Doctor thesis

Molecular Mechanisms and Genetic Behavior of *De novo*-activated Transcriptions in the *Arabidopsis* Genome

Graduate School of Life and Environmental Sciences

Kyoto Prefectural University

Takayuki Hata

Table of contents

Index

Abbreviations

cDNA: Complementary DNA

CDS: Coding sequence

ChIP: Chromatin immunoprecipitation

ChIP-seq: ChIP sequencing

CV: Coefficient of variation

DNA: Deoxyribonucleic acids

DSB: Double strand break

EGT: Endosymbiotic gene transfer

eRNA: Enhancer RNA

GFP: Green fluorescent protein

GTF: General transcription factor

HGT: Horizontal gene transfer

H2A.Z: Histone 2A.Z

H3K4me1: Lysine (K) 4 mono-methylation of histone H3

H3K4me3: Lysine (K) 4 tri-methylation of histone H3

H3K9me2: Lysine (K) 9 di-methylation of histone H3

H3K27ac: Lysine (K) 27 acetylation of histone H3

H3K36me3: Lysine (K) 36 tri-methylation of histone H3

Inr: Transcriptional initiator

LB: Left border sequence of T-DNA

LINE: Long interspersed nuclear element

LTR: Long terminal repeat

LUC: Luciferase

mC: Methylated cytosine

MBD-seq: methyl-CpG binding domain protein-enriched genome sequencing

MPRA: Massively parallel reporter assay

mRNA: messenger RNA

MS: Murashige and Skoog

NGS: Next generation sequencer/sequencing

NHEJ: Nonhomologous end-joining

NMD: nonsense-mediated decay

NPTII: Neomycin phosphotransferase II

ORF: Open reading frame

PCR: Polymerase chain reaction

PIC: Pre-initiation complex

Pol II: RNA polymerase II

qPCR: quantitative PCR

RB: Right border sequence of T-DNA

RNA: Ribonucleic acids

RNA-seq: RNA sequencing

SINE: Short interspersed nuclear element

TBP: TATA binding protein

T-DNA: Transferred DNA (from Ti-plasmid)

TF: Transcription factor

TRIP: Thousands of reporters integrated in parallel

TSS: Transcription start site

TSS-seq: TSS sequencing

UTR: Untranslated region

WT: Wild type

Chapter 1:

General introduction

General Perspective

The concepts of evolution have roots in antiquity. The philosophers in ancient Greek, Rome, and China noticed that an organism takes over some kind of information from its ancestor, and the information changes over time to make the birth of new species (Harris, 1981; Miller, 2008). Such very early thinking developed in the modern theory in the mid-19th century, led by the two remarkable works; the idea of the evolution driven by the natural selection by Charles Darwin (Darwin, 1859), and the introduction of the concept of *gene* by Gregor Mendel (Mendel, 1865). Since then, the process and underlying mechanism by which genetic novelty emerges during the evolution has attracted the interest of biologists in past years.

Studies about gene evolution advanced in the early 20th. Based on the microscopic observation of fly's chromosomes, Muller and Haldane suggested that chromosomal duplication might contribute to the new gene origination (Haldane, 1933; Muller, 1935). This early thinking lately confirmed and expanded into the duplication-diversification model (Ohno, 1970; Ohno, 1972). The model has been widely accepted, and it had been a consensus view that all the genes are derived from ancestral ones, and that *"the probability that a functional protein would appear de novo by random association of amino acids is practically zero"* (Jacob, 1977).

Upon considering studies about gene evolution, DNA sequencing technology is critical. This technology was firstly appeared in the 1970s (Sanger and Coulson, 1975; Sanger, Nicklen and Coulson, 1977; Maxam and Gilbert, 1977). It enables us to compare the sequence of DNA fragments that are derived from an individual organism or different species. In 2000, the pioneering study about gene duplication was published (Lynch and Conery, 2000). The authors utilized whole genomic data available at that time, and compared all translated coding frames to identify duplicated sequences within each genome (Lynch and Conery, 2000). They explored evolutionary 'young' duplicated genes by calculating the mutation rate of duplicated gene pairs (Figure 1.1) (Lynch and Conery, 2000). These methods have still been relevant today in the gene evolution research.

Due to the recent technological advancements on DNA sequencing technology that can read out more than trillions of bases in a couple of days, now researchers can compare the genome, transcriptome, translatome, and epigenome among closely related species. These genome-wide analyses provided 'very young' genes whose properties and evolutionary fates are

in concert with previous theoretical predictions of gene evolution, and moreover provided novel finding. One of the biggest findings was the *de novo* gene whose origin was assumed to be non-genic sequences in the ancestor's genome (Johnson *et al.*, 2001; Levine *et al.*, 2006). The finding opened up the new paradigm for gene origination processes; a new gene could emerge not only from pre-existing genetic materials but also *de novo* (Van Oss and Carvunis, 2019).

Whichever the processes by which new genes are originated, the gene must acquirer expression competency. The first step and prerequisite on the gene functionalization is to be transcribed into the RNA. If transcription does not occur, the information on the DNA sequences will never appear in the phenotype. Thus, in the process of gene evolution, it should be emphasized that the importance of acquiring transcriptional competency, namely, a promoter.

In this chapter, I review the current understandings about newly originated genes. Firstly, I review the gene evolution researches and summarize the manners of new gene origination processes in the genome. Secondly, I review the current view of gene promoter; its definition, properties, and evolution. In the third part, I summarize how new (or evolutionary young) genes have acquired their promoters. Finally, I clarify the central questions and aims of this thesis.

Manners of new gene origination

To date, many studies reported newly emerged genes in the organism's genome according to the phylogenetic comparison of genome sequences among closely related species. In this section, I review current understandings about the processes of gene origination events.

Duplication-diversification

The basic concept of *gene duplication* is that the copy of pre-existing gene can acquire new function, while the ancestral one maintains original function (Figure 1.2a). The first model of gene duplication was developed by Muller, based on the optical microscopic observation of chromosomes of *Drosophila* (Haldane, 1933; Muller, 1935). This early thinking lately expanded in the 1970s, and has been accepted widely (Ohno, 1970; Ohno, 1972; Jacob, 1977). Gene duplication occurs not only by DNA-mediated mechanism but also by RNA-mediated mechanism (see *retroposition*). The duplication event occurs either a single-gene or whole-genome scale. Whole-genome duplication played a key role in the evolution and diversification of eukaryotes, especially in plants (Clark and Donoghue, 2018).

The fates of duplicated genes are thought to be diversified into the following three patterns;

Pseudogenization

Losing the coding sequences or expression, duplicated copies will be pseudogenized by mutations, because they will become free from selection pressure against their function. Thus it is thought that the majority of the duplicated genes will be lost from the genome (Ohno, 1970; Ohno, 1972).

Neofunctionalization

Neofunctionalization is the process by which duplicated copy acquires a novel function. Ohno suggested that gene duplication could open up the opportunities in order for a new copy of a

gene to acquire a new function by mutation, fusion, fission, or shuffling, while the ancestral copy can maintain original function (Ohno, 1970; Ohno, 1972).

Subfunctionalization

Subfunctionalization occurs when a gene duplication event occurs on the gene with more than two different functions. After the divergences, the original and duplicated copy might retain a subset of their original ancestral function, independently (Force *et al.*, 1999; Stoltzfus, 1999).

Genomic rearrangements

Genomic rearrangements such as recombination, mutations or transposition of mobile elements could change the recipient genome either in single-nucleotide level or chromosomal level (Eichler, 2001; Özlem *et al.*, 2013). When genomic rearrangements occur within coding sequences, they sometimes result in the creation of novel chimeric coding sequences. In contrast, if they occur within non-coding sequences, they will sometimes cause mis-regulation of the pre-existing gene by disrupting its *cis*-regulatory elements. Such expressional variation can facilitate the formation of *de novo* genes (see *de novo gene*). Importantly, the creation of a new gene mediated by genomic rearrangements is often preceded by gene duplication. As gene duplication could maintain the original gene intact, the probability of the new gene being fixed would be increased (see *Gene duplication*).

The patterns of gene originations through genomic rearrangements are typically classified into following three types;

Gene fusion

Gene fusion is a process that two distinct genes are fused together and become a single transcription unit (Figure 1.2b). All the recombination events such as insertion, deletion, or inversion can cause gene fusion. Gene fusion even occurs between the gene pairs on the distinct chromosomes through meiotic recombination (Li, Qin and Li, 2018; Stewart and Rogers, 2019).

Gene fission

Conversely from gene fusion, new gene structures also could be formed via splitting a single gene mediated by recombination events (Figure 1.2b). The fates of individual genes are as noted above; they will independently gain distinct functions (neofunctionalization/subfunctionalization) or be pseudogenized (see *Gene duplication*).

Exon shuffling

Recombination of exons and domains within either single or multiple genes can emerge new gene structures (Figure 1.2c) (Gilbert, 1978). Exon shuffling is thought to play a major role in the formation of novel protein domains in the eukaryotic genome (Patthy, 1999).

Retroposition

Retroposition is a class of gene duplication mediated by RNA (Kaessmann, Vinckenbosch and Long, 2009; Kaessmann, 2010). In the process of retroposition, firstly a messenger RNA (mRNA) is transcribed from the parental gene. Then, mRNA is reverse transcribed into a complementary DNA (cDNA), and is accidentally inserted into the genome (Figure 1.2d). Thus, the retroposed gene copies usually have clear hallmarks; they lack introns and have a polyadenylated tail. The fates of retroposed copies are basically similar to those of DNA mediated duplicated genes (Kaessmann, 2010).

Remarkably, retroposed genes often lack their original *cis*-regulatory sequences, because of their origination process. However, several manners were reported that retroposed genes acquire their novel transcriptional competency (detailed in further below).

Horizontal gene transfer (HGT)

Horizontal (or lateral) gene transfer (HGT) is a kind of inter-species gene duplication. It is a process of gene (or a sub-chromosomal region) translocation between different organisms (Figure 1.2e). Thus, HGT often results in the generation of an anomalous phylogenetic tree. Viral transduction and bacterial conjugation and transformation are well-known HGT events that host

genome acquire novel genetic information from other organisms or from the environment (Soucy, Huang and Gogarten, 2015; Husnik and McCutcheon, 2018).

Despite its great contribution to the genome evolution among prokaryotes, archaea and virus, HGT between prokaryotes and eukaryotes has been thought to rarely occur, especially in animals because of the barrier in the germlines (Boucher *et al.*, 2003; Syvanen, 2012; Doolittle and Brunet, 2016). Noteworthy examples of HGT between prokaryotes to eukaryotes are nuclear-encoded genes that originally encoded by the endosymbiont bacterial genome (Bock, 2017). Especially, this gene flow from plastid to nucleus still goes on in plant genome (termed as endosymbiotic gene transfer (EGT)) (Matsuo *et al.*, 2005; Bock, 2017).

De novo gene origination

It has long been controversial whether a novel gene can be formed from non-coding sequences (Figure 1.2f) independently from duplication (DNA-mediated, RNA-mediated, or inter-species) of pre-existing genetic materials. However, in 2006, genes of which origins were thought to be ancestrally non-coding sequences were found in the *Drosophila* genome (Begun *et al.*, 2006). Such genes are referred to as *de novo* genes, and are now reported in many eukaryotic species including yeast (Cai *et al.*, 2008; Carvunis *et al.*, 2012; Lu, Leu and Lin, 2017), mammals (Knowles and McLysaght, 2009; Murphy and McLysaght, 2012), and plants (Xiao *et al.*, 2009; Li *et al.*, 2016).

