Plant genome response to incoming coding sequences: stochastic transcriptional activation independent of chromatin configuration

Soichirou Satoh[1], Takayuki Hata[1,2], Naoto Takada[1], Makoto Tachikawa[1], Mitsuhiro Matsuo[2], Sergei Kushnir[3], and Junichi Obokata[2]*

[1] Graduate School of Life and Environmental Sciences, Kyoto Prefectural University, Kyoto-shi, Kyoto, 606-8552, Japan.

[2] Faculty of Agriculture, Setsunan University, Hirakata-shi, Osaka, 573-0101, Japan.

[3] Plant Molecular Genetics, Teagasc, Crop Science Department, Oak Park, Carlow R93 XE12, Ireland

* Corresponding author

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

Soichirou Satoh and Takayuki Hata should be considered joint first author.

Running Head

Stochastic behavior of transgene transcription

Keywords

Transcriptional activation, chromatin remodeling, transgene, horizontal gene transfer, endosymbiotic gene transfer, promoter, artificial evolutionary experiment, massively parallel reporter assay, promoter-trap screening, *Arabidopsis thaliana*

ABSTRACT

Horizontal gene transfer can occur between phylogenetically distant organisms, such as prokaryotes and eukaryotes. In these cases, how do the translocated genes acquire transcriptional competency in the alien genome environment? According to the conventional view, specific loci of the eukaryotic genome are thought to provide transcriptional competency to the incoming coding sequences. To examine this possibility, we randomly introduced the promoterless luciferase (LUC)-coding sequences into the genome of *Arabidopsis thaliana* cultured cells and performed a genome-wide “transgene location vs. expression” scan. We found that one-third of the 4,504 mapped *LUC* genes were transcribed. However, only 10% of them were explained by conventional transcriptional fusions with the annotated genes, and the remainder of the genes exhibited novel transcription that occurred independently of the chromatin configuration or transcriptional activity inherent to the given chromosomal locus; rather, their transcriptional activation occurred stochastically at about 30% of each insertion event, but independent of the integration sites. We termed this activation phenomenon as an integration-dependent stochastic transcriptional activation, a new type of response of the plant genome to incoming coding sequences. We discuss the possible roles of this phenomenon in the evolution of eukaryotic genomes.

INTRODUCTION

Horizontal or endosymbiotic gene transfer (HGT/EGT) events greatly contributed to the evolution and diversification of terrestrial life and organisms (Soucy, Huang and Gogarten, 2015; Husnik and McCutcheon, 2018). In the plant phyla, many genes originally encoded by photosynthetic bacteria were transferred to the host nuclear genome to produce photosynthetic eukaryotes (Timmis *et al.*, 2004), even though the bacterial genome systems were quite different from those of eukaryotes. The genetic flow from the plastid to the nuclear genome is ongoing (Matsuo *et al.*, 2005; Bock, 2017).

However, the manner in which foreign genes obtained by HGT/EGT events acquired transcriptional competency and were integrated into the host nuclear genome remains unclear. This question has been addressed using the following two approaches. The first consisted in the examination of the DNA that escaped the plastid genome and was integrated into the nucleus, which proposed two possible scenarios: (1) the translocated genes were transcribed as fusion transcripts via the trapping of endogenous promoters with several mutational pathways; or (2) some of the plastid promoters were compatible with, and functional in the nuclear genome environment (Stegemann and Bock, 2006; Wang *et al.*, 2014). The second approach included promoter/gene-trapping experiments that resulted in a cryptic promoter hypothesis, i.e., hypothetical cryptic promoters were postulated to explain the phenomenon of transcription of exogenously inserted promoterless coding sequences without trapping any annotated genes/promoters (Fobert *et al.*, 1994; Topping *et al.*, 1994; Mollier *et al.*, 2000; Plesch, Kamann and Mueller-Roeber, 2000; Yamamoto *et al.*, 2003; Sivanandan *et al.*, 2005; Stangeland *et al.*, 2005).

We previously demonstrated that such unexpected transcriptional activation in promoter-trapping experiments occurred via two different mechanisms: (1) cryptic promoter capturing, in which exogenous DNA was transcribed by trapping a pre-existing promoter-like chromatin configuration; and (2) promoter *de novo* origination, in which promoter-like epigenetic landscapes were newly formed via chromatin remodelling triggered by the insertion of DNA (Kudo, Matsuo, and Satoh *et al.*, 2020). However, the experimental throughput of this promoter-trapping experiment was too limited to illustrate the generality and extensiveness of these transcriptional activation mechanisms.

Here, to overcome the low-throughput of the conventional promoter-trapping experiments,

we applied a massively parallel reporter assay (Akhtar *et al.*, 2013; Inoue and Ahituv, 2015) to this analysis. We carried out a genome-wide "transgene location vs. expression" scan, and examined the mechanisms via which transcriptionally inert transgenes become activated in *Arabidopsis thaliana* cultured cells. We found that the transcriptional activation of transgenes occurred in two ways: (1) transcriptional fusion, in which the transgenes recruited endogenous *cis*-regulatory elements of *A. thaliana* genes; and (2) integration-dependent stochastic transcriptional activation. The latter activation occurred at about 30% of each transgene integration event independently of the integration locus. This represents a new class of transcriptional activation mechanism that could explain the manner in which the incoming coding sequences in HGT/EGT acquire transcriptional competency in the host nuclear genome. The possible roles of epigenetic chromatin remodelling in this activation phenomenon are also discussed.

RESULTS

Genome-wide scanning of "transgene location vs. expression"

To understand the rules that govern the transcriptional activation of alien genes, we introduced thousands of promoterless luciferase (*LUC*) genes into *A. thaliana* T87 suspension-cultured cells via *Agrobacterium*-mediated transformation (Figure S1a). In the pools of transformed cells, 4,504 *LUC* genes were mapped onto the *A. thaliana* genome (Figure 1a, and Figures S1b, c, and S2). The *LUC* genes were evenly distributed across the length of the five *A. thaliana* chromosomes, with the exception of the pericentromeric regions, where the insertion frequency was significantly lower (Figure 1a). The relative abundance of the *LUC* genes inserted in the intergenic, genic, and promoter regions was proportional to the relative length of these genomic regions (Figure 1b, and Figure S3), indicating that *LUC* genes were unbiasedly integrated into the genome and, thus, were suitable for the genome-wide scanning of transgene transcriptional activation events.

RNA deep sequencing was used to detect and measure *LUC* expression. Unique barcode identifiers enabled us to link the distinct genomic locations and *LUC* transcription levels at each position (Figure 1c–e, and Figure S1b and d). We found that 1,355 of the 4,504 *LUC* genes identified were transcribed with a ~10^5-fold variation in *LUC* mRNA levels (Figure 1d). Some

barcodes could possibly behave as *cis*-regulatory elements and affect their own expression. However, our correlation analyses did not provide evidence of such function of barcode sequences (Figures S4 and S5).

Identification of two distinct pathways of transgene transcriptional activation in the plant genome

In the simplest case scenario, promoterless *LUC* transcription is a result of the trapping of endogenous transcription units. To test this model, we classified the 4,504 *LUC* loci into five insertion types in relation to the annotated genes: (1) sense and (2) antisense orientation within the gene-coding regions, (3) sense and (4) antisense orientation in the promoter regions, and (5) intergenic regions. According to this classification, 25%–30% of the *LUC* genes in individual types were transcribed without the genic-sense insertion type, which in turn accounted for about 50% of the transcribed fraction (Figure 1c). How can this over-representation be explained?

As shown in Figure 1d, the transcription levels of *LUC* genes in each insertion type ranged from 10^1 to 10^7 at the mean transcription level of 10^4, with only that of the genic sense type occurring at the level of 10^5 (Figure 1d). The comparison of the distribution profiles of the five types of insertion revealed that the genic-sense type had a superposed fraction at higher transcription levels (10^5–10^7) (light-blue fraction in Figure 1d), which accounted for ~10% of all of the transcribed *LUC* genes (Figure 1c and d). To address this phenomenon, we next examined the relationship between the *LUC* insertion sites relative to the annotated transcription start sites (TSSs) and *LUC* transcription levels (Figure 1e). We found that the abundance of *LUC* inserts with higher transcription levels (10^5–10^7) was over-represented at 0.2–2.4 kb downstream of the annotated TSSs ((iii) in Figure 1e and f). This over-represented fraction corresponded to ~10% of all transcribed *LUC* genes. In *A. thaliana*, the median length of the 5′ untranslated regions (UTRs) and mRNAs is ~70 and ~1,900 bp, respectively (as calculated from the TAIR10 database, https://www.arabidopsis.org/index.jsp). Therefore, it is very likely that the highly transcribed inserts at 0.2–2.4 kb downstream of the annotated TSSs, which accounted for 10% of all transcribed *LUC* inserts, represent conventional gene fusions with intrinsic *A. thaliana* genes. Without this superposed fraction of 10% of the inserts, the remainder of the genic-sense-type inserts showed remarkably similar transcription-level distribution profiles to those of the other insertion types (Figure 1c–f).

Our results indicate that the transcribed promoterless *LUC* reporters can be attributed to

two mechanisms. One is the so-called promoter/gene-trapping (Springer, 2000; Stanford, Cohn and Cordes, 2001), which generated transcriptional fusions of *LUC* genes with *A. thaliana* pre-existing genes/transcripts ((iii) in Figure 1e and f) and accounted for 10% of all transcribed *LUC* genes. The other was observed for the majority of the transcribed *LUC* genes, with transcription levels that were one order of magnitude lower than the promoter/gene-trap events (Figure 1d). Hereafter, we termed such events as ubiquitous *de novo* transcriptional activation. Even within the genic-sense insertion type (28.2% of all transcribed *LUC* inserts, Figure 1g), the majority of the transcribed *LUC* genes were most likely attributed to this *de novo* transcription (Figure 1d–f). Hence, we estimated that, in this study, the *de novo* transcriptional activation was responsible for the expression of more than 90% of transcribed *LUC* genes (Figure 1g). To confirm that this *de novo* transcription signal was not the background noise of the sequencing analysis, we reanalysed the inserted *LUC* genes that were represented by high-number reads in our DNA and RNA sequencing datasets (Figure S6). An increased stringency cut-off reduced the number of *LUC* genes from 4,504 (read number >5) to 2,057 (read number >100) or 140 (read number >1,000), but two distinct peaks corresponding to *de novo*-type and promoter-trap-type transcription events were clearly detected (Figure S6a and b), as in Figure 1d. In addition, we confirmed that three biologically independent samples gave remarkably similar *LUC* expression profiles (Figure S6c–e) to those depicted in Figure 1d. We concluded that the ubiquitous low-level *de novo* transcriptional activation of transgenes was not an experimental artefact; rather, it may be an outcome of the intrinsic *A. thaliana* genome properties, i.e., the transgene transcriptional activation mechanism.

Promoterless *LUC* genes were transcribed regardless of inherent transcriptional activities

Pervasive transcription throughout the genome characterizes eukaryotic organisms. We asked whether the transcription of the *LUC* genes could be primarily explained by integration within such pervasively transcribed regions. To define the genomic transcription landscape of the *A. thaliana* T87 cells studied here, we performed deep RNA sequencing of the wild-type (WT) cells. We classified the 4,504 *LUC* loci by comparing their transcription status between WT and transgenic cells (Figure 2a). Unexpectedly, only 6.9% of the *LUC* genes were transcribed in the inherently transcribed genomic regions (type (iii) in Figure 2a), whereas 23.1% of the *LUC* genes were transcribed in the transcriptionally inert genomic regions (type (i) in Figure 2a). Regarding the 6.9% of the *LUC* genes for which the integration sites were transcribed in both WT and

transgenic cells, we compared the transcription levels between the two types of cells, but found no correlation (Figure 2b, $r = 0.182$). Two conclusions were drawn from this analysis: (1) the transgene transcriptional activities do not reflect the inherent transcriptional activities; and (2) transgene transcriptional activation occurs independently of the transcriptional states of the given genomic regions.

Transcriptional activation of promoterless *LUC* genes was not affected by the inherent chromatin configuration

We wondered whether *LUC* transgenes would be able to overcome the effects of the silencing histone code. In *A. thaliana*, the dimethylation of the ninth lysine residue of histone H3 (H3K9me2) is thought to be associated with transcriptional silencing in the heterochromatin regions (Shu *et al.*, 2012; Bühler and Moazed, 2007; Grewal and Jia, 2007). A ChIP-Seq analysis of the WT *A. thaliana* T87 cells revealed that 15.6% of the genome was covered by H3K9me2-containing chromatin and was largely associated with pericentromeric heterochromatin. In the transgenic cells, only 120 *LUC* genes were inserted into the H3K9me2-containing chromatin regions (see the legend of Figure 3). The observation that this integration frequency was one-seventh of that observed in the H3K9me2-free genomic regions suggests that the heterochromatic region is less prone to transgene integration. However, 28% of the heterochromatin-inserted *LUC* genes were transcribed at levels similar to those of the transgenes that were inserted in the other regions (Figure 3). The transcription-level distribution of *LUC* genes in the heterochromatic region was also similar to that detected for genes with other insertion types, i.e., the genic-sense type (Figure 3a). The transcription levels of these *LUC* genes did not show any correlation with the degree of H3K9me2 modification (Figure 3b). Furthermore, two transcribed *LUC* genes were located 63 kb and 682 kb from the centromeres (Figure S7), although the pericentromeric heterochromatin regions extend over several megabases (Bernatavichute *et al.*, 2008). We concluded that transgene transcriptional activation at an approximate rate of 30% is not affected by the inherent chromatin configuration.

