F	ATACGGATGTTGGTCGTT	Primer set for ChIP-PCR in the T2:205 line. Primers aligned to the flanking region of *LUC* insert.
205_LUC-317_R	AGGTTTATCCAATTCCTCTTGAC	
205_LUC-43_F	GTACGGAGGCCTCGACGTTAT	Primer set for ChIP-PCR in the T2:205 line. Primers aligned over the flanking region of *LUC* insert and ORF of *LUC* gene.
205_LUC-43_R	CGCCGGGCCTTTCTTTATGTTT	
205_LUC+866_F	TCAAAGTGCGTTGCTAGTACC	Primer set for ChIP-PCR in the T2:205 line. Primers aligned to the ORF of *LUC* gene.
205_LUC+866_R	CCCCAGAAGCAATTTCGTGT	
205_LUC+1599_F	CTTACCGGAAAACTCGACGCAAGA	Primer set for ChIP-PCR in the T2:205 line. Primers aligned to the ORF of *LUC* gene.
205_LUC+1599_R	CGGCCGCTTTACAATTTGGACT	
205_LUC-43_WT_F	AGAACAAACGCGTACGGA	Primer set for ChIP-PCR in the T2:205 line. Primers aligned to the corresponding loci in WT allele where 205_LUC-43_F and 205_LUC-43_R aligned in the *LUC* allele.
205_LUC-43_WT_R	AGGGTAGCTGCTAAAAGGAC	
205_LUC+866_WT_F	CTGATGCAATCCGGCACAAA	Primer set for ChIP-PCR in the T2:205 line. Primers aligned to the corresponding loci in WT allele where 205_LUC+866_F and 205_LUC+866_R aligned in the *LUC* allele.
205_LUC+866_WT_R	TGTCGTTCTGGTAATGCCTCAGATG	
205_LUC+1599_WT_F	TTCCCATGCTTACACAGTCCA	Primer set for ChIP-PCR in the T2:205 line. Primers aligned to the corresponding loci in WT allele where 205_LUC+1599_F and 205_LUC+1599_R aligned in the *LUC* allele.
205_LUC+1599_WT_R	GATGAATGCTATGCCGGGCAAA	

ChIP-PCR (T2:161 line)

Name	Sequence (5' -> 3')	Descriptions
161_LUC-1542_F	ACACAGCCTGTAACACTCATC	Primer set for ChIP-PCR in the T2:161 line. Primers aligned to the flanking region of *LUC* insert.
161_LUC-1542_R	AGTTTGTTTTGCCGCGTGAA	
161_LUC_TSS-200_F	TCTCAAAACCTAGCTACGGA	Primer set for ChIP-PCR in the T2:161 line. Primers aligned to the flanking region of *LUC* insert.
161_LUC_TSS-200_R	GCACATATTTGCGTCTGACCT	
161_LUC_TSS_F	CACCGACCATCAGCTGAATCGAAA	Primer set for ChIP-PCR in the T2:161 line. Primers aligned to the flanking region of *LUC* insert.
161_LUC_TSS_R	TCCATGGAGACTTCTCTTATTCTCAGACAC	
161_LUC_TSS+200_F	CTACAAGTGGACCTAGCACGTTTACTG	Primer set for ChIP-PCR in the T2:161 line. Primers aligned to the flanking region of *LUC* insert.
161_LUC_TSS+200_R	TGAGCTAGCTGGACACTGCACA	
161_LUC-192_F	CTGTCCAAGATTTCCCTGTGGCAT	Primer set for ChIP-PCR in the T2:161 line. Primers aligned to the flanking region of *LUC* insert.
161_LUC-192_R	GGTCAGGCATAGACATTTGGTTGCT	
161_LUC-8_F	CCTCGCATATGGAAGACGCCAAA	Primer set for ChIP-PCR in the T2:161 line. Primers aligned over the flanking region of *LUC* insert and ORF of *LUC* gene.
161_LUC-8_R	CTCTCCAGCGGTTCCATCCTCTA	
161_LUC+866_F	TCAAAGTGCGTTGCTAGTACC	Primer set for ChIP-PCR in the T2:161 line. Primers aligned to the ORF of *LUC* gene.
161_LUC+866_R	CCCCAGAAGCAATTTCGTGT	
161_LUC+1599_F	CTTACCGGAAAACTCGACGCAAGA	Primer set for ChIP-PCR in the T2:161 line. Primers aligned to the ORF of *LUC* gene.
161_LUC+1599_R	CGGCCGCTTTACAATTTGGACT	
161_LUC-8_WT_F	TCCAGCATACACGACACCGAAA	Primer set for ChIP-PCR in the T2:161 line. Primers aligned to the corresponding loci in WT allele where 161_LUC-8_F and 161_LUC-8_R aligned in the *LUC* allele.
161_LUC-8_WT_R	CAATGGAGGTTCTTCGCCAGGTTA	
161_LUC+866_WT_F	ACCTCCGGTGAAAACAAAG	Primer set for ChIP-PCR in the T2:161 line. Primers aligned to the corresponding loci in WT allele where 161_LUC+866_F and 161_LUC+866_R aligned in the *LUC* allele.
161_LUC+866_WT_R	ATAGTACCTCCGACGCTT	
161_LUC+1599_WT_F	TACGCCGGACCATTTGCAGAAATC	Primer set for ChIP-PCR in the T2:161 line. Primers aligned to the corresponding loci in WT allele where 161_LUC+1599_F and 161_LUC+1599_R aligned in the *LUC* allele.
161_LUC+1599_WT_R	GACCAAACAGCAATGCTCGCTT	

ChIP control sites

Name	Sequence (5' -> 3')	Descriptions
FLC_-480_F	TGTAGAGTGGAGGTTCTTTCTG	Primer set for the validation of H2A.Z enrichment and normalization for the enrichment level in T2 plants.
FLC_-480_R	TTTTGGGGGTAAACGAGAGT	
FLC_+49_F	CGACAAGTCACCTTCTCCAA	Primer set for the validation of H2A.Z enrichment.
FLC_+49_R	TTGGAGAAGGTGACTTGTCG	
ACTIN_31_F	GAGCTATATATTCGCACATGTACTCG	Primer set for the validation of H3K36me3 enrichment and normalization for the enrichment level in T2 plants.
ACTIN_124_R	GATACAGAAGATTCGAGAAGCAGC	
ACTIN_-871_F	CGTAGTTGATATGTATCTTGCTCC	Primer set for the validation of H3K36me3 enrichment.
ACTIN_-773_R	GATTGATCGGTTTTCGTGATATATC	
at1g13410_F	AGGTGGACATTGGCGAAGTTGC	Primer set for the validation of mC enrichment and normalization for the enrichment level in T2 plants.
at1g13410_R	AGCCGGGTTTCTTGGTTCAAGC	
at1g22500_F	ATTGATGCCTGGCTCCGTTCTC	Primer set for the validation of mC enrichment.
at1g22500_R	ACCCGGTACAGGAACGAGATTG	