Above studies compared the genome sequences among closely related species, and survey genes without any homologies with other known genes. Such genes appear on the terminal of the phylogenetic tree, and thus are also termed as ‘orphan gene’ (Tautz and Domazet-Lošo, 2011). As increasing the genomic data to be compared, orphan genes sometimes will be no longer orphans, and hence they also referred to as ‘taxon-restricted genes’ (Khalturin *et al.*, 2009).

Importantly, orphan genes do not necessarily arise *de novo* from non-coding sequences, and instead may be generated through duplication-diversification of pre-existing genes (Wissler *et al.*, 2013). For instance, suppose that an orphan gene that arises from gene duplication. Rapid evolution of the duplicated genes may erase the sequence similarity among them, which makes

it difficult to elucidate their phylogenetic relationships (Schlötterer, 2015). Therefore, the definition of *de novo* genes sometimes includes such ambiguous classes.

Features of de novo genes

Although a *de novo* gene and orphan gene is not exactly the same, the properties of these young genes are similar. Typically, these young genes are shorter in length, contain fewer exons, and express lower than established genes (Werner *et al.*, 2018; Van Oss and Carvunis, 2019; Durand *et al.*, 2019). In the specific organs, their expression tends to be higher than that of established genes, especially in the male reproductive tissue (Kaessmann, 2010) (see *Out of testis hypothesis*).

Epigenetic status is also characteristic. Enhancer-like epigenetic configuration was found around the start sites of the young genes; open-chromatin region with enhancer-like histone modifications (Werner *et al.*, 2018; Vakirlis *et al.*, 2018; Majic and Payne, 2020).

ORF-first vs RNA-first model

During the process of *de novo* gene origination, two steps are necessary; acquisition of an ORF and the regulatory elements for active transcription. According to the sequence of events, two models are proposed; ORF-first and RNA-first (McLysaght and Hurst, 2016; Schmitz and Bornberg-Bauer, 2017).

ORF-first

In the ORF-first model, firstly a translatable ORF is newly formed by mutations in the previously non-coding region. The ORF occasionally becomes transcriptionally activated, and form a *de novo* gene.

RNA-first

In the RNA-first model, mutations occur in the region where the ancestral genome is transcribed but without any ORFs (or too much short to have biological significance). If a translatable ORF is generated, it will become a *de novo* gene.

Both models are mutually exclusive. In fact, *de novo* genes in either model were reported in the various species (i.e. ORF-first; fruit fly (Zhao *et al.*, 2014), and fish (Zhuang *et al.*, 2019)RNA-first; fruit fly (Reinhardt *et al.*, 2013), and mouse (Heinen *et al.*, 2009)) Noteworthy, based on the parallel comparison among genomes and transcriptomes of 13 closely related *Oryza* species, Zang *et al*. revealed the stepwise processes of *de novo* gene evolution in the rice genome (Zhang *et al.*, 2019). They analyzed when the *de novo* genes acquired their ORFs and expression, and reported that 91% and 6% of *de novo* genes in the rice genome were explained by the RNA-first and ORF-first model, respectively (Zhang *et al.*, 2019).

Pervasive transcription

Due to the recent development of sequencing technology, it has been widely accepted that the large fraction of eukaryotic genome is transcribed (Clark *et al.*, 2011). Moreover, it was reported that these pervasively/spuriously transcribed RNAs can be translated (Carvunis *et al.*, 2012; Durand *et al.*, 2019). Such a condition can be a trigger of a new gene origination agreed with either ORF-first/RNA-first model (see above). Even in the non-coding sequences, they will be under the selective pressure from the environment cells exposed (Schlötterer, 2015). There are many examples that pervasive transcription is thought as a driver of the *de novo* gene origination in many species including human (Ruiz-Orera *et al.*, 2015), mouse (Neme and Tautz, 2016), yeast (Carvunis *et al.*, 2012; Durand *et al.*, 2019), fruit fly (Zhou *et al.*, 2008), and rice (Zhang *et al.*, 2019).

Out of testis hypothesis

Evolutionally young genes such as *de novo* genes tended to be highly expressed in the male reproductive tissue. This characteristic was observed in many species including mammals (Marques *et al.*, 2005; Wu, Irwin and Zhang, 2011; Xie *et al.*, 2012), fruit flies (Betrán, Thornton and Long, 2002; Kondo *et al.*, 2017), and plants (Wu *et al.*, 2014; Wang *et al.*, 2016).

Importantly, male reproductive tissue is in a transcriptionally permissive status caused by the accessible chromatin configuration (Schmidt, 1996). Such promiscuous conditions for

transcription may be an ideal field for non-coding sequences to acquire their transcription and to make the birth of *de novo* genes (Kaessmann, 2010).

Promoter: center of transcription initiation

Before discussing how newly emerged genes become transcriptionally activated, I clarify what kind of genomic component has a function to initiate transcription.

Promoter

The transcriptional competency of a given DNA sequence is driven by a promoter, the region where the pre-initiation complex (PIC) is assembled and transcription initiates (Haberle and Stark, 2018; Andersson and Sandelin, 2020). Although eukaryotic genome has different classes of RNA polymerases, here, I specifically note the properties of the promoter of the RNA polymerase II (Pol II). Pol II transcribes all protein-coding genes and many non-coding genes into the RNA from the defined position called transcription start site (TSS). TSS is defined by general transcription factors (GTFs) that recognize specific DNA sequence elements called core promoter (or *cis*-element) that typically are defined as the +/- 50 bp centered on the TSS.

Sequence elements of promoter

Well-known core promoter elements are initiator (Inr) and TATA-box, which are widely conserved among eukaryotes (Haberle and Stark, 2018; Andersson and Sandelin, 2020). Inr is typically specified by pyrimidine-purine (Py-Pu) dinucleotide motif, of which the genomic position of the purine residue is the actual TSS (Javahery *et al.*, 1994). TATA-box is an AT-rich region located about 30 bp upstream from TSS, which is recognized by TATA-biding protein (a central component of PIC) (Patikoglou *et al.*, 1999). Importantly, sequence elements including core promoter elements have diverged among genes in order to respond to specific conditions, and many of them are typically found in the species-specific or gene/function-specific manners (Ames and Lovell, 2011; Ballester *et al.*, 2014).

Epigenetic configurations of promoter

In addition to the sequence elements, promoters have a specific chromatin configuration. Generally, nucleosome-depleted open chromatin region is observed around promoter, which enables PIC to access its binding site (Klemm, Shipony and Greenleaf, 2019).

The first nucleosome positioned just downstream of TSS (referred as to +1 nucleosome) contains specific post-translational modified histones and histone variant such as a lysine4 tri-methylation of the histone H3 (H3K4me3), lysine27 acetylation of the histone H3 (H3K27ac), and histone variant H2A.Z (Klemm, Shipony and Greenleaf, 2019).

DNA methylation on the cytosine residue is also important for transcriptional regulation by preventing transcription initiation (Neri *et al.*, 2017). Hence, the region where transcriptional status should be inactivated, such as a gene body, transposable elements, and heterochromatic regions are maintained at a higher methylated level, whereas promoter region is basically hypomethylated (Neri *et al.*, 2017).

Other elements of transcription regulation

Although the core promoter element is sufficient for transcriptional initiation, its complex regulation needs integration of the other proximal or distal signals mediated by the enhancer, transcription factors (TFs), and co-activators (Kadonaga, 2012; Haberle and Stark, 2018). Enhancer regulates transcription from core promoter independently of distances and orientations by binding specific TFs (Andersson and Sandelin, 2020). TF is a protein that regulates the rate of transcription either positively or negatively by binding to a specific sequence element within an enhancer (Spitz and Furlong, 2012). Connecting GTFs to TF, co-activator also controls transcription (Näär, Lemon and Tjian, 2001).

Property of transcripts from promoter

Pol II transcripts are processed before export by adding 5’-cap and 3’-polyadenylated tail at their ends, respectively, which contribute to the stability and translation efficiency of mRNA (Hocine, Singer and Grünwald, 2010). In addition, because 5’-cap is bound on the 5’ end of mRNA, it is

utilized for the determination of TSS. Several techniques were established for precise determination of TSSs based on the biochemical enrichment of mRNA with 5'-cap followed by DNA sequencing (Maruyama and Sugano, 1994; Shiraki *et al.*, 2003).

Bi-directionality of promoter

Promoters can generate transcripts in both directions. Nascent transcript sequencing technologies revealed that transcription occurred opposite orientation of coding sequences (Core, Waterfall and Lis, 2008; Churchman and Weissman, 2011). Such antisense transcripts from promoters are generally unstable, and rapidly degraded when they do not have a certain function (Neil *et al.*, 2009; Wei *et al.*, 2011). More recently, based on the nascent transcript mapping in the *Saccharomyces cerevisiae* genome containing foreign yeast DNA, Jin *et al.* suggested that promoter regions are intrinsically bi-directional (Jin *et al.*, 2017).

Enhancer

Enhancers can also recruit Pol II and initiate transcription as well as promoters (Natoli and Andrau, 2012; Lam *et al.*, 2014). The resultant RNA, called enhancer RNA (eRNA), is generally transcribed both directions of enhancer, and is rapidly degraded (Natoli and Andrau, 2012; Lam *et al.*, 2014). Because eRNAs are processed by the addition of 5'-cap, the conventional technique of TSS determination can be applicable for eRNA analysis (Hirabayashi *et al.*, 2019).

Enhancers also have specific epigenetic properties (Andersson and Sandelin, 2020). Typically, enhancers are found in an open chromatin region where TFs can access. The enhancer region is hypomethylated as well as that of the promoter. In addition, a nucleosome positioned in the vicinity of the enhancer has a specific histone mark; lysine4 mono-methylation of the histone H3 (H3K4me1).

Intrinsic homogeneity of promoter and enhancer

The classical view of the regulation of transcription initiation considers that promoters and enhancers have distinct molecular functions and abilities. However, recent studies suggested their close relativities. Based on the measurements of promoter activities of thousands of small genomic fragments across the entire human genome, van Arensbergen *et al.* showed that many enhancers have weak autonomous promoter activities (van Arensbergen *et al.*, 2017). Due to such promoter activities, enhancers can be one of the sources of new genes (see *de novo* gene origination).

On the other hand, promoters have enhancer activities. Arnold *et al*. reported that 4.5% of *Drosophila* promoters have enhancer activities according to the massive *in vitro* reporter assay (Arnold *et al.*, 2013). A similar observation was reported in the human genome (Dao *et al.*, 2017). This duality of promoters and enhancers are also found in their epigenetic configurations. Specifically, enhancers with higher transcription levels of eRNAs tend to be enriched by H3K4me3 that is known as a histone modification specifically found in promoters (Core *et al.*, 2014). Contrary, H3K4me1, an epigenetic landmark of enhancers is deposited around lower expressed promoters (Hirabayashi *et al.*, 2019). Therefore, although the enrichment of specific marker histones is typical properties of promoters and enhancers, they are not suitable for the identification code to discriminate these two elements. Based on such similarity and duality among enhancers and promoters on the ability of transcriptional initiation and regulation, they should not regard as mutually exclusive, but rather they have different degrees of two abilities (Andersson, Sandelin and Danko, 2015; Henriques *et al.*, 2018; Hirabayashi *et al.*, 2019)

What is a promoter? – The definition in this thesis

Promoters and enhancers have the capability to activate transcription of a given downstream sequence. However, their similarity and duality cannot be explained by the conventional view that assumes promoters and enhancers to be distinct elements according to their sequence and epigenetic marks. Then, how do we consider what the promoter is?