Integration-dependent stochastic activation of transgene transcription

As described above, transgene transcriptional activation was observed for 30% of the *LUC* inserts. Next, we wondered whether this phenomenon reflects the portion of activation-competent chromosomal regions. To understand whether some genomic locations have unknown properties that facilitate the transcriptional activation of the transgenes, we

analysed the transcriptional behaviour of *LUC* genes that were integrated in close neighbouring locations (Figure 4a). Theoretically, *LUC* pairs inserted in close proximity could result in three transcriptional fates: expression of both *LUC* genes (Fate A), expression of one *LUC* gene and silencing of the other (Fate B), and silencing of both *LUC* genes (Fate C) (Figure 4b). If the transgene transcriptional activation depends on the chromosomal locus, the transcriptional fates of neighbouring *LUC* inserts are expected to be similar (Figure 4c). Hence, in this scenario, only Fate A and C would be observed for the *LUC* pairs (Figure 4c). Moreover, the expected ratio between Fate A and C would be 30:70 (Figure 4c), assuming an average transcriptional activation rate of 30%, as measured by the described genome-wide analysis. Conversely, if the transgene transcriptional activation occurs stochastically at 30% of each integration event and is independent of the chromosomal locus, the distribution of the transcriptional fates of *LUC* pairs would fit the joint probability of two individual activation events (Figure 4d). In this model, the distribution ratio among Fates A, B, and C would be 9, 42, and 49, respectively (Figure 4d). According to these expectations, we examined which activation model fits the transcriptional activation of promoterless *LUC* genes. In our dataset, we identified 21 genomic locations in which independent *LUC* inserts were integrated within a 50 bp sliding window. Among these 21 *LUC* insert pairs, all three possible transcriptional fates were observed, as follows: Fate A, three cases; Fate B; five cases; and Fate C, 13 cases (Figure 4e, upper panel). This distribution fits the integration-dependent stochastic model of transcriptional activation (Figure 4d), rather than the chromosomal-locus-dependent model (Figure 4c). In fact, the expected values, i.e., Fate A (1.9 events), Fate B (8.8 events), and Fate C (10.2 events), were not significantly different from the observed rates (Fisher's exact test, $P = 0.55$) (Figure 4f). To perform a more rigorous test of the stochastic transcriptional activation model, we reduced the sliding window to 10 bp, which yielded 12 genomic locations (Figure 4e, lower panel). Similar to the results of the 50 bp sliding window analysis, the pairwise *LUC* comparison did not detect significant differences (Fisher's exact test, $P = 0.82$) between the observed and theoretical values (Fate A, 1.0 vs. 1.1; Fate B, 3.0 vs. 5.0; Fate C, 8.0 vs. 5.9) (Figure 4f). It should be emphasized that the individual *LUC* inserts used for this integration-site neighbourhood analysis stemmed from different, independently transformed cells that passed through at least 5–10 cell divisions before nucleic acid extraction. Thus, we concluded that the transgene transcriptional activation in a given genome location was likely to be the outcome of an integration-dependent stochastic phenomenon.

DISCUSSION

In this study, we performed an unbiased genome-wide screening of promoter-trapping events using a non-selective reporter. Collectively, the data led to the identification of integration-dependent stochastic transcriptional activation, which cannot be explained by the hijacking of the *A. thaliana cis*-regulatory elements, as occurs in the conventional transcriptional fusions. To date, we have not been able to find any specific motifs that were enriched at the 5′ proximal regions of the transcribed *LUC* genes, which was in contrast to the annotated gene promoters (Figure S8). To investigate the *cis*-regulatory elements of *de novo*-activated transcription in greater detail, TSS determination of the promoterless *LUC* genes is needed, which will be described elsewhere.

It is generally accepted that T-DNA integration requires the formation of double-stranded DNA breaks (DSBs), the repair of which could result in the joining of the T-DNA with the genomic DNA, predominantly via the activity of the non-homologous end-joining (NHEJ) DNA repair pathway (Magori and Citovsky, 2011; Kleinboelting *et al.*, 2015). The activity of the DNA repair machinery remodels chromatin at sites of DNA damage and leaves so-called DNA damage scars in the chromatin epigenetic configuration (Soria, Polo and Almouzni, 2012; Dabin, Fortuny and Polo, 2016). Thus, chromatin remodelling at the site of T-DNA integration may account, at least in part, for integration-dependent stochastic transcriptional activation. If this is the case, the expected chromatin remodelling could be the mechanism underlying the *de novo* promoter origination identified by us previously (Kudo, Matsuo, and Satoh *et al.*, 2020).

In the present study, we found and characterized a new class of transcription activation mechanism, i.e., integration-dependent stochastic transcriptional activation. This activation did not require fusion with pre-existing genes/transcripts, which would have harmful effects on the host gene network. Therefore, this finding sheds light on the long-standing question of how horizontally transferred genes acquired transcriptional competency in phylogenetically distant organisms. To scrutinize the biological significance of this novel transcriptional activation phenomenon in the plant genome, its molecular mechanisms and stability across generations need to be elucidated.

EXPERIMENTAL PROCEDURES

Construction of barcode-labelled plasmid libraries

The transformation vector plasmid was constructed using a modified pGreenII vector (Hellens *et al.*, 2000; Hirashima *et al.*, 2006) to encode 12 bp of random sequence ("barcode"), a promoterless firefly luciferase (*luc*+) coding sequence, a *nos* terminator sequence and an expression cassette of a kanamycin-resistant gene within the T-DNA region (Figure S1a and Methods S1).

Plant cell culture and transformation

A. thaliana T87 cells (Axelos *et al.*, 1992) were cultured in mJPL3 medium (Ogawa *et al.*, 2008) under continuous illumination (60 μE m^{-1} s^{-1}) at 22°C with shaking (120 rpm). One-week-old cultures were collected using a 10 μm nylon mesh, washed with H_2O twice and subjected to DNA, RNA and chromatin isolation and transformation.

Agrobacterium tumefaciens (GV3101) cells were transformed with the barcode-labelled libraries. *Agrobacterium*-mediated transformation of *A. thaliana* T87 cells was carried out according to the published protocol (Ogawa *et al.*, 2008). We obtained three independently transformed pools of T87 cells (termed TRIP pools hereafter), which were grown on mJPL3 plates containing 25 μg ml^{-1} meropenem (MEPM) and 30 μg ml^{-1} kanamycin (Km) at 22°C under continuous illumination for about 2 weeks. Green calli were cultured in liquid mJPL3 medium containing 12.5 μg ml^{-1} MEPM and 10 μg ml^{-1} Km with shaking under continuous illumination at 22°C for 2 weeks. Finally, the cells were transferred to fresh mJPL3 medium and grown for 1 additional week.

Determination of the insertion loci of *LUC* genes

Two micrograms of genomic DNA extracted from the TRIP pools using the DNeasy Plant Mini Kit (QIAGEN) were digested completely with *Dpn*II, purified using the QIAquick PCR purification kit (QIAGEN) and circularized with T4 DNA ligase. After purification using the QIAquick PCR purification kit, the circularized DNA was subjected to inverse PCR using primers that were designed to hybridize within the *LUC* gene (Figure S1b). From this point, we prepared two types of sequencing libraries: (1) The inverse PCR product was digested completely with *Apa*LI or *Sca*I to block the amplification of the vector-backbone-containing fragments in the subsequent steps.

Nested PCR was performed, followed by sequencing library preparation using the Nextera XT DNA Sample Prep Kit (Illumina); (2) The inverse PCR product was subjected to tailed-PCR and digestion with *Apa*LI or *Sca*I, followed by the addition of terminal adapters via one additional round of PCR, to prepare the sequencing libraries essentially according to Akhtar *et al*. (Akhtar *et al*., 2013). Sequencing was performed on an Illumina MiSeq sequencer with 301 bp paired-end reads.

The insertion loci of *LUC* genes were determined using an open-source software and custom Perl scripts (Figure S1b and c). Briefly, the sequencing reads were trimmed from the 3′ end with a phred-scaled quality score ≥30. Reads containing a *LUC* segment (31 bp), the barcode (12 bp) and a *LUC* flanking sequence (25–50 bp) were extracted. The *LUC* flanking sequences were mapped to the TAIR10 version of the *A. thaliana* genome using Bowtie (Langmead *et al*., 2009) with the following parameters; *bowtie -m 1 -v 3*. Subsequently, the 3′-junction sites of the mapped flanking sequences were defined as the genomic loci of the corresponding *LUC* inserts. Reliable locus–barcode pairs of *LUC* inserts were collected according to their read depth; at least three reads and 90% of individual mapped loci were occupied by an identical barcode sequence. We combined all *LUC* loci that were derived from three biologically independent TRIP pools, as well as from two of mapping libraries, and subjected them to subsequent analyses. For additional details, see Methods S1.

Determination of the relative transcription levels of *LUC* genes

RNA was extracted from the TRIP pool using the RNeasy Plant Mini Kit (QIAGEN) and treated with RNase-free DNase I (QIAGEN). cDNA was synthesized from 5 μg of the RNA using an oligo dT_{15} primer and SuperScript III Reverse Transcriptase (Thermo Fisher Scientific). Sequencing libraries were prepared by amplification of the barcode region using primers with an adapter extension, followed by tailed-PCR using Nextera XT Index Primers (Illumina) (Figure S1b). From an aliquot of DNA from the TRIP pools, sequencing libraries of the barcode region were prepared using the method described above. These cDNA and DNA libraries were sequenced on an Illumina MiSeq sequencer with 76 bp paired-end reads.

To determine the relative transcription levels of *LUC* genes, barcode sequences were extracted from sequencing reads and counted. Barcode sequences with a read number ≤5 in the DNA library were omitted. Moreover, barcode sequences with a read number ≤5 in the cDNA library were set as zero. For each library, the read number of each barcode was normalized to

the total reads of the library. To obtain an indicator of the RNA level per DNA molecule, the cDNA read number was divided by the corresponding DNA read number and multiplied by 10,000, which was used to indicate the transcription levels of the individual *LUC* genes. For additional details, see Methods S1.

RNA-Seq

RNA was extracted using the RNeasy Plant Mini Kit (QIAGEN) and treated with RNase-free DNase I (QIAGEN). RNA-Seq libraries were prepared using the SureSelect Strand-Specific RNA-Seq Kit (Agilent), according to the manufacturer's instructions. The libraries were sequenced on an Illumina MiSeq sequencer with 76 bp paired-end reads. The sequencing reads from two replicated experiments were combined. The transcribed regions and their expression levels were determined using TopHat2 (Kim *et al.*, 2013) and Cufflinks (Trapnell *et al.*, 2010), with the *A. thaliana* genome (TAIR10) as a reference for mapping.

Chromatin immunoprecipitation-Seq (ChIP-Seq)

The fixation of *A. thaliana* T87 cells, chromatin isolation and fragmentation and ChIP (antibody: anti-H3K9me2 (MABI, 308-32361)) were performed basically as described by Saleh *et al.* (Saleh, Alvarez-Venegas and Avramova, 2008). Successful enrichment of ChIPed DNA was validated according to To *et al.* (To *et al*., 2011). ChIP-Seq libraries were prepared using the DNA SMART ChIP-Seq Kit (Takara Clontech), according to the manufacturer's instructions. Libraries were sequenced on an Illumina MiSeq sequencer with 76 bp paired-end reads. The sequences derived from a template-switching reaction were trimmed from the reads. Subsequently, the reads from two replicated experiments were combined and mapped to the *A. thaliana* genome (TAIR10) using Bowtie2 (Langmead and Salzberg, 2012). Peaks corresponding to H3K9me2 enrichment were called using MACS (version 2) (Zhang *et al.*, 2008).

ACKNOWLEDGEMENTS

We thank A. Katahata, I. Ohshima and M. Mino for useful discussions. *Arabidopsis* T87 suspension-cultured cells (rpc00008) were provided by RIKEN BRC which is participating in the National Bio-Resource Project of the MEXT, Japan. This work was supported by Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Number [23125512, 23657040, 23117006

and 25650131 to J.O., 26660008 to S.S.], by Sekisui Chemical Co. Ltd., and by the Strategic Research Funds at Kyoto Prefectural University.

SUPPORTING INFORMATION

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

Figure S1. Precise workflow of the massive promoter-trap screening.

Figure S2. Validation of the *LUC* loci and barcode sequences.

Figure S3. Length of each genomic context.

Figure S4. Assessment of the effect of barcode sequences on the transcription levels of *LUC* genes.

Figure S5. Similarity/dissimilarity of the transcription levels of the randomly selected *LUC* pairs against the sequence identity of the 12-base barcode.

Figure S6. Expression profiles of *LUC* genes with high-number reads from the amplicon-sequencing data and of *LUC*s from biological replicates.

Figure S7. Two examples of transcribed *LUC* genes in the H3K9me2-marked regions located around the centromere.