Some researchers claimed an alternative definition; the open chromatin region should be an indicator of transcriptional regulatory (Andersson, Sandelin and Danko, 2015; Henriques *et al.*, 2018; Hirabayashi *et al.*, 2019). The open chromatin region is experimentally determined by the

chromatin accessibility assay by utilizing the nuclease or transposase such as DNaseI, Micrococcal nuclease (MNase), or Tn5 transposase (Tsompana and Buck, 2014). However, these techniques have several problems. For example, the reactions between the chromatin and such enzymes are generally affected by various factors including the reaction conditions, enzyme activities, or their sequence preferences, which cause experimental biases of determined chromatin accessibility (He *et al.*, 2014; Karabacak Calviello *et al.*, 2019; Chereji, Bryson and Henikoff, 2019). In addition, these experiments generally need many cells (typically, more than millions of cells), since lower cell number often provides in discrete and noisy results (Zhou *et al.*, 2019). The obtained chromatin accessibility is an average value of chromatin status of given cells. However, a recent study reported that nucleosome positioning among individual cells showed molecular diversity (Wang *et al.*, 2019; Shipony *et al.*, 2020). Such diversity may make the chromatin accessibility a lower resolution.

Because of the uncertainty of measurements of chromatin accessibility as described above, at least in this thesis, I do not adopt chromatin accessibility as a determinant of the transcriptional regulatory region. Instead, I only consider the region where the experimentally validated TSS exists. TSS definitely indicates that there should be an active promoter, or something with promoter activity. This criterion does not assume promoters and enhancers as mutually exclusive, because both can recruit Pol II. Moreover, the criterion does not consider any sequence elements or epigenetic markers. Thus, this criterion enables us to elucidate what kinds of factors exist around newborn genes with wide and neutral spectrum.

How do new genes obtain their promoter?

A term *gene* indicates a functional unit of a heritable genomic fraction. Sequences without any function are not adaptive, and will be rapidly disrupted by neutral mutations. More specifically, it is not until being a functional gene that a fraction of DNA containing an ORF is transcribed, the RNA is translated, the protein is folded and transported its destination, and such a series of complex systems is precisely regulated spatiotemporally. Even in the newly emerged genes, they should be satisfied above requirements.

Then, how have the new genes obtained their functionality during their evolution? Considering that the non-coding RNAs could evolve into the coding genes (Carvunis *et al.*, 2012), the very first step of their evolution might be to be transcribed, namely, to obtain their promoters. However, it is difficult to compare the promoter sequences among distant species, because their mutation rates are generally higher than those of coding sequences. Moreover, as I described above, eukaryotic promoter is not just a sequence, but a complex status including epigenetic configurations. Therefore, it had been overlooked how newborn genes acquired their promoters.

More recently, due to the extensive development of comparative functional genomics, researchers now can compare the genome, transcriptome, translatome, and epigenome among closely related species, which allows us to elucidate evolutionary young promoters. In this section, I summarized the reported mechanisms by which newborn genes acquired their individual promoters.

Utilizing ancestral promoter

New genes except for *de novo* originated ones could arise together with their ancestral promoters (Figure 1.3a). For instance, gene duplication could occur on the genomic fraction containing the coding sequence with its regulatory elements. In such cases, duplicated copies could immediately express.

Retroposed copies basically lack their promoter sequences during their duplication process (see *retroposition*). However, it was reported that when a parental promoter encloses broadly distributed multiple TSSs, the retroposed copy could sometimes have a sub-fraction of the

parental promoter (Okamura and Nakai, 2008). Subsequent genomic rearrangements (i.e spontaneous mutation) will diversify the promoter sequences of either original or duplicated copy and result in the pseudogenization, neofunctionalization, or subfunctionalization (Ohno, 1970; Ohno, 1972; Force *et al.*, 1999).

Horizontally transferred genes can translocate together with their parental promoters. However, differently from gene duplications, these promoters practically exhibit no or low functionality in the host genome (Cornelissen and Vandewiele, 1989; Silva, Loreto and Clark, 2004). This is mainly because of the inter-species (or sometimes inter-domain) barrier such as the differences of regulatory sequences and corresponding transcription factors between the gene donor and recipient genome (Keeling and Palmer, 2008). Transferred genes need to recruit more adapted promoter sequences through evolution for the regulated expression in the host genome.

Utilizing pre-existing promoter

New genes also become transcriptionally active by recruiting pre-existing promoters or promoter-like elements (Figure 1.3b).

Highjacking pre-existing promoter

In retroposition, new genes can be inserted into the pre-existing genes. Resulting chimeric genes sometimes can be transcribed as fusion transcripts with the host pre-existing genes. Particularly, retrocopies in the human genome are often found in the intron or 5'-UTR of the ancestral host genes and form splice variants, potentially avoiding deleterious effects on the host gene functions (Vinckenbosch, Dupanloup and Kaessmann, 2006). Even if a new gene is not inserted within the pre-existing genes, subsequent genomic rearrangements (i.e, deletion) may connect between a pre-existing promoter and the newly inserted coding sequence.

Analogous to this, promoter acquisition can also occur in the HGT process. Experimental simulation of HGT process reported that exogenously integrated promoter-less coding sequences became transcribed by forming transcriptional fusions with the pre-existing genes (Stangeland *et al.*, 2005).

Bidirectional promoter

In eukaryotic promoter, transcription initiates both directions (see *Bi-directionality of promoter*). Such bidirectional promoter is one of the hot spots for the new gene birth, because they can endow the new gene with transcriptional competency with relatively low deleterious effects on the pre-existing transcription units and networks. In the yeast genome, Vakirlis *et al*. reported that *de novo* genes were highly enriched at bidirectional promoters (Vakirlis *et al.*, 2018). Similar observations were reported in other species including human (Kalitsis and Saffery, 2009; Xie *et al.*, 2012; Gotea, Petrykowska and Elnitski, 2013) and mouse (Neme and Tautz, 2013).

Enhancer

Promoters and enhancers have similar capability against the transcriptional initiation (see *Intrinsic homogeneity of promoter and enhancer*). Hence, enhancers also could endow the new genes with transcriptional competency as promoters do (Kaessmann, Vinckenbosch and Long, 2009). Enhancer also contributes to the *de novo* gene origination (Gotea, Petrykowska and Elnitski, 2013; Long, Prescott and Wysocka, 2016). For instance, based on the comparative genomics among rodent species, Majic and Payne showed that several *de novo* genes in the mouse genome were derived from the ancestrally putative ORFs located in the enhancers (Majic and Payne, 2020). Furthermore, they showed that enhancers contributed in order for the *de novo* genes to immediately integrate into the pre-existing regulatory networks (Majic and Payne, 2020). A study on the nematode genome supported the idea of the enhancer-driven *de novo* gene origination; the epigenetic patterns of *de novo* genes in the nematode genome were similar to those of enhancers than those of promoters (Werner *et al.*, 2018).

Open chromatin

Open chromatin region where nucleosomes are depleted is generally observed in the promoter and enhancer in the eukaryotic genome. The region is transcriptionally permissive because its loose chromatin conformation allows DNA-binding proteins to access. According to the parallel comparison of the genome, transcriptome, and epigenome among a wide range of mammalian genomes (from dog to human), evolutionally young TSSs tended to appear in the ancestrally

accessible and transcriptionally active chromatin regions (Li, Lenhard and Luscombe, 2018). Analogous to this finding, new genes frequently observed in the open chromatin region in the mouse (Majic and Payne, 2020), and nematode (Werner *et al.*, 2018) genome.

Repetitive elements

Repeat sequences account for a large fraction of the eukaryotic genome (Treangen and Salzberg, 2011). The vast majority of them are derived from retrotransposition of LINE (long interspersed nuclear element), SINE (short interspersed nuclear element), and LTR (long terminal repeat) transposable elements. In the mammalian genome, it is well known that transcription initiates from these repetitive elements (Faulkner *et al.*, 2009; Young *et al.*, 2015). Because LTR harbors internal promoter activity, promoters within associated LTRs likely drove such TSSs. In contrast, it is still unclear how non-LTR associated TSSs became activated. Notably, Li *et al.* reported that the evolutionally young genes in the mammalian genome were enriched nearby repetitive elements, and suggested that these repeat elements could be a source of promoters for new genes (Li, Lenhard and Luscombe, 2018). Furthermore, the TSSs within such repetitive elements rapidly evolved by mutations because of the instability of repeat sequences (Li, Lenhard and Luscombe, 2018).

Pervasive transcription

Pervasive transcription in the eukaryotic genome can arise *de novo* genes (see *de novo gene origination*). As I mentioned, the comparative genomics revealed that the evolutionally young genes were enriched in the region where transcriptional activity previously existes without meaningful ORFs. Functional ORFs were subsequently formed through mutations (see *RNA-first model*). Particularly, male reproductive tissue is a hot spot for the *de novo* gene origination, because of its transcriptionally permissive status that allows occurring pervasive transcription (see *out-of-testis hypothesis*). While these supported the RNA-first model for the *de novo* gene origination, pervasive transcription also could contribute to the ORF-first type of gene birth events by being a source of transcriptional competency. For instance, if the horizontally transferred genes are inserted into the transcriptionally inert regions, they can be occasionally transcribed by forming transcriptional fusions with pervasive transcripts (Husnik and

McCutcheon, 2018). Thus, pervasive transcription can provide initial round of transcription to the newborn coding sequences, which will open up subsequent adaptive evolution.

De novo origination

Transcriptional regulatory sequences of new genes can emerge *de novo* from transcriptionally inert regions (Figure 1.3c). Mutations such as nucleotide substitutions can arise novel sites for TF binding. Based on the comparative genomics of the closely related *Drosophila* strains, Zhao *et al*. reported that *de novo* genes are transcriptionally activated by the spontaneous substitutions that generated *cis*-acting sequences in the vicinity of the originated coding sequences (Zhao *et al.*, 2014). This type of promoter acquisition mechanism of *de novo* genes was also reported in the hominoid genomes (Ruiz-Orera *et al.*, 2015). Several retroposed genes in the *Drosophila* genome obtained their promoter elements as with similar manner (Betrán and Long, 2003).

TSS can also emerge without any mutations in the previously non-promoter region. Kudo *et al*. reported the promoter *de novo* origination: exogenously inserted promoter-less coding sequence acquired a brand-new promoter-like epigenetic status and TSSs at the 5' proximal region of the insert (Kudo, Marsuo, and Satoh et al. 2020). However, the example of such event was so limited to illustrate the generality and extensiveness of this transcriptional activation mechanism (Kudo, Marsuo, and Satoh et al. 2020).

Conclusion of Historical Review

To generate new genes is a fundamental property of the genome. Studies on this property have developed based on the analysis of 'young genes' characterized by the phylogenetic comparison of the genomic sequences among various species (Figure 1.1). Such comparative genomics has provided great insights about new gene originations such as how their novel coding sequences are generated, and how they acquired their transcriptional competency. However, despite extensive developments in this field of research, the time-resolution of comparative genomics has intrinsic limitations. Practically, 'young' genes are relatively young among species to be compared, but have already experienced at least hundreds of thousands of years. Because such young genes already have been selected and fixed in the genome, we cannot know the actual population of newborn genes had arisen. Moreover, it is open to question whether such 'young' genes still have maintained the appearances since their birth. Therefore, to elucidate the appearances of 'truly newborn' genes, we need an alternative approach to analyze much younger genes directly, instead of the circumstantial evidences from the relatively young genes according to the phylogenetic comparison.

In this thesis, I aimed to study the appearances of newborn genes whose time-resolution cannot be approached by comparative genomics. On that account, I performed an artificial gene evolutionary experiment in the plant genome. In order to lead new coding sequences to arise in the genome, I carried out promoter-trap screening as a mimic of HGT process. Firstly, I refined a transformation protocol of plant cell culture to obtain massive number of transformants for the promoter-trap screening. By using the massive transgenic cell lines, I analyzed where, how, and how often the plant genome endows newborn coding sequences transcriptional competency. Based on the TSS analysis of these activated transcription, I proposed a model to explain the gene origination process in the plant genome. In addition, I demonstrated the genetic behavior of such activated transcripts in the plant genome. Finally, I summarized the obtained results and discussed future subjects.