Figure S8. Distribution of *cis*-regulatory elements in the upstream region of *LUC* integration sites.

Table S1. All primers used in this study were listed.

Methods S1. Detailed experimental procedures.

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 paralle. *Cell,* **154**, 914–927.

https://doi.org/10.1016/j.cell.2013.07.018

Axelos, M., Curie, C., Mazzolini, L., Bardet, C. and Lescure, B. (1992) A protocol for transient gene expression in Arabidopsis thaliana protoplasts isolated from cell suspension cultures. *Plant Physiol. Biochem.,* **30**, 123–128.

Bernatavichute, Y. V., Zhang, X., Cokus, S., Pellegrini, M. and Jacobsen, S. E. (2008) Genome-wide association of histone H3 lysine nine methylation with CHG DNA methylation in Arabidopsis thaliana. *PLoS One,* **3**, e3156. https://doi.org/10.1371/journal.pone.0003156

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

Bühler, M. and Moazed, D. (2007) Transcription and RNAi in heterochromatic gene silencing. *Nat Struct Mol Biol,* **14**, 1041–1048. https://doi.org/10.1038/nsmb1315

Dabin, J., Fortuny, A. and Polo, S. E. (2016) Epigenome Maintenance in Response to DNA Damage. *Mol Cell,* **62**, 712–727. https://doi.org/10.1016/j.molcel.2016.04.006

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

Grewal, S. I. and Jia, S. (2007) Heterochromatin revisited. *Nat Rev Genet,* **8**, 35-46. https://doi.org/10.1038/nrg2008

Hellens, R. P., Edwards, E. A., Leyland, N. R., Bean, S. and Mullineaux, P. M. (2000) pGreen: a versatile and flexible binary Ti vector for Agrobacterium-mediated plant transformation. *Plant Mol Biol,* **42**, 819–832. https://www.ncbi.nlm.nih.gov/pubmed/10890530

Hirashima, M., Satoh, S., Tanaka, R. and Tanaka, A. (2006) Pigment shuffling in antenna systems achieved by expressing prokaryotic chlorophyllide a oxygenase in Arabidopsis. *J Biol Chem,* **281**, 15385–15393. https://doi.org/10.1074/jbc.M602903200

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

Inoue, F. and Ahituv, N. (2015) Decoding enhancers using massively parallel reporter assays. *Genomics,* **106**, 159–164. https://doi.org/10.1016/j.ygeno.2015.06.005

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. and Salzberg, S. L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol,* **14**, R36. https://doi.org/10.1186/gb-2013-14-4-r36

Kleinboelting, N., Huep, G., Appelhagen, I., Viehoever, P., Li, Y. and Weisshaar, B. (2015) The Structural Features of Thousands of T-DNA Insertion Sites Are Consistent with a Double-Strand Break Repair-Based Insertion Mechanism. *Mol Plant,* **8**, 1651–1664. https://doi.org/10.1016/j.molp.2015.08.011

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. and Salzberg, S. L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods,* **9**, 357–359. https://doi.org/10.1038/nmeth.1923

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

Magori, S. and Citovsky, V. (2011) Epigenetic control of Agrobacterium T-DNA integration. *Biochim Biophys Acta,* **1809**, 388–394. https://doi.org/10.1016/j.bbagrm.2011.01.007

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

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

Ogawa, Y., Dansako, T., Yano, K., Sakurai, N., Suzuki, H., Aoki, K., Noji, M., Saito, K. and Shibata, D. (2008) Efficient and high-throughput vector construction and

Agrobacterium-mediated transformation of Arabidopsis thaliana suspension-cultured cells for functional genomics. *Plant Cell Physiol,* **49**, 242–250. https://doi.org/10.1093/pcp/pcm181

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

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

Shu, H., Wildhaber, T., Siretskiy, A., Gruissem, W. and Hennig, L. (2012) Distinct modes of DNA accessibility in plant chromatin. *Nat Commun,* **3**, 1281. https://doi.org/10.1038/ncomms2259

Sivanandan, C., Sujatha, T. P., Prasad, A. M., Resminath, R., Thakare, D. R., Bhat, S. R. and Srinivasan (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

Soria, G., Polo, S. E. and Almouzni, G. (2012) Prime, repair, restore: the active role of chromatin in the DNA damage response. *Mol Cell,* **46**, 722–734. https://doi.org/10.1016/j.molcel.2012.06.002

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**, 1007–1020. 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

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

To, T. K., Kim, J. M., Matsui, A., Kurihara, Y., Morosawa, T., Ishida, J., Tanaka, M., Endo, T., Kakutani, T., Toyoda, T., Kimura, H., Yokoyama, S., Shinozaki, K. and Seki, M. (2011) Arabidopsis HDA6 regulates locus-directed heterochromatin silencing in cooperation with MET1. *PLoS Genet,* **7**, e1002055. https://doi.org/10.1371/journal.pgen.1002055

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

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J. and Pachter, L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol,* **28**, 511–515. https://doi.org/10.1038/nbt.1621

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

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

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W. and Liu, X. S. (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol,* **9**, R137. https://doi.org/10.1186/gb-2008-9-9-r137

FIGURE LEGENDS

Figure 1. Massively parallel promoter-trapping analysis of the *Arabidopsis thaliana* genome. (**a**) Genomic positions of the inserted promoterless *LUC* genes and respective transcription levels. The bars represent the 4,504 mapped *LUC* genes regarding their orientation towards the upper (+) or bottom (–) DNA strands of the five *A. thaliana* chromosomes. The colour scheme discriminates *LUC* genes according to their expression levels. (**b**) Relative abundance of the *LUC* gene insertion types relative to the annotated gene locations. (**c**) Percentage of transcribed *LUC* genes within the respective insertion types. S and AS indicate the sense and antisense orientations, respectively. (**d**) Distribution profiles of the *LUC* genes of respective insertion types according to the transcription level, with the total frequency of each insertion type normalized to be 100%. The light-blue area indicates the superposed fraction in the genic-sense insertion type, which accounted for ~10% of all transcribed *LUC* genes. (**e**) Abundance of the *LUC* genes with the indicated transcription levels in relation to the distance from the genic TSS. (**f**) Expression profiles of the *LUC* genes in the four regions indicated in (e). The transcription levels of the *LUC* genes are as in (d). (**g**) Classification of the transcribed *LUC* genes according to their insertion types, as in (b).

Figure 2. Transcription states of the *LUC* loci in WT and transgenic cells. (**a**) The 4,504 *LUC* loci were clustered into four groups according to the combination of on/off transcription states in WT and transgenic cells. The local transcription landscape in WT cells was determined based on the RNA-Seq analysis. (**b**) Comparison of the transcription levels between WT and transgenic cells for the *LUC* loci that were transcribed in both WT and transgenic cells.

Figure 3. Transcription states of the *LUC* genes in the heterochromatic regions. (**a**) The upper panel shows the transcription profile of the *LUC* genes in the heterochromatic regions. The middle and bottom panels are derived from Figure 1d and represent the transcription profiles of the genic-sense type and all of the *LUC* genes, respectively. H3K9me2-marked heterochromatic regions covered 18.6 Mb in total and accounted for ~15.6% of the genome, where 120 *LUC* genes were inserted. About 80% of the H3K9me2-marked regions lay within the pericentromere. (**b**) Transcription levels of the *LUC* genes relative to the increased enrichment of H3K9me2. The

transcription levels and H3K9me2 enrichment are both shown as percentiles based on all of the *LUC* genes located in the H3K9me2-marked heterochromatic regions.

Figure 4. Transcriptional states of neighbouring *LUC* insert pairs located in close proximity. (**a**) *LUC* pairs inserted in close proximal chromosomal regions were used for integration-neighbourhood analysis. (**b**) Three possible fates of the transcription of *LUC* pairs: Fate A, expression of both *LUC* genes; Fate B, expression of one *LUC* gene and silencing of the other; and Fate C: silencing of both *LUC* genes. (**c** and **d**) Expected ratio of the three transcriptional fates classified in (b) for *LUC* pairs obeying (**c**) locus-dependent activation or (**d**) integration-dependent stochastic activation. (**e**) Transcriptional states of neighbouring *LUC* pairs inserted in the different cells. The distances between each neighbouring *LUC* insert were <50 bp (upper panel, n = 21) and less than 10 bp (lower panel, n = 12). (**f**) Measured and expected number of *LUC* pairs with Fate A, Fate B, and Fate C, as described in (e). The expected number was calculated according to the integration-dependent activation mechanism.

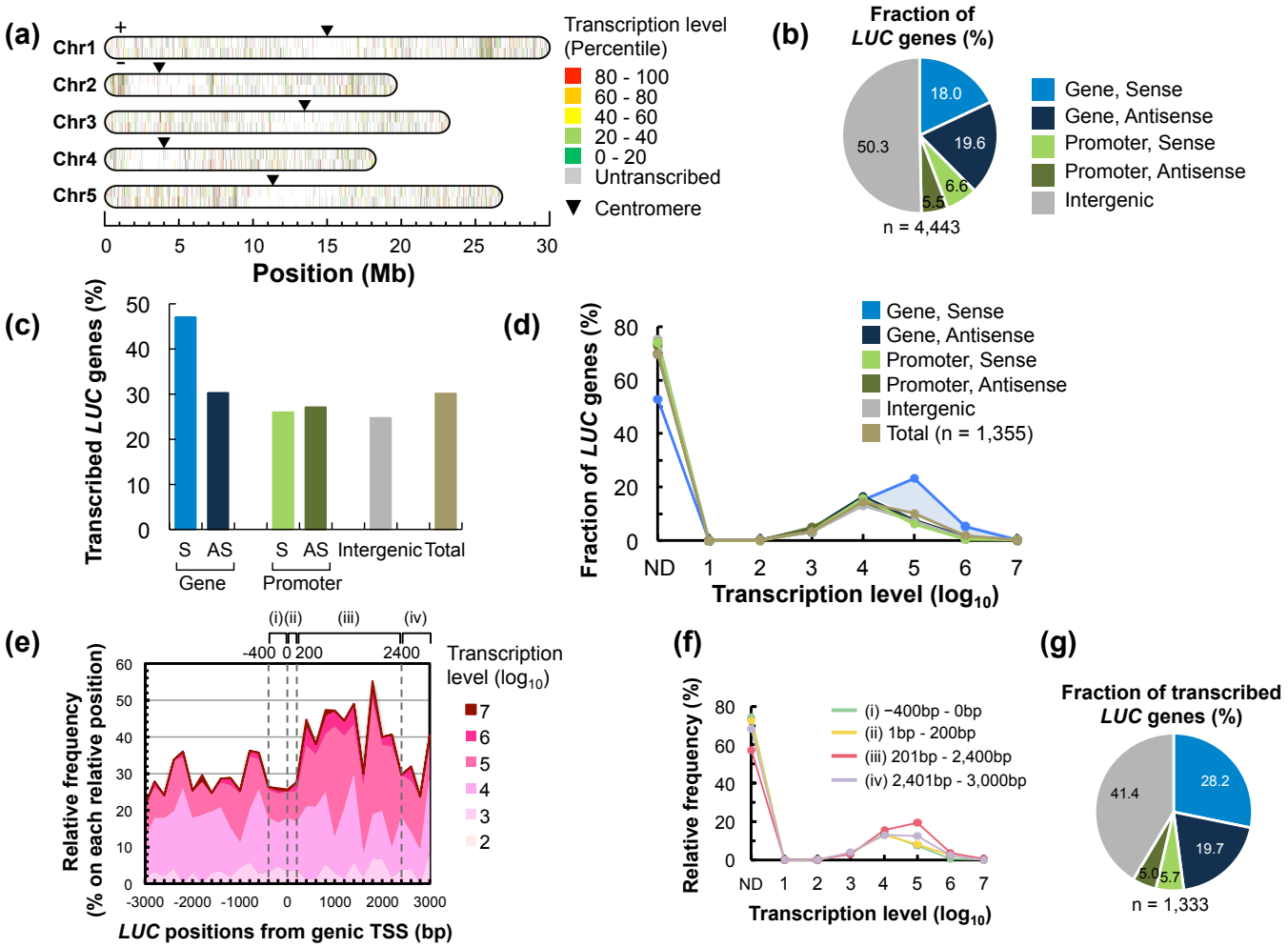


Figure 1. Massively parallel promoter-trapping analysis of the *Arabidopsis thaliana* genome. (**a**) Genomic positions of the inserted promoterless *LUC* genes and respective transcription levels. The bars represent the 4,504 mapped *LUC* genes regarding their orientation towards the upper (+) or bottom (–) DNA strands of the five *A. thaliana* chromosomes. The colour scheme discriminates *LUC* genes according to their expression levels. (**b**) Relative abundance of the *LUC* gene insertion types relative to the annotated gene locations. (**c**) Percentage of transcribed *LUC* genes within the respective insertion types. S and AS indicate the sense and antisense orientations, respectively. (**d**) Distribution profiles of the *LUC* genes of respective insertion types according to the transcription level, with the total frequency of each insertion type normalized to be 100%. The light-blue area indicates the superposed fraction in the genic-sense insertion type, which accounted for ~10% of all transcribed *LUC* genes. (**e**) Abundance of the *LUC* genes with the indicated transcription levels in relation to the distance from the genic TSS. (**f**) Expression profiles of the *LUC* genes in the four regions indicated in (e). The transcription levels of the *LUC* genes are as in (d). (**g**) Classification of the transcribed *LUC* genes according to their insertion types, as in (b).