Outline of this thesis

The central question of this thesis is the molecular mechanism by which newborn coding sequences become functional in the genome. As I noted, the first step of functionalization of newborn coding sequences is to be transcribed. To reveal how newborn coding sequences become transcribed, we carried out an experimental simulation of gene origination events in the plant genome (see below) focusing on the following three objectives.

Objective 1:

Objective 1 aims to reveal how, how often, and where the plant genome can transcriptionally activate newly originated coding sequences. For this aim, we establish a high-throughput promoter-trap screening technique in order to elucidate the transcriptional fates of artificially mimicked newborn genes that are exogenously inserted into the genome.

Objective 2:

Objective 2 aims to infer the very initial process of the gene evolution focusing on how newborn coding sequences acquire their initial transcription. For this purpose, we analyze the promoter architecture of transcriptionally activated newborn coding sequences in the artificial evolutionary experiment.

Objective 3:

Objective 3 aims to elucidate whether the activated transcription can transmit to the next generation. For this aim, we carry out the artificial evolutionary experiment in the plants and analyze T2 generation (one generation after from origination of coding sequence) of them.

Experimental evolution approach

To date, gene evolution researches have mainly led by the comparative genomics, which has provided knowledge about new (strictly, young) genes, i.e., their region preferences, birth rates, and evolutionary fates. However, because the comparative genomics can treat the pre-selected and pre-fixed genes in the genome, it is still unclear what extent of the new genes is actually originated including those rapidly eliminated. How could we overcome this situation?

Experimental evolution approach could be an alternative, providing direct information of just newborn genes in a controlled condition (Garland and Rose, 2009). In plants, exogenously introduced coding sequences that mimic the originated genes through HGT/EGT event provided insights about how newborn coding sequences become transcribed. One is simulation of EGT process based on the transplastomic approach (Bock, 2017). In this system, a reporter gene is introduced into the plastid genome. The reporter gene is transcriptionally inert in the plastid genome but can be active in the nuclear genome. Note that plastid DNAs are constantly integrated into the nuclear genome (Matsuo *et al.*, 2005). Therefore, if such escaped DNA from the plastid to the nucleus includes the reporter gene, its expression will be observed. Previous studies suggested that transferred plastid genes become transcriptionally active by trapping neighbouring eukaryotic promoters, or utilizing the prokaryotic plastid promoter sequences (Stegemann and Bock, 2006; Wang *et al.*, 2014).

Another one is based on the promoter-trap screening. This is an experimental method to capture and analyze previously unknown promoters in the genome (Friedrich and Soriano, 1991). More practically, a promoter-less coding sequence of a reporter gene is introduced by the stable transformation method. If the promoter-less construct expresses, it should 'trap' a promoter at a given genomic locus. Interestingly, many studies reported that exogenously inserted promoter-less coding sequences became transcribed without trapping any annotated promoters (Fobert *et al.*, 1994; Topping *et al.*, 1994; Plesch, Kamann and Mueller-Roeber, 2000; Mollier *et al.*, 2000; Yamamoto *et al.*, 2003; Sivanandan *et al.*, 2005; Stangeland *et al.*, 2005). More recently, Kudo *et al.* demonstrated that such unexpected transcriptional activation in the promoter-trap experiment occurred at least by two different mechanisms; (1) cryptic promoter capturing, in which exogenous DNA was transcribed by trapping pre-existing promoter-like chromatin configuration, and (2) promoter *de novo* origination, in which promoter-like epigenetic landscapes were newly formed via chromatin remodeling triggered by DNA insertion (Kudo,

Matsuo, and Satoh *et al.*, 2020).

As described above, the experimental evolution approach provided valuable insights about how newly originated coding sequences become transcriptionally activated in the foreign genome environment. However, due to the low throughput of these techniques, it is difficult to illustrate the generality and extensiveness of the observed transcriptional activation events.

Here, we carry out this promoter-trap screening as a mimic of gene origination event and analyze how such newly originated coding sequences become transcribed. In addition, to improve the throughput of this technique, we apply massively parallel reporter assay (MPRA) to this experiment (see below).

Massively parallel reporter assay (MPRA)

MPRA is a high through-put technology that enables us to analyze transcriptional activities of the thousands of reporter genes by utilizing NGS technology (Akhtar *et al.*, 2013; Inoue and Ahituv, 2015). Specifically, in the MPRA, each reporter construct is indexed by individual unique sequence tags (often termed as 'barcode') in advance of introduction to the genome. Because each reporter gene harbors a distinct barcode sequence, transcripts of individual reporter genes will be indexed uniquely as well, which allows us to analyze expression of each transgenic cell line *in silico* without establishing individual isogenic lines.

As the essential components of the MPRA are only the barcoded library and NGS technology, a variety of sequence arrangements can be applicable depending on what to analyze, i.e. enhancer activities (Arnold *et al.*, 2013; Arnold *et al.*, 2014; Arnold *et al.*, 2017), promoter activities (Patwardhan *et al.*, 2009; van Arensbergen *et al.*, 2017), mRNA stability, or chromatin position effect (Akhtar *et al.*, 2013). In this thesis, an MPRA-based high throughput promoter-trap screening was designed in order to analyze how newborn genes become functional in the plant genome.

Chapter 2:

To perform a genome-wide promoter-trap screening to reveal the mechanism by which newborn

genes become functional, massive number of transformants harboring individual reporter constructs along the entire chromosomes are needed. For this purpose, in Chapter 2, we attempted to improve the transformation efficiency of *Arabidopsis thaliana* T87 cell culture. We examined suitable media which could stably induce the T87 cells transformation-competent high-efficiently, and presented a refined transformation protocol that could be useful for MPRA, or other transgenic-based analysis.

Chapter 3:

In order to study how, how often, and where newborn genes become transcriptionally active in the plant genome, in chapter 3, we carried out an artificial evolutionary experiment based on the MPRA adapted to the conventional promoter-trap screening. Specifically, we introduced promote-less firefly luciferase gene (LUC) as a model of newborn coding sequences into the genome of *A. thaliana* T87 cells and analyzed what kind of the genomic property (i.e. position, or transcription and epigenetic status) was responsible for their transcriptional activation. We found a novel class of plant genome response, i.e., integration-dependent stochastic transcriptional activation, which occurs stochastically at a certain frequency of each insertion event but independently of the chromosomal locus in respect to the pre-existing genes, inherent transcribed regions, or heterochromatic regions.

Chapter 4:

To obtain insights into the molecular basis of the integration-dependent stochastic transcriptional activation, we mapped precise positions of TSSs of the inserted promoter-less LUC genes. Due to the systematic characterization of determined TSSs, we found *de novo* transcriptional initiation; transcription occurs *de novo* about 100 bp upstream of newborn coding sequences with avoiding pre-existing Kozak-containing reading frames. These features could be a clue to elucidate a first selection gate for newborn transcripts to evolve into functional genes. Based on the above results, we propose a model to explain the gene origination process in the plant genome.

Chapter 5:

To infer whether the *de novo* activated transcription can transmit to the next generation, we carried out an artificial evolutionary experiment in the T2 generation of *Arabidopsis* plants under the similar experimental scheme of the previous study using cultured cells. By comparing the results between plants and cultured cells, we concluded that *de novo* transcriptional activation along with chromatin rewiring should be an inheritable phenomenon of plant genome.

Chapter 6:

In chapter 5, I summarize obtained results of our studies. Remained questions and possible approaches are also presented.

References of Chapter 1

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(4), 914–927.

Ames, R. M. and Lovell, S. C. (2011) Diversification at transcription factor binding sites within a species and the implications for environmental adaptation. *Mol Biol Evol,* 28(12), 3331–3344.

Andersson, R. and Sandelin, A. (2020) Determinants of enhancer and promoter activities of regulatory elements. *Nat Rev Genet,* 21(2), 71–87.

Andersson, R., Sandelin, A. and Danko, C. G. (2015) A unified architecture of transcriptional regulatory elements. *Trends Genet,* 31(8), 426–433.

Arnold, C. D., Gerlach, D., Spies, D., Matts, J. A., Sytnikova, Y. A., Pagani, M., Lau, N. C. and Stark, A. (2014) Quantitative genome-wide enhancer activity maps for five Drosophila species show functional enhancer conservation and turnover during cis-regulatory evolution. *Nat Genet,* 46(7), 685–692.

Arnold, C. D., Gerlach, D., Stelzer, C., Boryń, Ł., Rath, M. and Stark, A. (2013) Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science,* 339(6123), 1074–1077.

Arnold, C. D., Zabidi, M. A., Pagani, M., Rath, M., Schernhuber, K., Kazmar, T. and Stark, A. (2017) Genome-wide assessment of sequence-intrinsic enhancer responsiveness at single-base-pair resolution. *Nat Biotechnol,* 35(2), 136–144.

Ballester, B., Medina-Rivera, A., Schmidt, D., Gonzàlez-Porta, M., Carlucci, M., Chen, X., Chessman, K., Faure, A. J., Funnell, A. P., Goncalves, A., Kutter, C., Lukk, M., Menon, S., McLaren, W. M., Stefflova, K., Watt, S., Weirauch, M. T., Crossley, M., Marioni, J. C., Odom, D. T., Flicek, P. and Wilson, M. D. (2014) Multi-species, multi-transcription factor binding highlights conserved control of tissue-specific biological pathways. *Elife,* 3, e02626.

Begun, D. J., Lindfors, H. A., Thompson, M. E. and Holloway, A. K. (2006) Recently evolved genes identified from Drosophila yakuba and D. erecta accessory gland expressed sequence

tags. *Genetics,* 172(3), 1675–1681.

Betrán, E. and Long, M. (2003) Dntf-2r, a young Drosophila retroposed gene with specific male expression under positive Darwinian selection. *Genetics,* 164(3), 977–988.

Betrán, E., Thornton, K. and Long, M. (2002) Retroposed new genes out of the X in Drosophila. *Genome Res,* 12(12), 1854–1859.

Bock, R. (2017) Witnessing Genome Evolution: Experimental Reconstruction of Endosymbiotic and Horizontal Gene Transfer. *Annu Rev Genet,* 51, 1–22.

Boucher, Y., Douady, C. J., Papke, R. T., Walsh, D. A., Boudreau, M. E., Nesbø, C. L., Case, R. J. and Doolittle, W. F. (2003) Lateral gene transfer and the origins of prokaryotic groups. *Annu Rev Genet,* 37, 283–328.

Cai, J., Zhao, R., Jiang, H. and Wang, W. (2008) De novo origination of a new protein-coding gene in Saccharomyces cerevisiae. *Genetics,* 179(1), 487–496.

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(7407), 370–374.

Chereji, R. V., Bryson, T. D. and Henikoff, S. (2019) Quantitative MNase-seq accurately maps nucleosome occupancy levels. *Genome Biol,* 20(1), 198.

Churchman, L. S. and Weissman, J. S. (2011) Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature,* 469(7330), 368–373.

Clark, J. W. and Donoghue, P. C. J. (2018) Whole-Genome Duplication and Plant Macroevolution. *Trends Plant Sci,* 23(10), 933–945.

Clark, M. B., Amaral, P. P., Schlesinger, F. J., Dinger, M. E., Taft, R. J., Rinn, J. L., Ponting, C. P., Stadler, P. F., Morris, K. V., Morillon, A., Rozowsky, J. S., Gerstein, M. B., Wahlestedt, C., Hayashizaki, Y., Carninci, P., Gingeras, T. R. and Mattick, J. S. (2011) The reality of pervasive transcription. *PLoS Biol,* 9(7), e1000625; discussion e1001102.