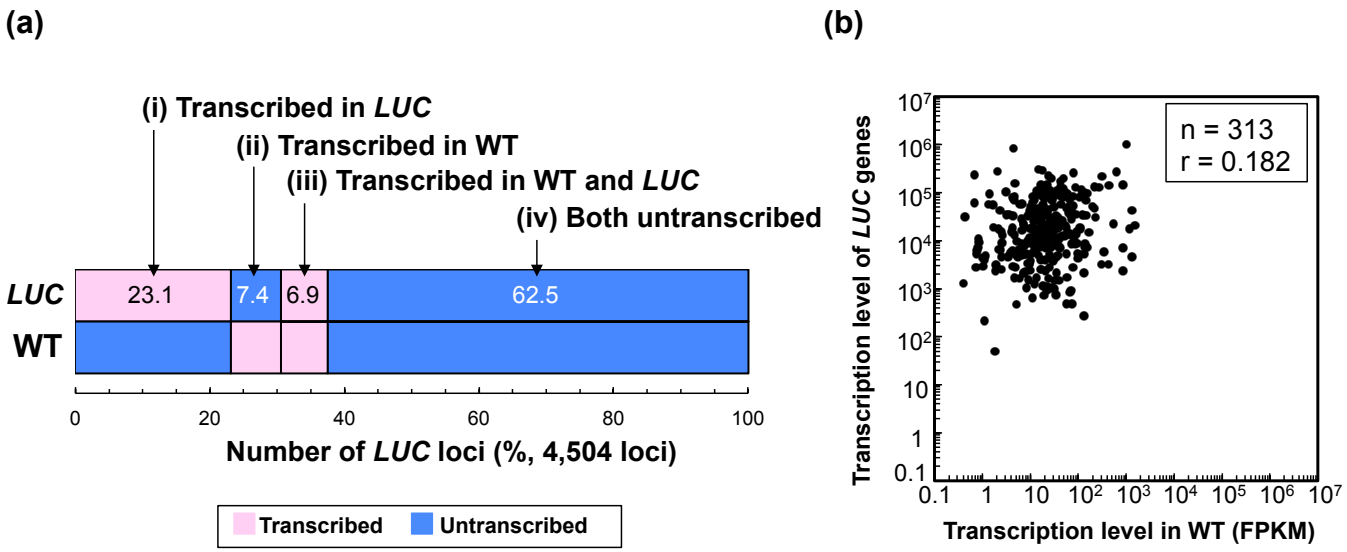


Figure 2. Transcription states of the *LUC* loci in WT and transgenic cells. (**a**) The 4,504 *LUC* loci were clustered into four groups according to the combination of on/off transcription states in WT and transgenic cells. The local transcription landscape in WT cells was determined based on the RNA-Seq analysis. (**b**) Comparison of the transcription levels between WT and transgenic cells for the *LUC* loci that were transcribed in both WT and transgenic cells.

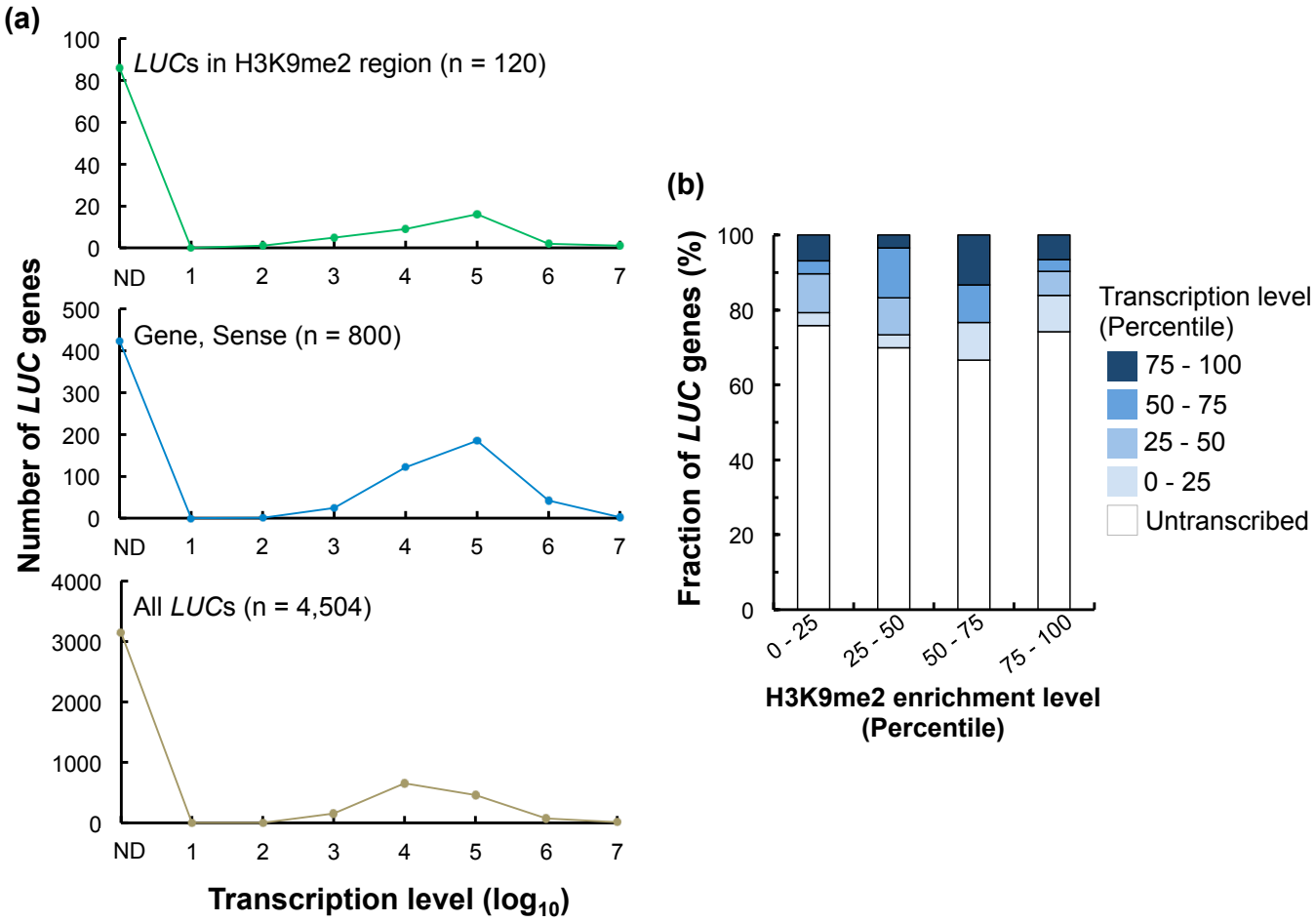
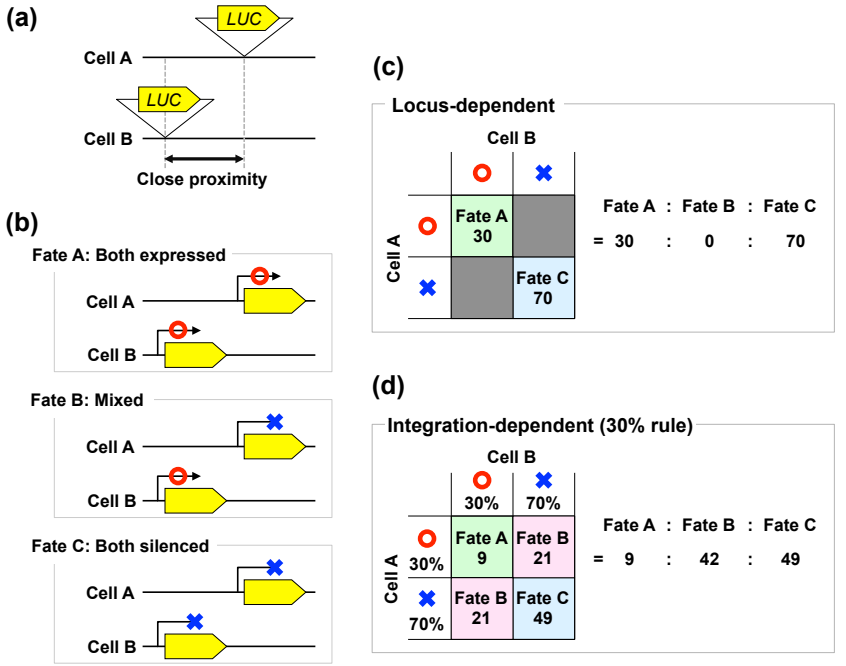
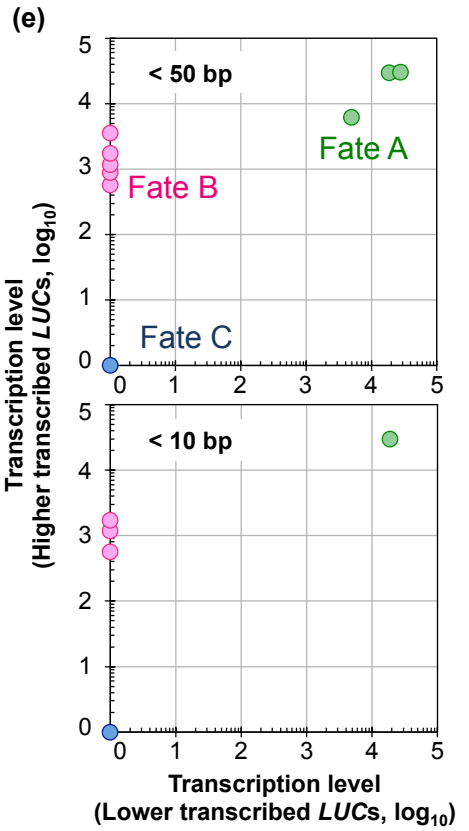


Figure 3. Transcription states of the *LUC* genes in the heterochromatic regions. (**a**) The upper panel shows the transcription profile of the *LUC* genes in the heterochromatic regions. The middle and bottom panels are derived from Figure 1d and represent the transcription profiles of the genic-sense type and all of the *LUC* genes, respectively. H3K9me2-marked heterochromatic regions covered 18.6 Mb in total and accounted for ~15.6% of the genome, where 120 *LUC* genes were inserted. About 80% of the H3K9me2-marked regions lay within the pericentromere. (**b**) Transcription levels of the *LUC* genes relative to the increased enrichment of H3K9me2. The transcription levels and H3K9me2 enrichment are both shown as percentiles based on all of the *LUC* genes located in the H3K9me2-marked heterochromatic regions.





(f)

	Distance between two *LUC*s <50 bp		Distance between two *LUC*s <10 bp	
	Measured	Expected (30% rule)	Measured	Expected (30% rule)
Fate A	3	1.9	1	1.1
Fate B	5	8.8	3	5.0
Fate C	13	10.2	8	5.9
Total	21	21	12	12
Fischer's exact test	$P = 0.55$		$P = 0.82$	

Figure 4. Transcriptional states of neighbouring *LUC* insert pairs located in close proximity. (**a**) *LUC* pairs inserted in close proximal chromosomal regions were used for integration-neighbourhood analysis. (**b**) Three possible fates of the transcription of *LUC* pairs: Fate A, expression of both *LUC* genes; Fate B, expression of one *LUC* gene and silencing of the other; and Fate C: silencing of both *LUC* genes. (**c** and **d**) Expected ratio of the three transcriptional fates classified in (b) for *LUC* pairs obeying (**c**) locus-dependent activation or (**d**) integration-dependent stochastic activation. (**e**) Transcriptional states of neighbouring *LUC* pairs inserted in the different cells. The distances between each neighbouring *LUC* insert were <50 bp (upper panel, n = 21) and less than 10 bp (lower panel, n = 12). (**f**) Measured and expected number of *LUC* pairs with Fate A, Fate B, and Fate C, as described in (e). The expected number was calculated according to the integration-dependent activation mechanism.

Methods S1.

Construction of barcoded plasmid libraries.

The transformation vector plasmid was constructed using a modified pGreenII vector (Hellens *et al.*, 2000; Hirashima *et al.*, 2006) to encode a promoter-less reporter cassette and a kanamycin-resistant cassette between the right (RB) and left border (LB) (Figure S1a). The reporter cassette consisted of a 12-base random barcode sequence, the firefly luciferase (LUC^+) gene, and a *nos* terminator sequence, and contained the short sequence 5′–AGGCCTCGAGGTTATCAGCTTACAG–3′ (the *Xho*I site is underlined) between the RB and the random barcode. This short sequence was inserted for the sake of introducing the barcode sequence and also for the construction of amplicon-sequencing libraries. The kanamycin-resistant cassette contained the *NptII* gene with a *nos* promoter and *nos* terminator. The LB was modified to be repeated four times (Figure S1a), to suppress the integration of the vector backbone sequence into the plant genome (Kuraya *et al.*, 2004). The modified LB sequence was 5′–ATCCTGCCAGTTACACCACAATATATCCTGCCAGTTACACCACAATATATCCTGCCAGTTACACCACAATATATCCTGCCAGTTACACCACAATATATCCTGCCA–3′, and the first 9 bases were added through the construction step. To obtain a plasmid library that contained the random barcode sequence, the 5′-end fragment of the luciferase gene was amplified using two primers (5′–AAAGTCGACGTTATCAGCTTACAGNNNNNNNNNNNNATGGAAGACGCCAAAAACAT–3′ and 5′–TTAGGTAACCCAGTAGATCCAGAGG–3′ (the *Sal*I site is underlined)), digested with *Sal*I and *Eco*RI (the *Eco*RI site was located on the amplified *LUC* fragment), and inserted into the *Xho*I and *Eco*RI sites of the transformation vector.