Core, L. J., Martins, A. L., Danko, C. G., Waters, C. T., Siepel, A. and Lis, J. T. (2014)

Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat Genet,* 46(12), 1311–1320.

Core, L. J., Waterfall, J. J. and Lis, J. T. (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science,* 322(5909), 1845–1848.

Cornelissen, M. and Vandewiele, M. (1989) Nuclear transcriptional activity of the tobacco plastid psbA promoter. *Nucleic Acids Res,* 17(1), 19–29.

Dao, L. T. M., Galindo-Albarrán, A. O., Castro-Mondragon, J. A., Andrieu-Soler, C., Medina-Rivera, A., Souaid, C., Charbonnier, G., Griffon, A., Vanhille, L., Stephen, T., Alomairi, J., Martin, D., Torres, M., Fernandez, N., Soler, E., van Helden, J., Puthier, D. and Spicuglia, S. (2017) Genome-wide characterization of mammalian promoters with distal enhancer functions. *Nat Genet,* 49(7), 1073–1081.

Darwin, C. (1859) On The Origin of Species by Means of Natural Selection, or Preservation of Favoured Races in the Struggle for Life. London: John Murray.

Doolittle, W. F. and Brunet, T. D. (2016) What Is the Tree of Life? *PLoS Genet,* 12(4), e1005912.

Durand, É., Gagnon-Arsenault, I., Hallin, J., Hatin, I., Dubé, A. K., Nielly-Thibault, L., Namy, O. and Landry, C. R. (2019) Turnover of ribosome-associated transcripts from de novo ORFs produces gene-like characteristics available for de novo gene emergence in wild yeast populations. *Genome Res,* 29(6), 932–943.

Eichler, E. E. (2001) Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends Genet,* 17(11), 661–669.

Faulkner, G. J., Kimura, Y., Daub, C. O., Wani, S., Plessy, C., Irvine, K. M., Schroder, K., Cloonan, N., Steptoe, A. L., Lassmann, T., Waki, K., Hornig, N., Arakawa, T., Takahashi, H., Kawai, J., Forrest, A. R., Suzuki, H., Hayashizaki, Y., Hume, D. A., Orlando, V., Grimmond, S. M. and Carninci, P. (2009) The regulated retrotransposon transcriptome of mammalian cells. *Nat Genet,* 41(5), 563–571.

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(4), 567–577.

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. L. and Postlethwait, J. (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics,* 151(4), 1531–1545.

Friedrich, G. and Soriano, P. (1991) Promoter traps in embryonic stem cells: a genetic screen to identify and mutate developmental genes in mice. *Genes Dev,* 5(9), 1513–1523.

Garland, T. and Rose, M. R. (2009) Experimental Evolution: Concepts, Methods, and Applications of Selection Experiments. University of California Press Books.

Gilbert, W. (1978) Why genes in pieces? *Nature,* 271(5645), 501.

Gotea, V., Petrykowska, H. M. and Elnitski, L. (2013) Bidirectional promoters as important drivers for the emergence of species-specific transcripts. *PLoS One,* 8(2), e57323.

Haberle, V. and Stark, A. (2018) Eukaryotic core promoters and the functional basis of transcription initiation. *Nat Rev Mol Cell Biol,* 19(10), 621–637.

Haldane, J. B. S. (1933) The Part Played by Recurrent Mutation in Evolution. *The American Naturalist*. 67(708), 5–19.

Harris, C., L. (1981) Evolution, genesis and revelations, with readings from Empedocles to Wilson. Albany : State University of New York Press.

He, H. H., Meyer, C. A., Hu, S. S., Chen, M. W., Zang, C., Liu, Y., Rao, P. K., Fei, T., Xu, H., Long, H., Liu, X. S. and Brown, M. (2014) Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. *Nat Methods,* 11(1), 73–78.

Heinen, T. J., Staubach, F., Häming, D. and Tautz, D. (2009) Emergence of a new gene from an intergenic region. *Curr Biol,* 19(18), 1527–1531.

Henriques, T., Scruggs, B. S., Inouye, M. O., Muse, G. W., Williams, L. H., Burkholder, A. B., Lavender, C. A., Fargo, D. C. and Adelman, K. (2018) Widespread transcriptional pausing and elongation control at enhancers. *Genes Dev,* 32(1), 26–41.

Hirabayashi, S., Bhagat, S., Matsuki, Y., Takegami, Y., Uehata, T., Kanemaru, A., Itoh, M., Shirakawa, K., Takaori-Kondo, A., Takeuchi, O., Carninci, P., Katayama, S., Hayashizaki,

Y., Kere, J., Kawaji, H. and Murakawa, Y. (2019) NET-CAGE characterizes the dynamics and topology of human transcribed cis-regulatory elements. *Nat Genet,* 51(9), 1369–1379.

Hocine, S., Singer, R. H. and Grünwald, D. (2010) RNA processing and export. *Cold Spring Harb Perspect Biol,* 2(12), a000752.

Husnik, F. and McCutcheon, J. P. (2018) Functional horizontal gene transfer from bacteria to eukaryotes. *Nat Rev Microbiol,* 16(2), 67–79.

Inoue, F. and Ahituv, N. (2015) Decoding enhancers using massively parallel reporter assays. *Genomics,* 106(3), 159–164.

Jacob, F. (1977) Evolution and tinkering. *Science,* 196(4295), 1161–1166.

Javahery, R., Khachi, A., Lo, K., Zenzie-Gregory, B. and Smale, S. T. (1994) DNA sequence requirements for transcriptional initiator activity in mammalian cells. *Mol Cell Biol,* 14(1), 116–127.

Jin, Y., Eser, U., Struhl, K. and Churchman, L. S. (2017) The Ground State and Evolution of Promoter Region Directionality. *Cell,* 170(5), 889–898.e10.

Johnson, M. E., Viggiano, L., Bailey, J. A., Abdul-Rauf, M., Goodwin, G., Rocchi, M. and Eichler, E. E. (2001) Positive selection of a gene family during the emergence of humans and African apes. *Nature,* 413(6855), 514–519.

Kadonaga, J. T. (2012) Perspectives on the RNA polymerase II core promoter. *Wiley Interdiscip Rev Dev Biol,* 1(1), 40–51.

Kaessmann, H. (2010) Origins, evolution, and phenotypic impact of new genes. *Genome Res,* 20(10), 1313–1326.

Kaessmann, H., Vinckenbosch, N. and Long, M. (2009) RNA-based gene duplication: mechanistic and evolutionary insights. *Nat Rev Genet,* 10(1), 19–31.

Kalitsis, P. and Saffery, R. (2009) Inherent promoter bidirectionality facilitates maintenance of sequence integrity and transcription of parasitic DNA in mammalian genomes. *BMC Genomics,* 10, 498.

Karabacak Calviello, A., Hirsekorn, A., Wurmus, R., Yusuf, D. and Ohler, U. (2019)

Reproducible inference of transcription factor footprints in ATAC-seq and DNase-seq datasets using protocol-specific bias modeling. *Genome Biol,* 20(1), 42.

Keeling, P. J. and Palmer, J. D. (2008) Horizontal gene transfer in eukaryotic evolution. *Nat Rev Genet,* 9(8), 605–618.

Khalturin, K., Hemmrich, G., Fraune, S., Augustin, R. and Bosch, T. C. (2009) More than just orphans: are taxonomically-restricted genes important in evolution? *Trends Genet,* 25(9), 404–413.

Klemm, S. L., Shipony, Z. and Greenleaf, W. J. (2019) Chromatin accessibility and the regulatory epigenome. *Nat Rev Genet,* 20(4), 207–220.

Knowles, D. G. and McLysaght, A. (2009) Recent de novo origin of human protein-coding genes. *Genome Res,* 19(10), 1752–1759.

Kondo, S., Vedanayagam, J., Mohammed, J., Eizadshenass, S., Kan, L., Pang, N., Aradhya, R., Siepel, A., Steinhauer, J. and Lai, E. C. (2017) New genes often acquire male-specific functions but rarely become essential in Drosophila. *Genes Dev,* 31(18), 1841–1846.

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., *bioRxiv* [posted 2020 Nov 28]. Available from: https://www.biorxiv.org/content/10.1101/2020.11.28.399337v1 doi: 10.1101/2020.11.28.399337

Lam, M. T., Li, W., Rosenfeld, M. G. and Glass, C. K. (2014) Enhancer RNAs and regulated transcriptional programs. *Trends Biochem Sci,* 39(4), 170–182.

Levine, M. T., Jones, C. D., Kern, A. D., Lindfors, H. A. and Begun, D. J. (2006) Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression. *Proc Natl Acad Sci U S A,* 103(26), 9935–9939.

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(5), 676–688.

Li, Z., Qin, F. and Li, H. (2018) Chimeric RNAs and their implications in cancer. *Curr Opin*

Genet Dev, 48, 36–43.

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(7), 2190–2202.

Long, H. K., Prescott, S. L. and Wysocka, J. (2016) Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell,* 167(5), 1170–1187.

Lu, T. C., Leu, J. Y. and Lin, W. C. (2017) A Comprehensive Analysis of Transcript-Supported De Novo Genes in Saccharomyces sensu stricto Yeasts. *Mol Biol Evol,* 34(11), 2823–2838.

Lynch, M. and Conery, J. S. (2000) The evolutionary fate and consequences of duplicate genes. *Science,* 290(5494), 1151–1155.

Majic, P. and Payne, J. L. (2020) Enhancers Facilitate the Birth of De Novo Genes and Gene Integration into Regulatory Networks. *Mol Biol Evol,* 37(4), 1165–1178.

Marques, A. C., Dupanloup, I., Vinckenbosch, N., Reymond, A. and Kaessmann, H. (2005) Emergence of young human genes after a burst of retroposition in primates. *PLoS Biol,* 3(11), e357.

Maruyama, K. and Sugano, S. (1994) Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. *Gene,* 138(1-2), 171–174.

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(3), 665–675.

Maxam, A. M. and Gilbert, W. (1977) A new method for sequencing DNA. *Proc Natl Acad Sci U S A,* 74(2), 560–564.

McLysaght, A. and Hurst, L. D. (2016) Open questions in the study of de novo genes: what, how and why. *Nat Rev Genet,* 17(9), 567–578.

Mendel, J. G. (1865) Versuche über Planzenhybriden. Verhandlungen des naturforschenden Vereines in Brünn, Bd.

Miller, J. (2008) Daoism and Nature. *Nature Across Cultures*, 393–409

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(11), 1076–1083.

Muller, H. J. (1935) The origination of chromatin deficiencies as minute deletions subject to insertion elsewhere. *Genetica*. 17, 237–252

Murphy, D. N. and McLysaght, A. (2012) De novo origin of protein-coding genes in murine rodents. *PLoS One,* 7(11), e48650.

Natoli, G. and Andrau, J. C. (2012) Noncoding transcription at enhancers: general principles and functional models. *Annu Rev Genet,* 46, 1–19.

Neil, H., Malabat, C., d'Aubenton-Carafa, Y., Xu, Z., Steinmetz, L. M. and Jacquier, A. (2009) Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. *Nature,* 457(7232), 1038–1042.

Neme, R. and Tautz, D. (2013) Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. *BMC Genomics,* 14, 117.

Neme, R. and Tautz, D. (2016) Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to de novo gene emergence. *Elife,* 5, e09977.

Neri, F., Rapelli, S., Krepelova, A., Incarnato, D., Parlato, C., Basile, G., Maldotti, M., Anselmi, F. and Oliviero, S. (2017) Intragenic DNA methylation prevents spurious transcription initiation. *Nature,* 543(7643), 72–77.