The constructed vector was transformed into *Escherichia coli* strain NEB 10-beta (New England Bilabs) by electroporation, and approximately 420,000 transformant colonies were obtained; this number suggests the initial diversity of the barcode clones. The transformed *E. coli* cells were cultured in liquid LB medium and subjected to plasmid DNA extraction.

Mapping of the *LUC* genomic loci.

The high-throughput sequencing data of the mapping libraries were trimmed from the 3′ end using fastq_quality_trimmer (http://hannonlab.cshl.edu/fastx_toolkit/) with a phred-scaled quality score ≥30 and were used for the mapping of *LUC* genes with the aid of open-source software and custom Perl scripts (Figure S1c). In *Agrobacterium*-mediated DNA integration, the 3′-terminal 3 bp of the RB is usually the junction between the T-DNA and the plant genome (Windels, De Buck and Depicker, 2008). Therefore, the transformation vector sequence from the 3′-terminal 3 bp of the RB to the ATG initiation codon of *LUC* was used as the *LUC* segment, to search for the *LUC* flanking genomic sequences. The searching methods were slightly different between the two types of mapping libraries. (1) In the Nextera-based libraries, both paired-end reads were used to obtain *LUC* flanking sequences. Sequenced reads that included the *LUC* segment plus more than 25 bp of its flanking sequence were screened, and the flanking sequences and their corresponding *LUC* barcodes were extracted. The flanking sequences obtained were trimmed up to 50 bp using fastx_trimmer (http://hannonlab.cshl.edu/fastx_toolkit/). (2) In the tailed-PCR libraries, only forward reads from the paired-end reads were used for the extraction of *LUC* flanking sequences, and their 3′-terminal 157 bp segments were removed using fastx_trimmer. From the obtained reads, 25-bp flanking sequences of the *LUC* segments

and their corresponding barcodes were extracted as described above.

The flanking sequences obtained above were mapped to the *Arabidopsis* genome of TAIR10 version using Bowtie (Langmead *et al.*, 2009), on the condition that, at most, three mismatches were allowed and that individual sequences were associated with a unique genomic locus (Bowtie settings, -m 1, -v 3). Subsequently, the 3′-junction sites of the mapped flanking sequences were defined as the genomic loci of the corresponding *LUC* insertion sites. We also applied the following rules: 1) *LUC* genes that mapped at a single locus but for which the sequence reads were less than 3 were discarded and 2) cases in which very similar barcodes were mapped to identical loci (data not shown) suggested that an error occurred in the high-throughput sequencing of the barcode. Therefore, barcodes that occupied more than 90% of the reads at their respective genomic loci were retained, and their *LUC* genes were mapped to the respective loci.

Finally, we combined all *LUC* loci that were derived from three biologically independent TRIP pools, as well as from two kinds of mapping libraries, and subjected them to the following analyses.

Determination of the transcription levels of *LUC* genes.

Bioinformatics analyses of the *LUC* expression data were performed using a custom Perl script, the Microsoft Excel software, and the R package (http://www.R-project.org).

To compare the expression levels of individual *LUC* genes, their relative transcript levels were determined as follows (Figures S1b and d). The cDNA and DNA libraries that were specifically prepared for the expression analysis were subjected to amplicon sequencing on an

Illumina MiSeq sequencer with 76 bp pair-end reads. The barcode sequences obtained were verified by the corresponding reads of each pair-end. The read numbers of each barcode sequences was counted in each sequencing library. We should note that, after sequencing on the MiSeq apparatus, *LUC* genes with a DNA read number ≤5 were omitted from the subsequent analysis. Besides, when the cDNA read number was ≤5, the transcript levels of the *LUC* genes were set as zero. Subsequently, the cDNA and DNA read numbers of the individual barcodes were normalized to the total cDNA and DNA read numbers of all barcodes, respectively. Then, the normalized cDNA barcode number was divided by the corresponding normalized DNA barcode number, to give an indicator of the RNA level per DNA molecule. This indicative number was multiplied by 10,000 and was used to indicate the transcription levels of the individual *LUC* genes. Obtained transcription levels of each *LUC* gene were then assigned to individual insertion loci described above according to the barcode sequences. *LUC* loci were omitted from subsequent analysis when transcription levels were not assigned.

References

Crooks, G. E., Hon, G., Chandonia, J. M. and Brenner, S. E. (2004) WebLogo: a sequence logo generator. *Genome Res,* 14, 1188-1190.

Hellens, R. P., Edwards, E. A., Leyland, N. R., Bean, S. and Mullineaux, P. M. (2000) pGreen: a versatile and flexible binary Ti vector for Agrobacterium-mediated plant transformation. *Plant Mol Biol,* 42, 819-832.

Hirashima, M., Satoh, S., Tanaka, R. and Tanaka, A. (2006) Pigment shuffling in antenna systems achieved by expressing prokaryotic chlorophyllide a oxygenase in Arabidopsis. *J Biol Chem,* 281, 15385-15393.

Kuraya, Y., Ohta, S., Fukuda, M., Hiei, Y., Murai, N., Hamada, K., Ueki, J., Imaseki, H. and Komari, T. (2004) Suppression of transfer of non-T-DNA 'vector backbone' sequences by

multiple left border repeats in vectors for transformation of higher plants mediated by *Agrobacterium tumefaciens*. *Molecular Breeding,* 14, 309–320.

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.

Windels, P., De Buck, S. and Depicker, A. (2008) Agrobacterium Tumefaciens-Mediated Transformation: Patterns of T-DNA Integration Into the Host Genome. In *Agrobacterium: From Biology to Biotechnology* (Tzfira, T. and Citovski, V., eds.). Springer, New York, NY, pp. 441–481.

Yamamoto, Y. Y., Ichida, H., Matsui, M., Obokata, J., Sakurai, T., Satou, M., Seki, M., Shinozaki, K. and Abe, T. (2007) Identification of plant promoter constituents by analysis of local distribution of short sequences. *BMC Genomics,* 8, 67.

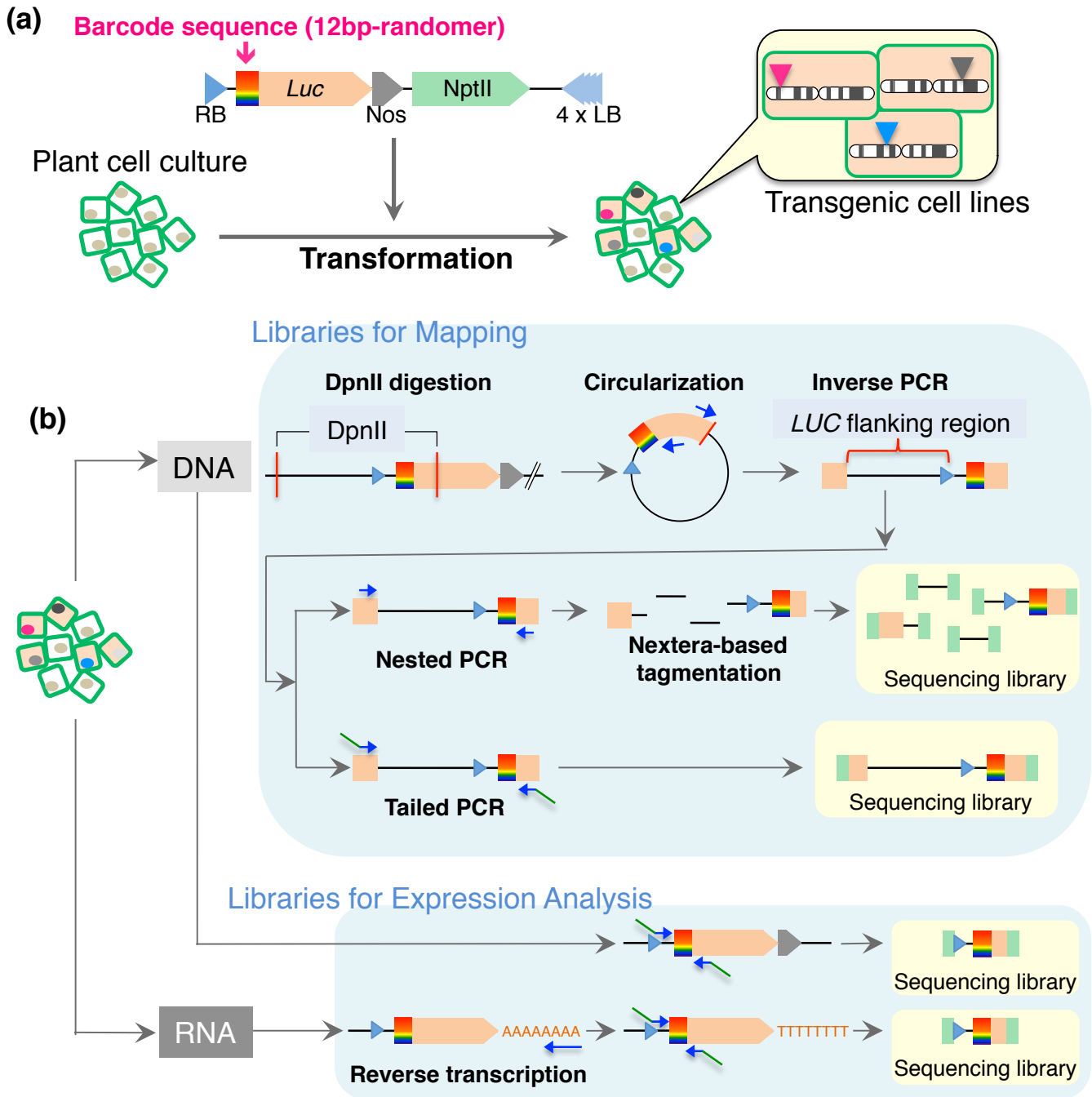


Figure S1. Precise workflow of the promoter analysis that was performed using the TRIP system. (**a**) Transformation of multiplexed barcoded vectors into *Arabidopsis* T87 suspension-cultured cells. (**b**) Preparation of sequencing libraries for the mapping and expression analyses. To prepare the mapping libraries, two different methods were employed after inverse PCR. In the first method, nested PCR products were fragmented and tagged with sequencing adapters using a Nextera-based method. In another method, inverse PCR products were subjected to tailed PCR, to add the sequencing adapters. To prepare libraries for the expression analysis, the barcode regions of both cellular DNA and cDNA were PCR amplified, followed by the addition of sequencing adapters using tailed PCR. cDNAs were prepared via an oligo(dT)-primed RT reaction. The libraries obtained were applied to a high-throughput sequencing analysis. (**c**) Workflow of the data-analysis pipeline that was used for the mapping of *LUC* genes. The flanking sequences of the *LUC* genes were extracted from the Nextera-based mapping library and tailed-PCR-based mapping library using slightly different methods. The *LUC* loci obtained were combined in the final step. (**d**) Flow diagram used for the determination of *LUC* transcription levels. The transcription level data obtained for individual barcodes were associated with the respective mapped *LUC* genes and used in subsequent analyses.

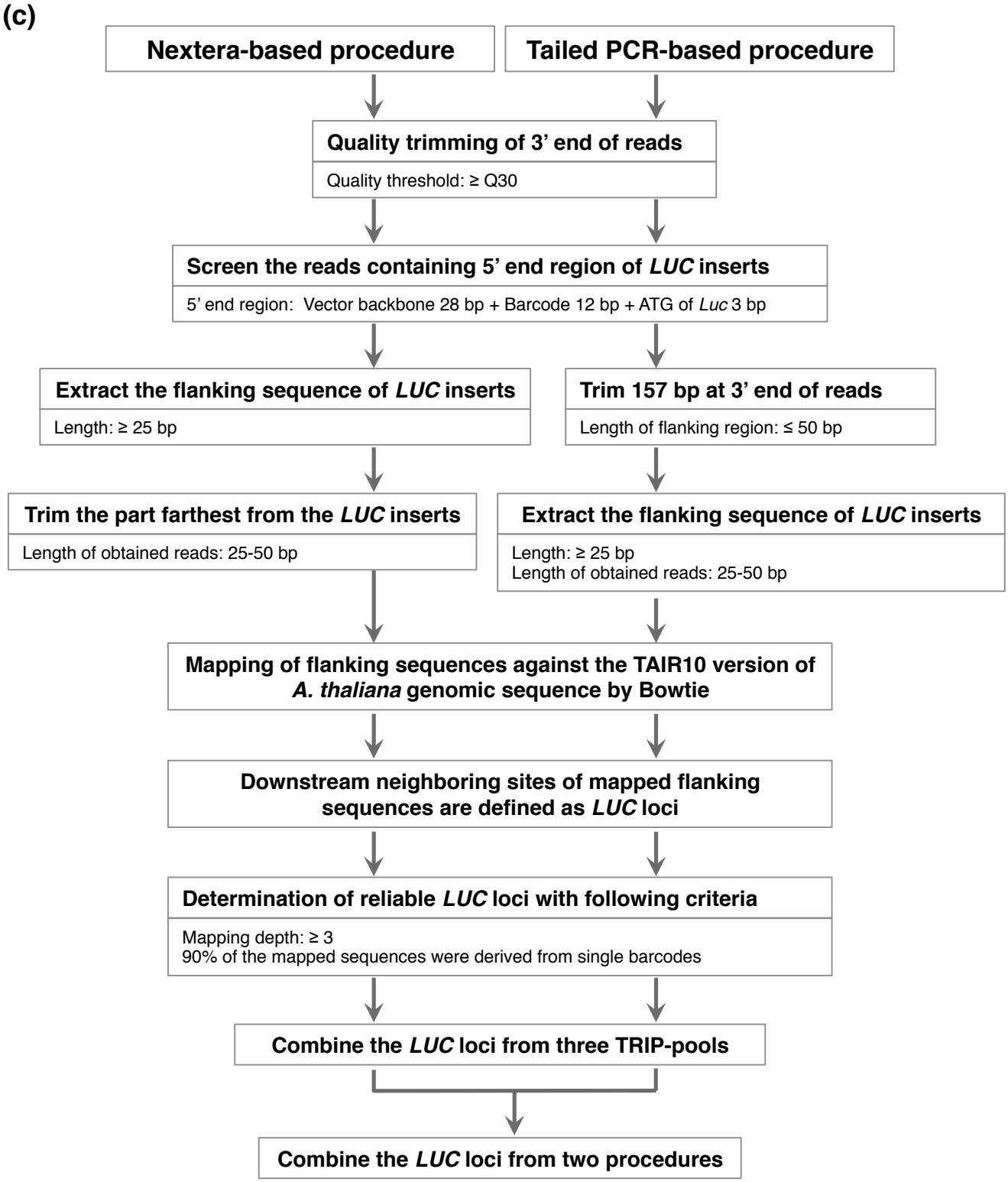


Figure S1. Precise workflow of the promoter analysis that was performed using the TRIP system. (**c**) Workflow of the data-analysis pipeline that was used for the mapping of *LUC* genes. The flanking sequences of the *LUC* genes were extracted from the Nextera-based mapping library and tailed-PCR-based mapping library using slightly different methods. The *LUC* loci obtained were combined in the final step.

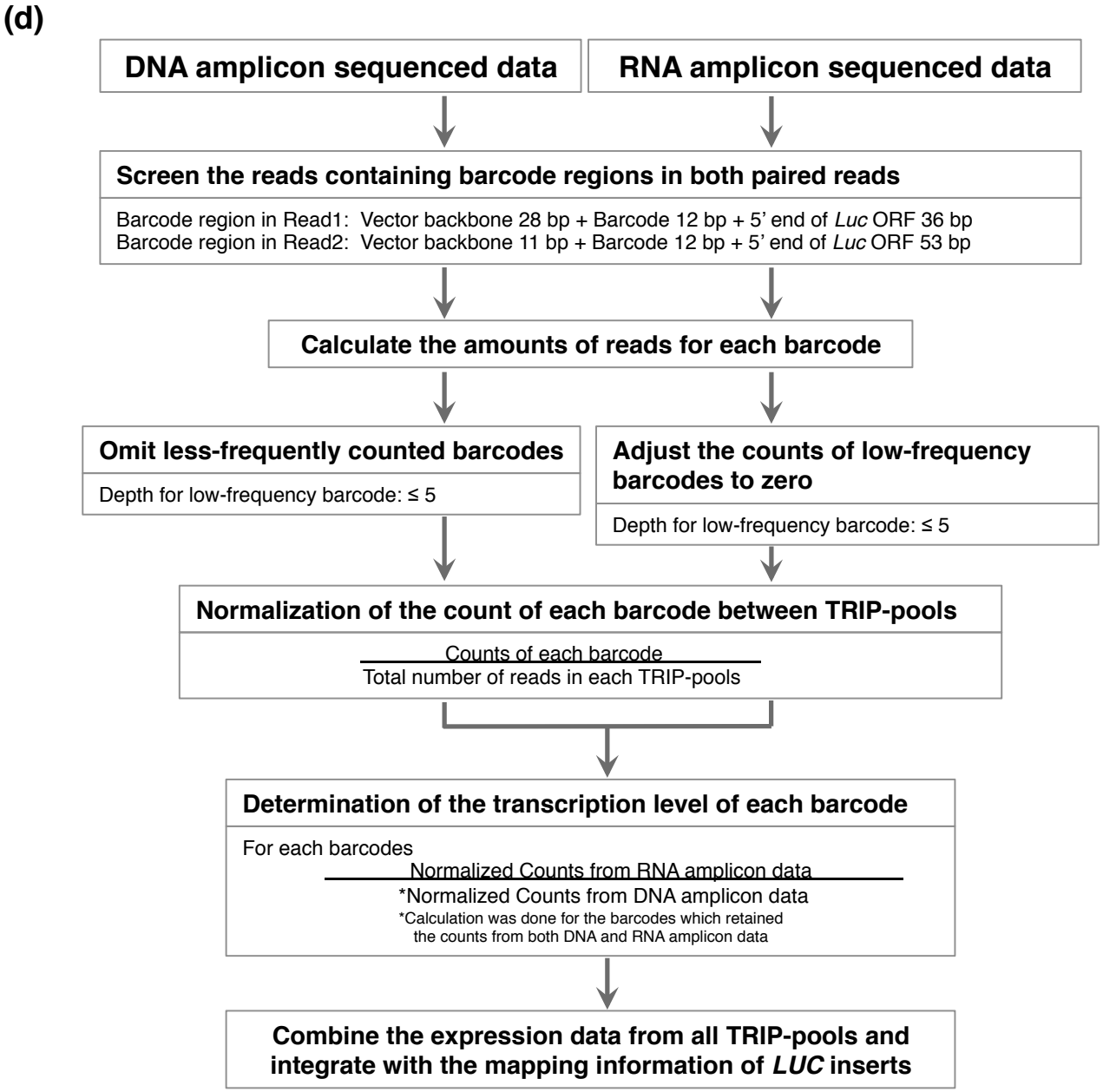


Figure S1. Precise workflow of the promoter analysis that was performed using the TRIP system. (**d**) Flow diagram used for the determination of *LUC* transcription levels. The transcription level data obtained for individual barcodes were associated with the respective mapped *LUC* genes and used in subsequent analyses.

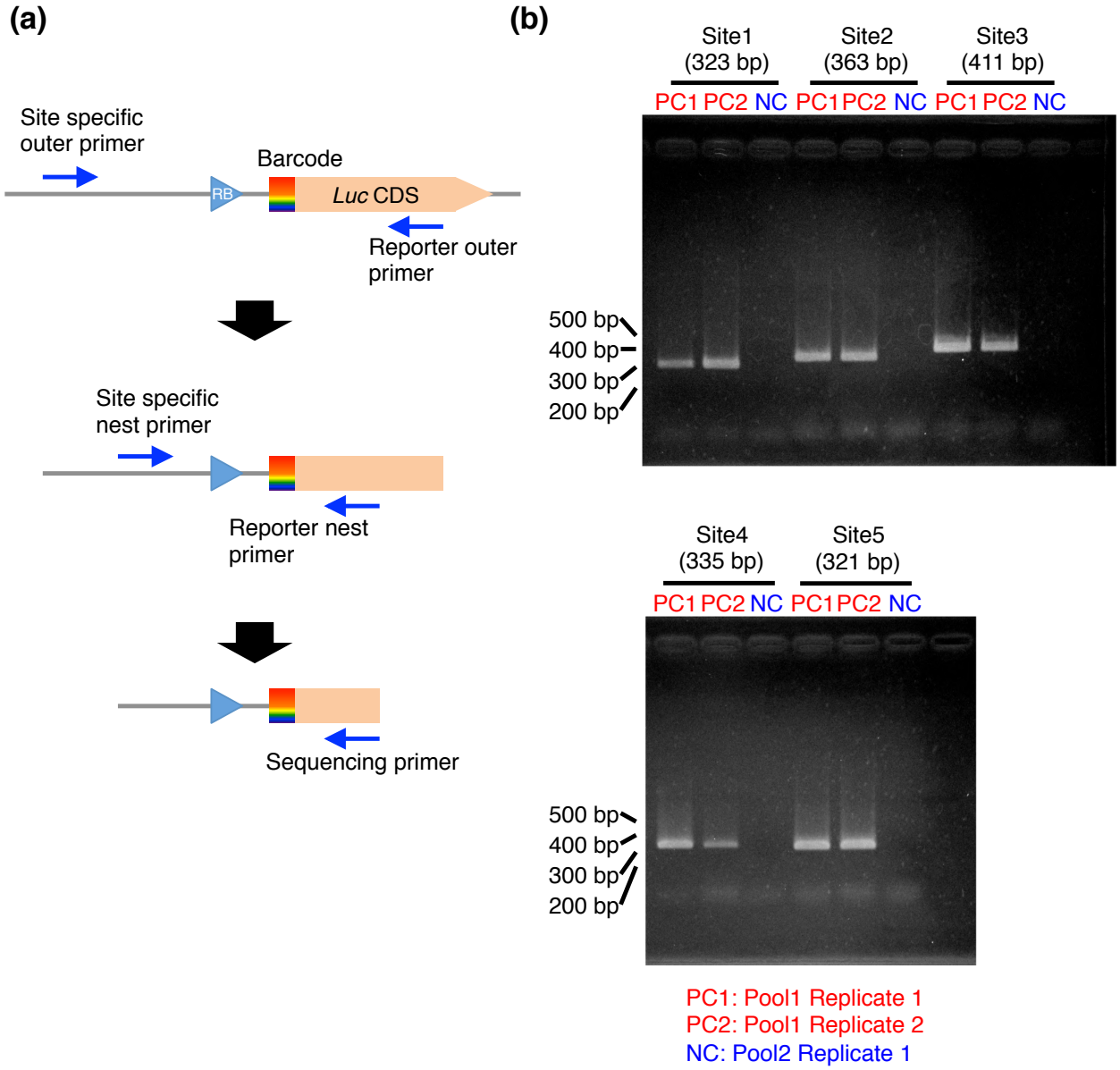


Figure S2. Validation of the *LUC* mapped loci and barcode sequences via PCR amplification in five representative samples. (**a**) Schematic diagram of the nested PCR that was performed using insertion-site-specific and *LUC*-specific primers. (**b**) Five *LUC* genes were chosen from the TRIP-Pool1 and detected by PCR. PC1 and PC2 are technical replicates of the PCR using the template DNA from TRIP-Pool1 cells. NC is the PCR product from the DNA of TRIP-Pool2 and was used as a negative control. The PCR products were loaded onto a 2% agarose gel. The expected size of the PCR products is shown at the top of the gel, in parentheses. The PCR products obtained were Sanger sequenced for verification of the barcode sequences.

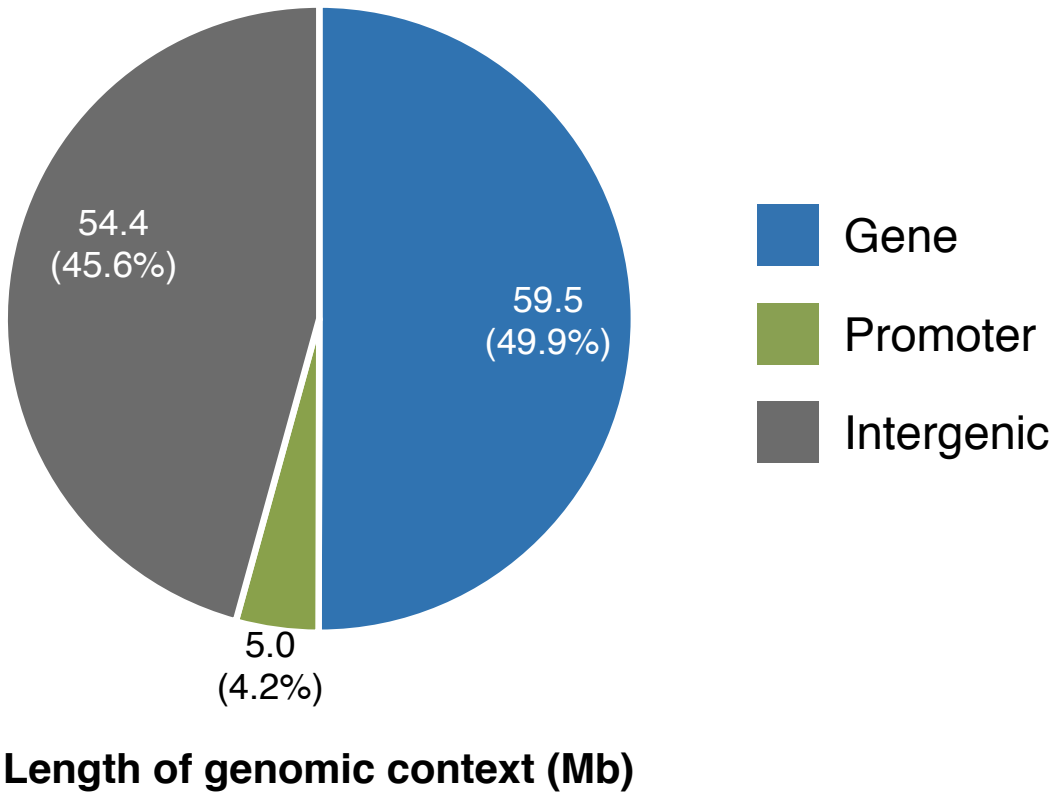


Figure S3. Length of each genomic context. The total length of the respective genomic contexts and their percentage in the whole genome are shown. The 200 bp segments 5′-proximal to the genic region (CDS plus UTR regions according to TAIR10) were defined as promoter regions, and the remaining sequences were defined as intergenic regions. When neighboring promoter and genic regions were overlapped, those parts were omitted from the statistical analyses described above (their sum was 0.23 Mb, 0.2% of the whole genome).