Näär, A. M., Lemon, B. D. and Tjian, R. (2001) Transcriptional coactivator complexes. *Annu Rev Biochem,* 70, 475–501.

Ohno, S. (1970) *Evolution by gene duplication.* Springer, Berlin, Heidelberg.

Ohno, S. (1972) So much "junk" DNA in our genome. *Brookhaven Symp Biol,* 23, 366–370.

Okamura, K. and Nakai, K. (2008) Retrotransposition as a source of new promoters. *Mol Biol Evol,* 25(6), 1231–1238.

Özlem, B., Mehmet, K., Derya, Y. and Medine, G. (2013) Genomic Rearrangements and Evolution. Current Progress in Biological Research. chapter2, pp, 19–39

Patikoglou, G. A., Kim, J. L., Sun, L., Yang, S. H., Kodadek, T. and Burley, S. K. (1999) TATA element recognition by the TATA box-binding protein has been conserved throughout evolution. *Genes Dev,* 13(24), 3217–3230.

Patthy, L. (1999) Genome evolution and the evolution of exon-shuffling--a review. *Gene,* 238(1), 103–114.

Patwardhan, R. P., Lee, C., Litvin, O., Young, D. L., Pe'er, D. and Shendure, J. (2009) High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat Biotechnol,* 27(12), 1173–1175.

Plesch, G., Kamann, E. and Mueller-Roeber, B. (2000) Cloning of regulatory sequences mediating guard-cell-specific gene expression. *Gene,* 249(1-2), 83–89.

Reinhardt, J. A., Wanjiru, B. M., Brant, A. T., Saelao, P., Begun, D. J. and Jones, C. D. (2013) De novo ORFs in Drosophila are important to organismal fitness and evolved rapidly from previously non-coding sequences. *PLoS Genet,* 9(10), e1003860.

Ruiz-Orera, J., Hernandez-Rodriguez, J., Chiva, C., Sabidó, E., Kondova, I., Bontrop, R., Marqués-Bonet, T. and Albà, M. M. (2015) Origins of De Novo Genes in Human and Chimpanzee. *PLoS Genet,* 11(12), e1005721.

Sanger, F. and Coulson, A. R. (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol,* 94(3), 441–448.

Sanger, F., Nicklen, S. and Coulson, A. R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A,* 74(12), 5463–5467.

Schlötterer, C. (2015) Genes from scratch--the evolutionary fate of de novo genes. *Trends Genet,* 31(4), 215–219.

Schmidt, E. E. (1996) Transcriptional promiscuity in testes. *Curr Biol,* 6(7), 768–769.

Schmitz, J. F. and Bornberg-Bauer, E. (2017) Fact or fiction: updates on how protein-coding genes might emerge. *F1000Res,* 6, 57.

Shipony, Z., Marinov, G. K., Swaffer, M. P., Sinnott-Armstrong, N. A., Skotheim, J. M., Kundaje, A. and Greenleaf, W. J. (2020) Long-range single-molecule mapping of chromatin

accessibility in eukaryotes. *Nat Methods,* 17(3), 319–327.

Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Kodzius, R., Watahiki, A., Nakamura, M., Arakawa, T., Fukuda, S., Sasaki, D., Podhajska, A., Harbers, M., Kawai, J., Carninci, P. and Hayashizaki, Y. (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc Natl Acad Sci U S A,* 100(26), 15776–15781.

Silva, J. C., Loreto, E. L. and Clark, J. B. (2004) Factors that affect the horizontal transfer of transposable elements. *Curr Issues Mol Biol,* 6(1), 57–71.

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(3), 202–208.

Soucy, S. M., Huang, J. and Gogarten, J. P. (2015) Horizontal gene transfer: building the web of life. *Nat Rev Genet,* 16(8), 472–482.

Spitz, F. and Furlong, E. E. (2012) Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet,* 13(9), 613–626.

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(419), 2495–2505.

Stegemann, S. and Bock, R. (2006) Experimental reconstruction of functional gene transfer from the tobacco plastid genome to the nucleus. *Plant Cell,* 18(11), 2869–2878.

Stewart, N. B. and Rogers, R. L. (2019) Chromosomal rearrangements as a source of new gene formation in Drosophila yakuba. *PLoS Genet,* 15(9), e1008314.

Stoltzfus, A. (1999) On the possibility of constructive neutral evolution. *J Mol Evol,* 49(2), 169–181.

Syvanen, M. (2012) Evolutionary implications of horizontal gene transfer. *Annu Rev Genet,* 46, 341–358.

Tautz, D. and Domazet-Lošo, T. (2011) The evolutionary origin of orphan genes. *Nat Rev Genet,* 12(10), 692–702.

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(6), 895–903.

Treangen, T. J. and Salzberg, S. L. (2011) Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet,* 13(1), 36–46.

Tsompana, M. and Buck, M. J. (2014) Chromatin accessibility: a window into the genome. *Epigenetics Chromatin,* 7(1), 33.

Vakirlis, N., Hebert, A. S., Opulente, D. A., Achaz, G., Hittinger, C. T., Fischer, G., Coon, J. J. and Lafontaine, I. (2018) A Molecular Portrait of De Novo Genes in Yeasts. *Mol Biol Evol,* 35(3), 631–645.

van Arensbergen, J., FitzPatrick, V. D., de Haas, M., Pagie, L., Sluimer, J., Bussemaker, H. J. and van Steensel, B. (2017) Genome-wide mapping of autonomous promoter activity in human cells. *Nat Biotechnol,* 35(2), 145–153.

Van Oss, S. B. and Carvunis, A. R. (2019) De novo gene birth. *PLoS Genet,* 15(5), e1008160.

Vinckenbosch, N., Dupanloup, I. and Kaessmann, H. (2006) Evolutionary fate of retroposed gene copies in the human genome. *Proc Natl Acad Sci U S A,* 103(9), 3220–3225.

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(6), 1327–1334.

Wang, J., Tao, F., Marowsky, N. C. and Fan, C. (2016) Evolutionary Fates and Dynamic Functionalization of Young Duplicate Genes in Arabidopsis Genomes. *Plant Physiol,* 172(1), 427–440.

Wang, Y., Wang, A., Liu, Z., Thurman, A. L., Powers, L. S., Zou, M., Zhao, Y., Hefel, A., Li, Y., Zabner, J. and Au, K. F. (2019) Single-molecule long-read sequencing reveals the chromatin basis of gene expression. *Genome Res,* 29(8), 1329–1342.

Wei, W., Pelechano, V., Järvelin, A. I. and Steinmetz, L. M. (2011) Functional consequences of bidirectional promoters. *Trends Genet,* 27(7), 267–276.

Werner, M. S., Sieriebriennikov, B., Prabh, N., Loschko, T., Lanz, C. and Sommer, R. J. (2018) Young genes have distinct gene structure, epigenetic profiles, and transcriptional regulation. *Genome Res,* 28(11), 1675–1687.

Wissler, L., Gadau, J., Simola, D. F., Helmkampf, M. and Bornberg-Bauer, E. (2013) Mechanisms and dynamics of orphan gene emergence in insect genomes. *Genome Biol Evol,* 5(2), 439–455.

Wu, D. D., Irwin, D. M. and Zhang, Y. P. (2011) De novo origin of human protein-coding genes. *PLoS Genet,* 7(11), e1002379.

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(10), 2822–2829.

Xiao, W., Liu, H., Li, Y., Li, X., Xu, C., Long, M. and Wang, S. (2009) A rice gene of de novo origin negatively regulates pathogen-induced defense response. *PLoS One,* 4(2), e4603.

Xie, C., Zhang, Y. E., Chen, J. Y., Liu, C. J., Zhou, W. Z., Li, Y., Zhang, M., Zhang, R., Wei, L. and Li, C. Y. (2012) Hominoid-specific de novo protein-coding genes originating from long non-coding RNAs. *PLoS Genet,* 8(9), e1002942.

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(2), 273–283.

Young, R. S., Hayashizaki, Y., Andersson, R., Sandelin, A., Kawaji, H., Itoh, M., Lassmann, T., Carninci, P., Bickmore, W. A., Forrest, A. R., Taylor, M. S. and Consortium, F. (2015) The frequent evolutionary birth and death of functional promoters in mouse and human. *Genome Res,* 25(10), 1546–1557.

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(4), 679–690.

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(6172), 769–772.

Zhou, Q., Zhang, G., Zhang, Y., Xu, S., Zhao, R., Zhan, Z., Li, X., Ding, Y., Yang, S. and Wang, W. (2008) On the origin of new genes in Drosophila. *Genome Res,* 18(9), 1446–1455.

Zhou, W., Ji, Z., Fang, W. and Ji, H. (2019) Global prediction of chromatin accessibility using small-cell-number and single-cell RNA-seq. *Nucleic Acids Res,* 47(19), e121.

Zhuang, X., Yang, C., Murphy, K. R. and Cheng, C. C. (2019) Molecular mechanism and history of non-sense to sense evolution of antifreeze glycoprotein gene in northern gadids. *Proc Natl Acad Sci U S A,* 116(10), 4400–4405.

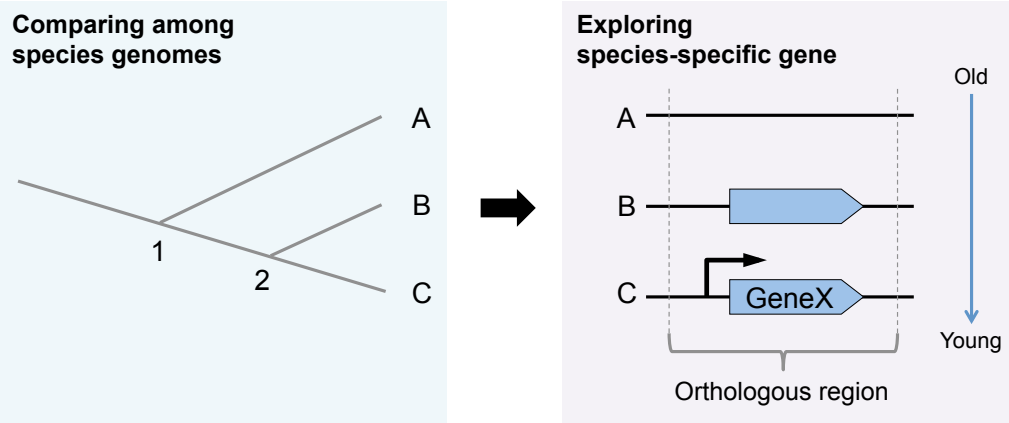


Figure 1.1. Scheme of comparative genomics approach. Genome, transcriptome, translatome, and epigenome are compared among closely related species. Coding sequence or transcription unit without any homologies with other known genes are explored.