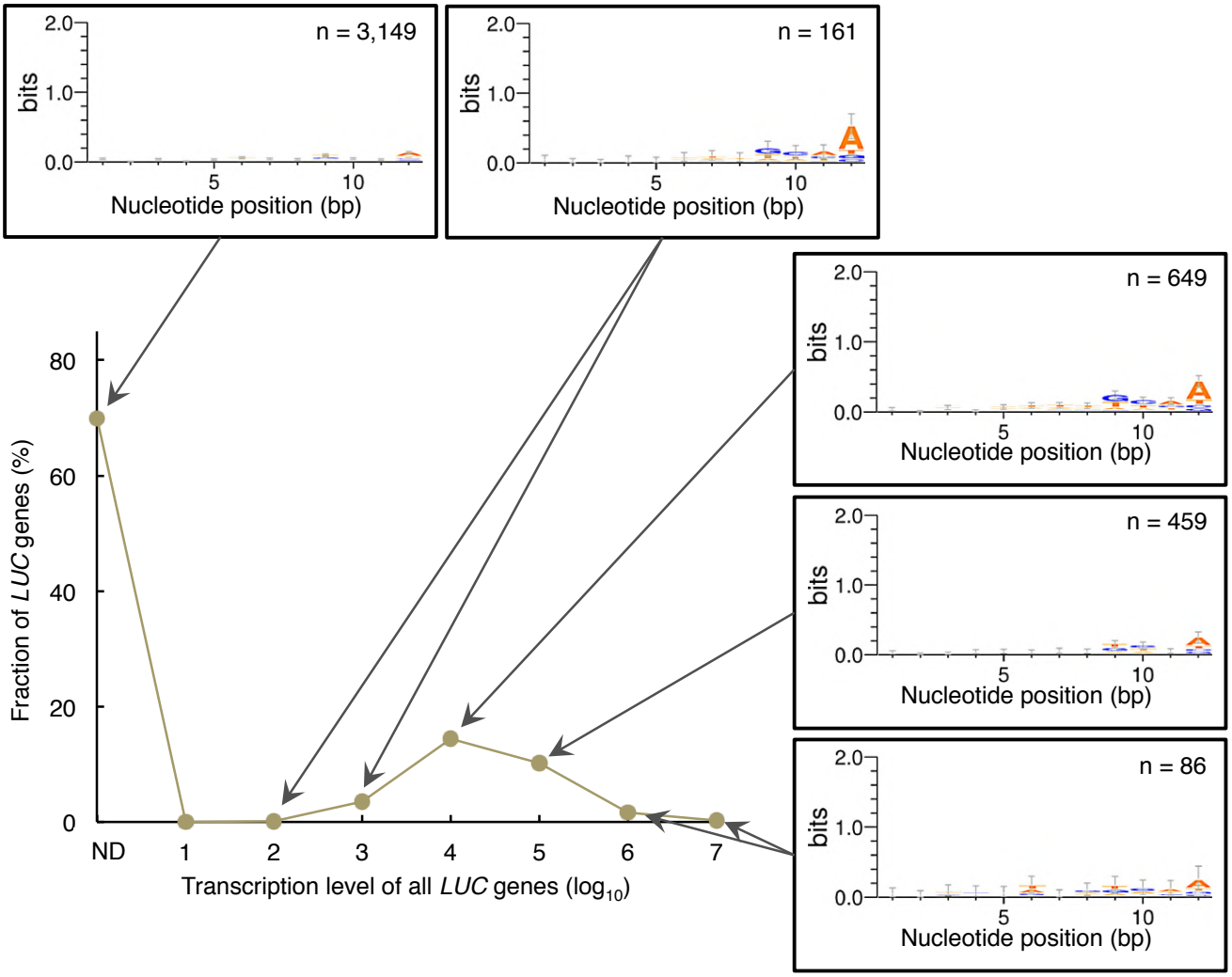


Figure S4. Assessment of the effect of barcode sequences on the LUC transcription levels. Frequently observed barcode motifs in the LUC insert of indicated transcription levels were analyzed using WebLogo3 (Crooks *et al.*, 2004). The transcription levels of all the LUC genes are shown as in Figure 1d. A weak positional preference for 'A' was found at the 3′-terminal position on the barcode. However, the frequency of 'A' at this position did not correlate with the strength of transcription.

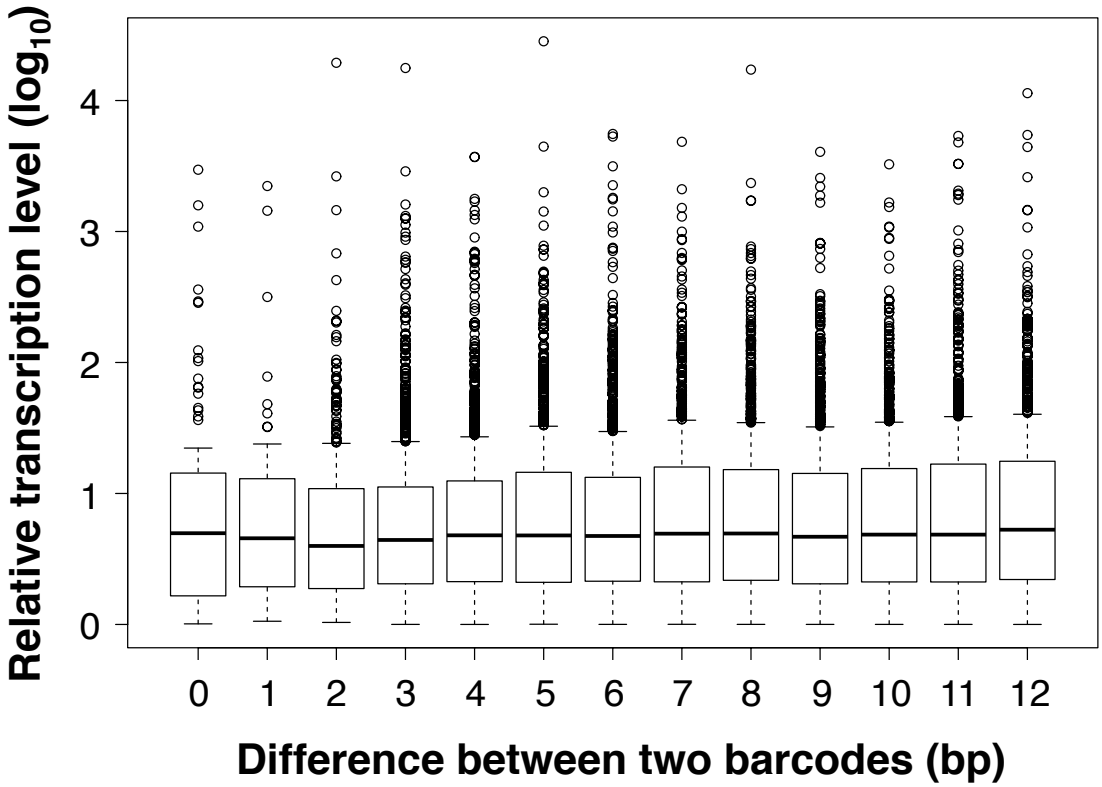


Figure S5. Similarity/dissimilarity of the transcription levels of the randomly selected LUC pairs against the sequence identity of the 12-base barcode. A pair of LUC genes was randomly selected from the 4,504 mapped LUC genes, and the similarity/dissimilarity of their transcription levels is shown as the ratio of their RNA levels in a logarithmic scale; the ratio was calculated by dividing the higher RNA level by the lower level (i.e., log(ratio) ≥0). The similarity/diversity of the barcode is indicated by the number of mismatched nucleotides at the corresponding positions. This graph is the summary of the analysis of 10,566 LUC pairs and indicates the absence of a correlation between the similarity of the barcode sequence and that of the transcription level. In other words, the barcode sequence does not affect the transcription level of LUC genes.

Methods note: 1) When randomly selected LUC pairs were located within 100 kb on the same chromosome, they were omitted from the analysis, lest their positional effect should influence their transcription levels. 2) One thousand LUC pairs were analyzed each for the indicated number of mismatches in the barcode. However, for mismatch numbers of 0, 1, and 2, the number of LUC pairs analyzed was 92, 51, and 423, respectively. This is because the number of such highly homologous barcodes in the total population of 4,504 LUC inserts was limited, and these are all the LUC genes that fulfilled the given requests. 3) The LUC inserts of the identical barcodes were derived from different TRIP pools, because LUC mapping in a given TRIP pool had been conducted so that the individual LUC genes were mapped to a unique locus, with omission of those that were mapped to more than one locus.

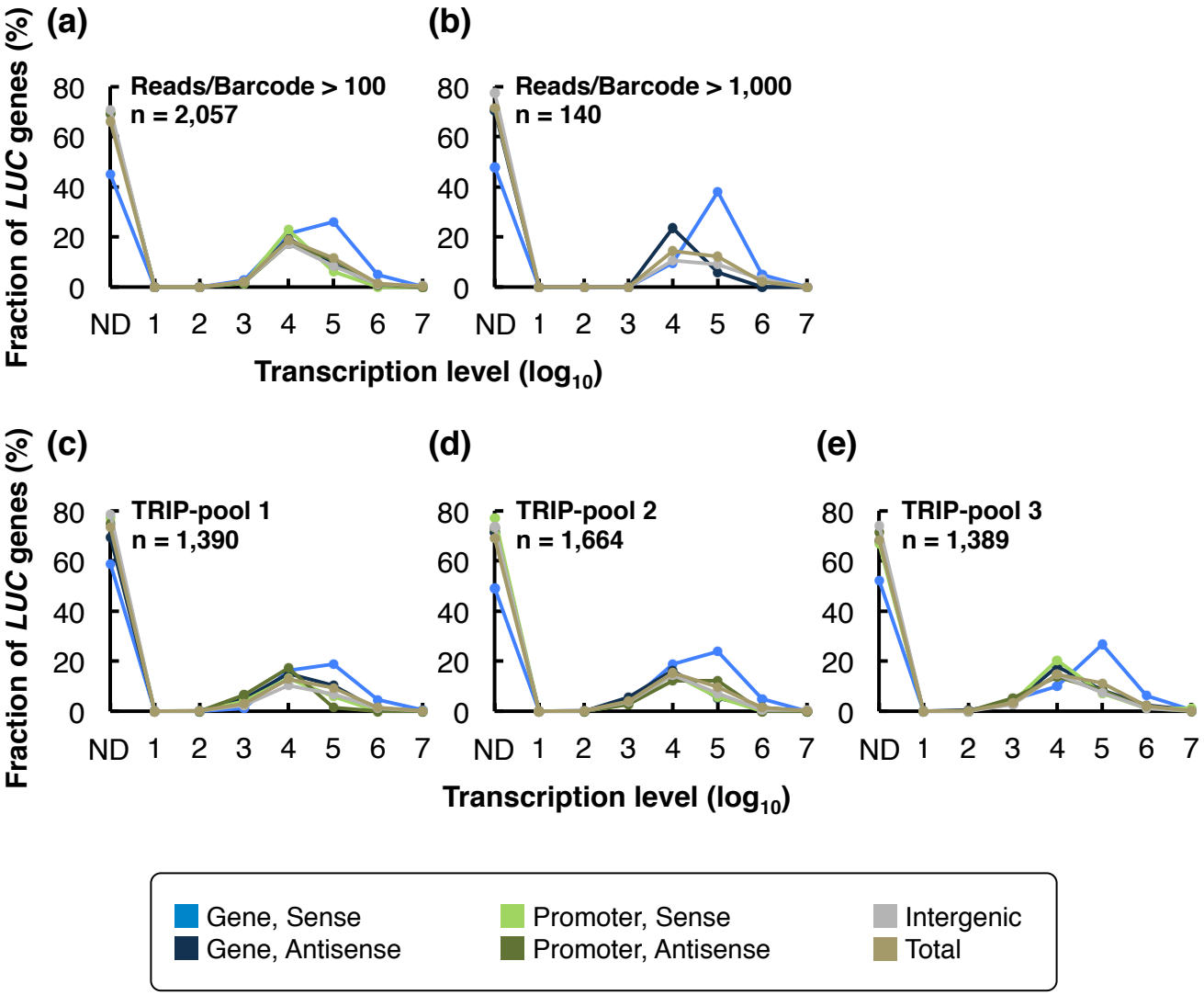


Figure S6. Expression profiles of *LUC* genes with high-number reads from the amplicon-sequencing data and of *LUC* genes from biological replicates. (**a** and **b**) For each barcode, when the number of reads from DNA amplicon sequencing was up to (**a**) 100 or (**b**) 1,000, the barcode was omitted from the analysis. The number of reads for each barcode obtained from RNA amplicon sequencing was redefined as zero, if the number of reads was below such thresholds. The subsequent processes used in this analysis were same as those used in Figure 1d. The expression profiles of the *LUC* genes located in promoter regions were omitted from (b), because the number of such *LUC* genes was insufficient to represent their profiles. (**c**–**e**) Expression profiles of three biological replicates. The numbers of *LUC* genes shown in all graphs are the total amount of *LUC* genes used for their analysis.

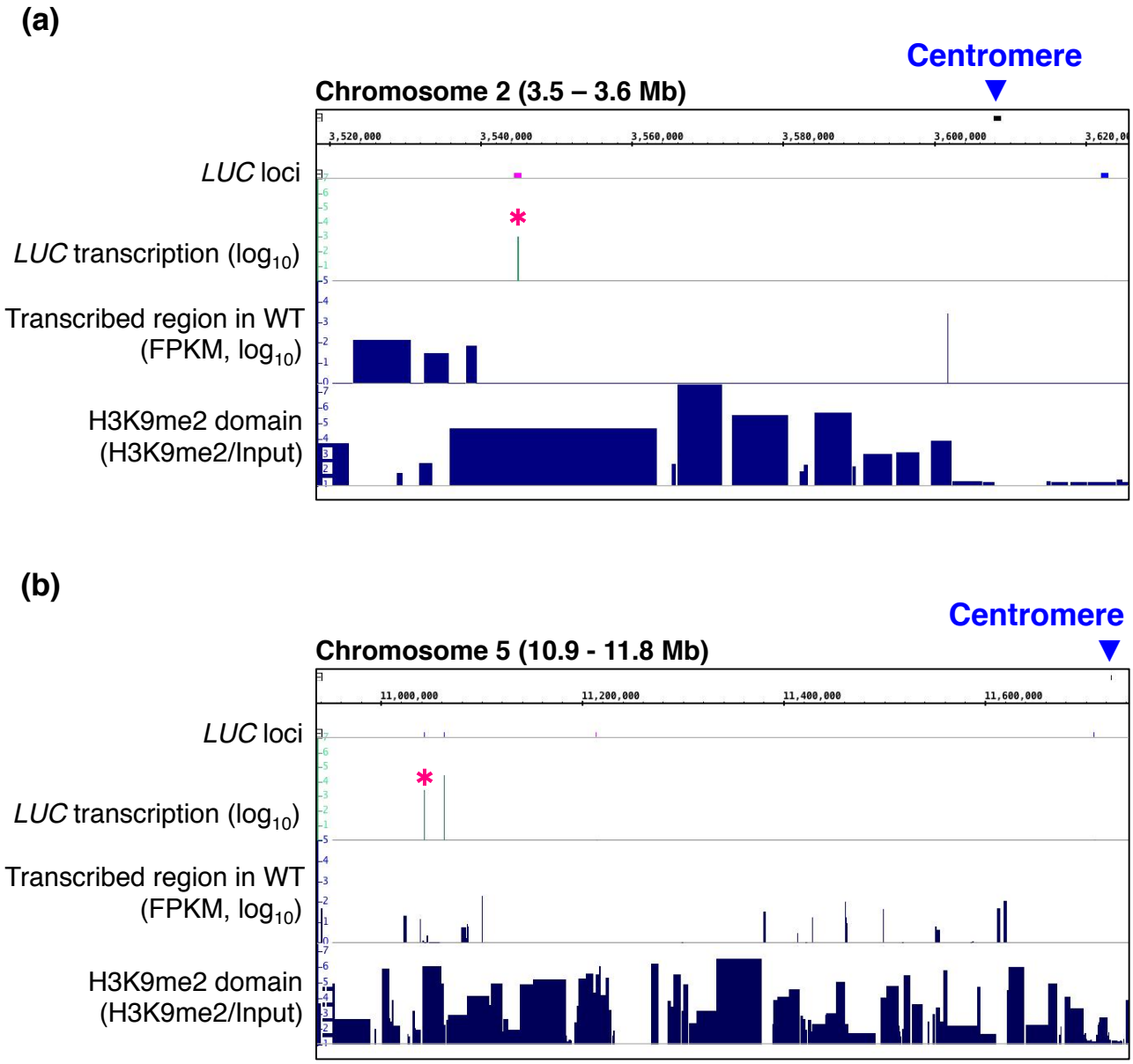


Figure S7. Two examples of transcribed LUC genes in the H3K9me2-marked regions located around the centromere. Transcribed LUC genes (asterisk) were found 63 kb and 682 kb away from the centromeres of chromosomes 2 and 5, respectively. The respective H3K9me2 levels of these loci were 80 and 91 percentiles. In WT T87 cells, transcripts were very scarce in these heterochromatic regions.

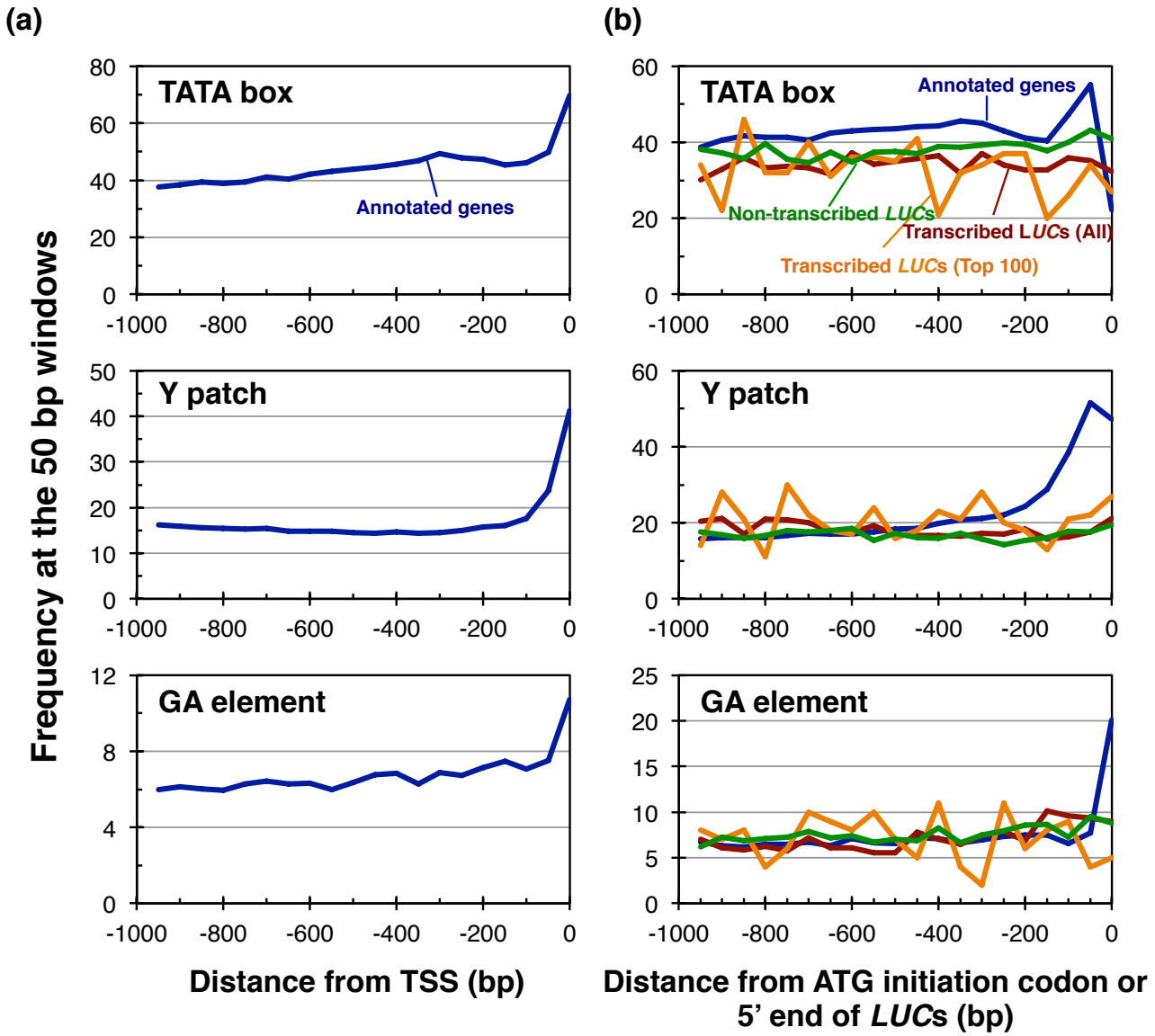


Figure S8. Distribution of *cis*-regulatory elements in the upstream region of *LUC* integration sites. (**a** and **b**) The frequency of TATA-box, Y patch, and GA elements in the upstream region of the (**a**) TSS of annotated genes, or (**b**) of ATG initiation codons of annotated genes and 5′ ends of *LUC* inserts were analyzed according to Yamamoto *et al.* (Yamamoto *et al.*, 2007) using a window size of 50 bp for the high-sensitive detection of the motifs. The Y-axis represents the fraction of genes or *LUC* genes that contained the indicated motifs.

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 (TAILed-PCR based)

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	Primer 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 locus determination (Nextera based)

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_LUC_iPCR_F2.1	CATTTCGCAGCCTACCGTAGTGTTT	Primer set for the nested-PCR following inverse PCR to specifically amplify *LUC*-including DNAs.
TRIP_LUC_iPCR_R2.2	CATTTCGAAGTATTCCGCGTACGTG	

Sequencing library preparation for the transcription level analysis

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

Primer sets for validation of *LUC* mapped loci and barcode

Name	Sequence (5' -> 3')	Descriptions
TRIP_LUC_iPCR_R1	GTTTTCACTGCATACGACGATTCTG	Reporter outer primer (see Figure S2).
TRIP_LUC_iPCR_R2.2	CATTTCGAAGTATTCCGCGTACGTG	Reporter nest primer (see Figure S2).
C1_CGGAAAGACCAA_AS_F1	TCCTCAATGAGTCTGGTGACTTC	Site1 specific outer primer (see Figure S2).
C1_CGGAAAGACCAA_AS_F2	CTCATTGCCCTCAGGTTGGT	Site1 specific nest primer (see Figure S2).
C2_GCACAAAGTCTA_S_F1	TCACTGCTCAATGCGATCTCC	Site2 specific outer primer (see Figure S2).
C2_GCACAAAGTCTA_S_F2	TTAGTGTCGCAACAACGAACCG	Site2 specific nest primer (see Figure S2).
C3_CTAGGGGACTCA_AS_F1	TTCGATCCTTCAAAGCGCATCAC	Site3 specific outer primer (see Figure S2).
C3_CTAGGGGACTCA_AS_F2	CAAGGAGCTTGTCTGGAGAGAG	Site3 specific nest primer (see Figure S2).
N1_TGATGATGTCCA_S_F1	GACTACAAATCATTCATCAACCACG	Site4 specific outer primer (see Figure S2).
N1_TGATGATGTCCA_S_F2	TAGTTGATTCCTCTCGTTCGGC	Site4 specific nest primer (see Figure S2).
T1_TTAGTTGGTCAA_AS_F1	CCAATCTGACACAAAATAGGTCTCT	Site5 specific outer primer (see Figure S2).
T1_TTAGTTGGTCAA_AS_F2	TTAAAGAGGAGTCACGATCATCGGT	Site5 specific nest primer (see Figure S2).

H3K9me2 ChIP validation

Name	Sequence (5' -> 3')	Descriptions
55670F1	CGTTGCTGACGACGGGTTTATGG	Primer set for validation of H3K9me2-ChIP according to To et al., 2011.
55670R1	GTTTCTAGATCCCGCTTCGTCGTTC	
63935F1	CGTTGTAGGTCAGGGTTCTTGC	Primer set for validation of H3K9me2-ChIP according to To et al., 2011.
63935R1	GCCATAGATGCATCACGAACCG	
44070F1	ACTTCCTCGACCTCTTATCTCC	Primer set for validation of H3K9me2-ChIP according to To et al., 2011.
44070R1	CTTCGGTTTAACCCAGAGAGATG	
ACT2F2	GATCTCCAAGGCCGAGTATGAT	Primer set for validation of H3K9me2-ChIP according to To et al., 2011.
ACT2R2	CCCATTCATAAAACCCCAGC	
67105F1	TGTCTCCAGTTTGATCCGGATTTG	Primer set for validation of H3K9me2-ChIP according to To et al., 2011.
67105R1	GTAACAGAAGATCCGATATGTAATCGG	
G683F1	TCCGATCTGAGATCGGTAGCCG	Primer set for validation of H3K9me2-ChIP according to To et al., 2011.
G683R1	CGAAACAAACCCACGACACTCC	