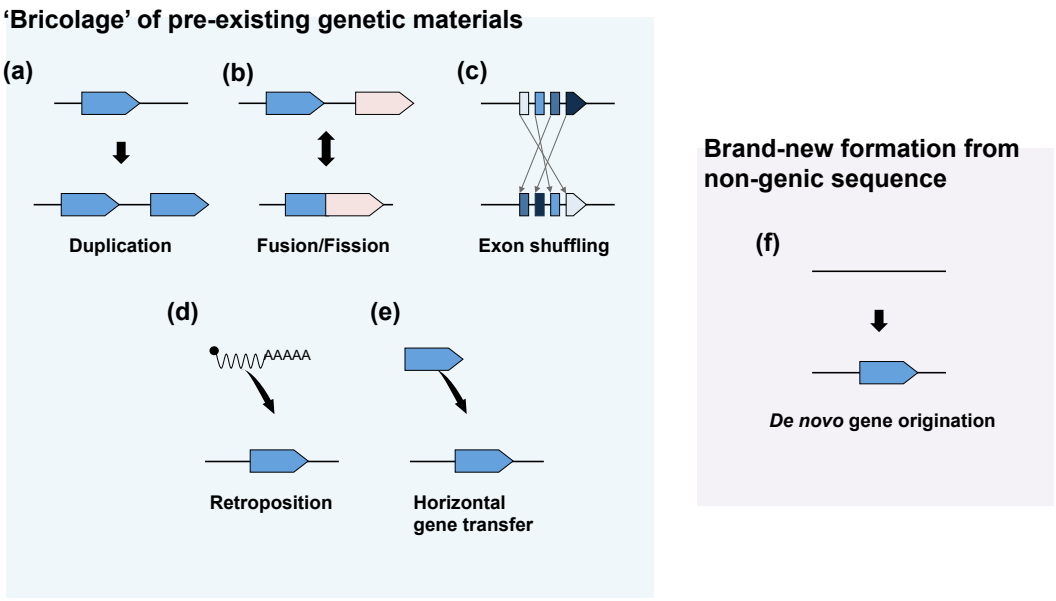


Figure 1.2. Schematic illustration of gene origination manners. (**a**) Gene duplication. (**b**) Gene fusion/ fission. (**c**) Exon shuffling. (**d**) Retroposition. (**e**) Horizontal gene transfer. (**f**) *de novo* gene origination. New gene origination manners can be categorized into two groups; 'bricolage' of pre-existing genetic materials (a–e), and *de novo* originated from non-coding DNA (f).

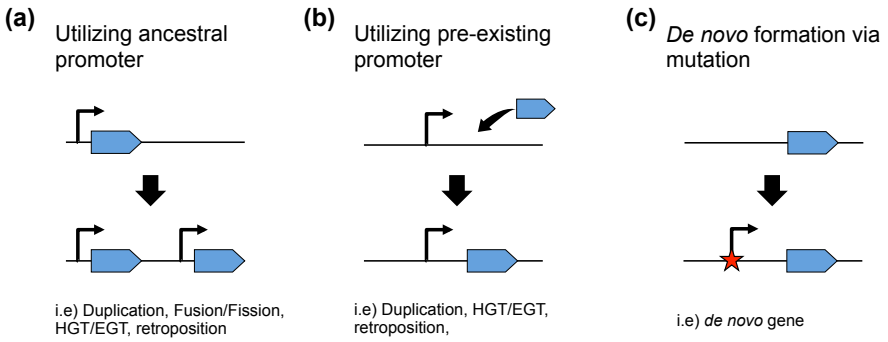


Figure 1.3. Schematic illustration of promoter acquisition mechanisms of newborn genes. (**a**) New genes except for *de novo* originated ones can arise together with their ancestral promoters. (**b**) New genes can acquire their promoter by 'high jacking' pre-existing promoters or transcripts. (**c**) New promoter can occur through sequence mutations.

Chapter 2:

Preculture in an enriched nutrient medium greatly enhances the *Agrobacterium*-mediated transformation efficiency in *Arabidopsis* T87 cultured cells

Summary of Chapter 2

The *Arabidopsis* T87 cell line has been widely used in both basic and biotechnological plant sciences. *Agrobacterium*-mediated transformation of this cell line was reported to be highly efficient when precultured in Gamborg's B5 medium for a few days. However, because we could not obtain the expected efficiency in our laboratory, we further examined the preculture conditions of *Arabidopsis* T87 cells in the *Agrobacterium*-mediated transformation. As a result, we found that preculture in an excess amount of Murashige and Skoog (MS) macronutrients before cultivation in the B5 medium enhanced the transformation efficiency up to 100-fold, based on the transformed callus number on selective gellan gum plates. In this study, transformants were labeled with green fluorescent protein (GFP), and we found multiple fluorescent spots on individual transgenic calli. Therefore, the actual number of transgenic clones seems much more than that of transgenic calli. In our MS macronutrient-rich culture condition, T87 cells tended to aggregate and formed bigger cell clumps, a change that might be related to the enhancement of transformation efficiency. Based on these results, we report an improved protocol of *Agrobacterium*-mediated transformation of *Arabidopsis* T87 cells with high efficiency.

Introduction

Plant cell culture is a useful system not only for basic plant sciences but also for genetic engineering to produce useful substances (Nagata et al. 1992, Ochoa-Villarreal et al. 2016, Eibl et al. 2018). Because the cell culture systems are sui to prepare cell populations of homogeneous physiological properties, they could provide highly reproducible experimental systems, compared with plant bodies or tissue samples.

The *Arabidopsis* T87 cell line (Axelos et al. 1992) is one of the widely used cultured cells for the following reasons: having photosynthetic ability under light illumination, transformation protocol is established and availability of highly reliable genomic data (Li et al. 2012, Li et al. 2018, Kwiatkowska et al. 2014). Ogawa et al. reported a highly efficient *Agrobacterium*-mediated transformation protocol of T87 cells (Ogawa et al. 2008), in which preculture in Gamborg's B5 medium (Gamborg et al. 1968) for a few days before cocultivation with *Agrobacterium* was crucial to obtain high transformation efficiency. In this study, we further examined the preculture conditions of this cell line in the *Agrobacterium*-mediated transformation and found that preculture with an excess amount of MS macronutrient (Murashige and Skoog 1962) before cultivation in B5 medium enhanced the transformation efficiency at least 100-fold. Incorporating this new finding, we now report an improved transformation protocol of *Arabidopsis* T87 cells.

Materials and Methods

Plant cell culture and transformation.

In this study, culture and transformation of *Arabidopsis* T87 cultured cells (Axelos et al. 1992) were carried out essentially according to Ogawa et al. (Ogawa et al. 2008) with slight modifications. The cells were cultured in mJPL3 medium (Ogawa et al. 2008) at 22 °C with shaking (120 rpm) under continuous light (50–70 μE m^{-2} s^{-1}). Two-week-old cultured cells were sieved through 1 mm stainless mesh and diluted to 60-fold by the following media (Table 2.1); mJPL3 medium, mJPL3+1/3MSmacro [JPL A (stock A of Axelos et al. 1992), 1/3 strength of Murashige and Skoog Plant Salt Mixture (392-00591, Nihon Pharmaceutical), Murashige and Skoog Vitamin Solution (M3900, Sigma), 15 g l^{-1} sucrose (30404-45, Nacalai Tesque), 0.1 g l^{-1}

casamino acids (392-00655, Nihon Pharmaceutical), 1 μM NAA (161-04021, Wako), 1% (v/v) 250 mM MES (pH5.9) (345-01625, Dojindo)] or mJPL3+MS [JPL A, Murashige and Skoog Plant Salt Mixture, Murashige and Skoog Vitamin Solution, 15 g l^{-1} sucrose, 0.1 g l^{-1} casamino acids, 1 μM NAA, 1% (v/v) 250 mM MES (pH 5.9)]. The detailed composition of the media used in this study is shown in Table 2.S1. Cells in the respective media were cultured at 22 °C with shaking under continuous light for 1 week, then harvested and 0.5 g wet weight aliquots were resuspended in 100 ml of B5 medium [Gamborg's B5 medium salt mixture (399-00621, Nihon Pharmaceutical), Gamborg's B5 vitamin mix (G-1019, Sigma), 1 μM NAA, 30 g l^{-1} sucrose, pH 5.9] or mJPL3+MS medium and cultured for 2 days. Subsequently, 5 ml aliquots of the cell cultures were cocultivated with 5 μL of *Agrobacterium* (GV3101) culture harboring pGreenII MH2 vector (Hellens et al. 2000, Hirashima et al. 2006) in a six-well plate. After 40 to 48 h of cocultivation, cells were washed three times with mJPL3 medium supplemented with 25 mg l^{-1} of meropenem (133-15671, Wako), then cultured on gellan gum (G1910, Sigma-Aldrich) plates (3 g l^{-1}) containing mJPL3 medium supplemented with 25 mg l^{-1} of meropenem and 30 mg l^{-1} of Kanamycin (113-00343, Wako). After two weeks of culture, green resistant calli were counted. White and yellowish calli were not counted because they were dead or escaped cells against the Kanamycin-based selection.

Results

Figure 2.1a represents the number of Kanamycin-resistant green calli on the plates from the cells precultured in the respective media, showing that increasing nutrient salt concentration in the preculture media resulted in a higher transformation efficiency. Transformants were hardly obtained when precultured in mJPL3, while those increased *ca* 30-fold and 100-fold when precultured in mJPL3+1/3MSmacro and mJPL3+MS, respectively.

Figure 2a, 2b, 2c shows the cells on the plates, corresponding to the three treatment samples in Figure 2.1. The cells precultured in mJPL3 hardly grew on the Kanamycin-containing plate and turned white (Figure 2.2a), those in mJPL3+1/3MSmacro grew to form calli (Figure 2.2b) and those in mJPL3+MS formed bigger green calli (Figure 2.2c).

Figure 2.2d, 2.2e represents the fluorescence and bright-field images of the cells as in Figure 2.2c, respectively. We found many green fluorescent protein (GFP) fluorescent spots on the

callus, indicating that the number of transformed cells was much higher than that of the green calli. Therefore, we expect that the preculture in the mJPL3+MS medium enhanced the transformation efficiency far more than 100-fold compared with that in mJPL3.

These results indicate that mJPL3+MS medium greatly enhances the *Agrobacterium*-mediated transformation efficiency when used for preculture medium. Next, we examined if mJPL3+MS also has an enhancing effect when used as a coculture medium. For this purpose, *Arabidopsis* T87 cells precultured in the mJPL3+MS medium for 1 week were transferred to fresh mJPL3+MS medium instead of B5 medium, and after 2 days, cocultivated with *Agrobacterium*. However, as shown in Figure 2.1b, we could not obtain a successful transformation. We observed that *Agrobacterium* could proliferate in mJPL4+MS medium, suggesting this condition may not suitable for the infection of *Agrobacterium*.

Based on these results, we propose an improved protocol for the highly efficient transformation of T87 cells (Figure 2.3). In this protocol, cells are precultured for 1 week in mJPL3+MS medium instead of mJPL3 medium. Subsequent steps of the transformation protocol are essentially the same as Ogawa's method (Ogawa et al. 2008). If isogenic clones are required rather than a massive number of transformants, we suggest a much shorter time of *Agrobacterium* cocultivation because *Agrobacterium* can sufficiently introduce T-DNA to plant genome as early as 6 h postinfection (Shilo et al. 2017).

Discussion

This protocol is very useful for plant biotechnology but raises the question of how preculture conditions affected the transformation efficiency. In this respect, we should first compare the composition of the tested media as shown in Table 2.1. These media share very similar components, and the difference mainly lies in their concentration (Table 2.S1). Though mJPL3 and mJPL3+1/3MSmacro have the same concentration of MS micronutrients (Table 2.1), their transformation efficiencies were quite different (Figure 2.1a). Therefore, the MS micronutrients' concentrations are less effective for the transformation efficiency, but the MS macronutrients should be the critical factor.

As another angle of the explanation of the transformation efficiency, we are interested in the cell clump size. When the clump size of the T87 cells was bigger, the introduction of the Cre

enzyme by electroporation was reported to be highly efficient (Furuhata et al. 2019). Analogous to this, *Agrobacterium*-mediated transformation efficiency may also be affected by the cell clump size. In this study, cell clump size tended to be bigger when cultured in mJPL3+MS (about 0.5 mm) (Figure 2.2f) than in mJPL3 medium (less than 0.1 mm) (Figure 2.2g). A possible explanation from this angle remains to be examined.

The protocol we propose in this study was really useful when we prepared massive transformants for a large-scale experiment utilizing next-generation sequencing and bioinformatics. Transformation efficiency of the cells is one of the critical factors for preparing transformant libraries for large-scale analysis (Akhtar et al. 2013, Inoue and Ahituv 2015). In this respect, this improved protocol could contribute to the advancement of future plant sciences.

References of Chapter 2

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(4), 914–927.

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.

Eibl, R., Meier, P., Stutz, I., Schildberger, D., Hühn, T. and Eibl, D. (2018) Plant cell culture technology in the cosmetics and food industries: current state and future trends. *Appl Microbiol Biotechnol,* 102(20), 8661–8675.

Furuhata, Y., Sakai, A., Murakami, T., Morikawa, M., Nakamura, C., Yoshizumi, T., Fujikura, U., Nishida, K. and Kato, Y. (2019) A method using electroporation for the protein delivery of Cre recombinase into cultured Arabidopsis cells with an intact cell wall. *Sci Rep,* 9(1), 2163.

Gamborg, O. L., Miller, R. A. and Ojima, K. (1968) Nutrient requirements of suspension cultures of soybean root cells. *Exp Cell Res,* 50(1), 151–158.

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(6), 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(22), 15385–15393.

Inoue, F. and Ahituv, N. (2015) Decoding enhancers using massively parallel reporter assays. *Genomics,* 106(3), 159–164.

Kwiatkowska, A., Zebrowski, J., Oklejewicz, B., Czarnik, J., Halibart-Puzio, J. and Wnuk, M. (2014) The age-dependent epigenetic and physiological changes in an Arabidopsis T87 cell suspension culture during long-term cultivation. *Biochem Biophys Res Commun,* 447(2), 285–291.

Li, B., Takahashi, D., Kawamura, Y. and Uemura, M. (2012) Comparison of plasma membrane proteomic changes of Arabidopsis suspension-cultured cells (T87 Line) after cold and ABA treatment in association with freezing tolerance development. *Plant Cell Physiol,* 53(3), 543–54.

Li, B., Takahashi, D., Kawamura, Y. and Uemura, M. (2018) Plasma Membrane Proteomics of Arabidopsis Suspension-Cultured Cells Associated with Growth Phase Using Nano-LC-MS/MS. *Methods Mol Biol,* 1696, 185–194.

Murashige, T. and Skoog, F. (1962) A Revised Medium for Rapid Growth and Bio Assays with Tobacco Tissue Cultures. *Physiologia Plantarum,* 15, 473–497

Nagata, T., Nemoto, Y. and Hasezawa, S. (1992) Tobacco BY-2 cell line as the “HeLa” cell in the cell biology of higher plants. *International Review of Cytology*, 132, 1–30

Ochoa-Villarreal, M., Howat, S., Hong, S., Jang, M. O., Jin, Y. W., Lee, E. K. and Loake, G. J. (2016) Plant cell culture strategies for the production of natural products. *BMB Rep,* 49(3), 149–158.

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(2), 242–250.

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(7), e1006875.

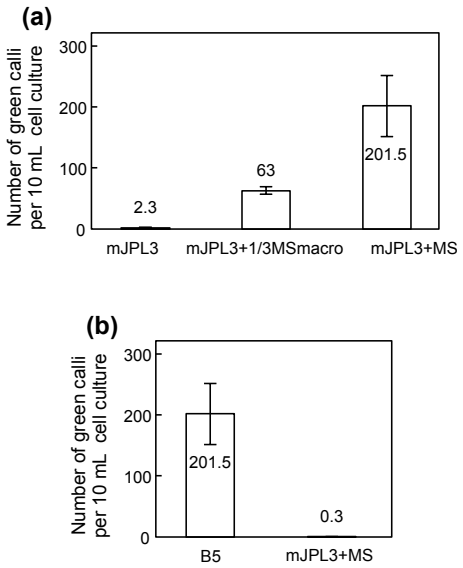


Figure 2.1. The number of Kanamycin-resistant green calli obtained per 10 ml of cell culture.
(**a**) Cells were pretreated by incubation for a week in the indicated media. (**b**) Cells were pretreated by incubation in the mJPL3+MS for a week. After that, the media were replaced the indicated media and cultured for two days, and then cocultivated with Agrobacterium. Mean ± SD of six independent plates are indicated.

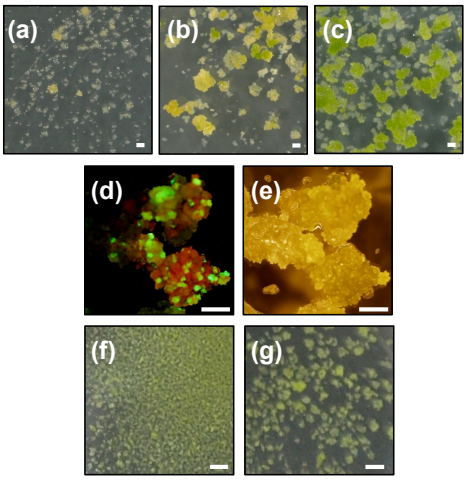


Figure 2.2. Photographs of the T87 cells.
(**a**–**c**) Green calli observed after two weeks of culture on the Kanamycin-containing plates. White and yellowish calli were not transformants. The preculture media were (a) mJPL3, (b) mJPL3+1/3MSmacro and (c) mJPL3+MS. (**d**) The green calli in (c) was observed using an OLYMPUS SZX7 system. Green and red spots indicate GFP fluorescence and fluorescence from chloroplasts, respectively. (**e**) Bright-field image of (d). (**f**–**g**) Cell clumps when cultured in mJPL3 (f), and in mJPL3+MS (g) for one week. Scale bar = 1 mm.

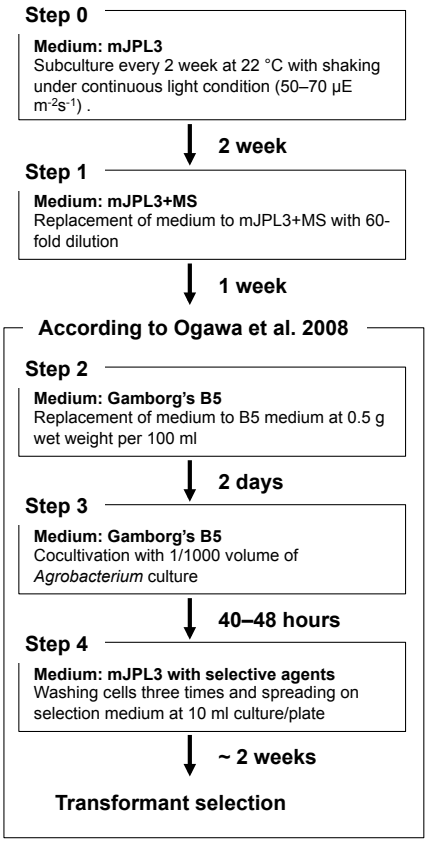


Figure 2.3. An improved protocol of efficient transformation of Arabidopsis T87 cells.
One week preculture in the mJPL+MS medium (step 1) is critical to obtain high efficiency.

Table 2.1. The concentration of MS nutrients in preculture media tested in this study

		mJPL3	mJPL3+1/3MSmacro	mJPL3+MS	B5
MS macronutrients	N	0.3	0.7	1.3	0.4
	K	1.0	1.3	2.0	1.2
	P	0.3	0.6	1.3	0.8
	Ca	0.3	0.6	1.3	0.3
	Mg	0.3	0.6	1.3	0.7
MS micronutrients		0.3	0.3	1.0	0.8

Each concentration was normalized by the original composition of MS medium (Murashige and Skoog 1962) as 1× strength.

Supplementary information of Chapter 2

Table 2.S1. The components of media used in this study.

Component (mg l^{-1})		MS (Murashige and Skoog 1962)	mJPL3 (Ogawa et al. 2008)	mJPL3+1/3×MS	mJPL3+MS	B5 (Gamborg et al. 1968)
MS macronutrients	KNO_3	1900	1965	2598	3865	2500
	NH_4NO_3	1650	-	550	1650	-
	$(NH_4)_2SO_4$	-	-	-	-	134
	KH_2PO_4	170	51	108	221	-
	$NaH_2PO_4 \cdot H_2O$	-	-	-	-	150
	$CaCl_2 \cdot 2H_2O$	440	132	279	572	150
	$MgSO_4 \cdot 7H_2O$	370	111	234	481	250
MS micronutrients	H_3BO_3	6.2	2.1	2.1	6.2	3.0
	$MnSO_4 \cdot 5H_2O$	22.3	-	7.4	22.3	-
	$MnSO_4 \cdot H_2O$	-	5.6	-	-	10.0
	$ZnSO_4 \cdot 7H_2O$	8.6	2.9	2.9	8.6	2.0
	KI	0.83	0.28	0.28	0.83	0.75
	$Na_2MoO_4 \cdot 2H_2O$	0.25	0.08	0.08	0.25	0.25
	$CoCl_2 \cdot 6H_2O$	0.025	0.01	0.01	0.025	0.025
	$CuSO_4 \cdot 5H_2O$	0.025	0.01	0.01	0.025	0.025
	$Na_2 \cdot EDTA$	37.3	12.4	12.4	37.3	37.3
	$FeSO_4 \cdot 7H_2O$	27.8	9.3	9.3	27.8	27.8
Vitamins	Myoinositol	100	100	100	100	100
	Glycine	2.0	2.0	2.0	2.0	-
	Nicotinic acid	0.50	0.50	0.50	0.50	1.0
	Pyridoxine・HCl	0.50	0.50	0.50	0.50	1.0
	Thiamine・HCl	0.10	0.10	0.10	0.10	10
Amino acids	Casamino acids	-	100	100	100	-
Plant hormone	NAA・K	-	0.22	0.22	0.22	-
Carbon source	Sucrose	-	15000	15000	15000	-

Adjust the pH of the each medium by adding 1% (v/v) of 250 mM MES (pH5.9) in mJPL3, mJPL3+1/3×MS, and mJPL3+MS, or to 5.9 using 1N KOH in MS and B5.

Chapter 3:

Plant genome response to incoming coding sequences: stochastic transcriptional activation independent of integration loci

Summary of Chapter 3

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 eukaryotic genome? 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 mapped 4,504 promoterless *LUC* inserts on the *A. thaliana* chromosomes, and found that about 30% of them were transcribed. Only a small portion of them were explained by the conventional transcriptional fusions with the annotated genes, and the remainder occurred in a quite different manner; (1) they occurred all over the chromosomal regions, (2) independently of the insertion sites relative to the annotated gene loci, inherent transcribed regions, or heterochromatic regions, and (3) with one magnitude lower transcriptional level than the conventional transcriptional fusions. This type of transcriptional activation occurred at about 30% of the inserts, raising a question as to what this 30% means. We tested two hypotheses: the activation occurred at 30% of the entire chromosomal regions, or stochastically at 30% of each insertion event. Our experimental analysis indicates that the latter model could explain this transcriptional activation, a new type of plant genome response to the incoming coding sequences. We discuss the possible mechanisms and evolutionary roles of this phenomenon in the plant genome.

Introduction

The process via which genetic novelty emerges has been a fundamental question of evolutionary biology. Because of the advancement of comparative genomics, our knowledge of new gene origination has been expanded; genes can be generated through the “bricolage” of pre-existing genetic materials, or can be originated *de novo* from non-coding DNA (Kaessmann, 2010; Cardoso-Moreira and Long, 2012; McLysaght and Guerzoni, 2015; Van Oss and Carvunis, 2019).

An essential question of gene birth is how newly originated gene sequences acquire their transcriptional competency, because it is a prerequisite for the mere sequences to become genes. Transcriptional competency is driven by a promoter, in which a specific sequence of elements and chromatin configuration exist for pre-initiation complex (PIC) binding and the initiation of transcription at a precise genomic position (Haberle and Stark, 2018; Andersson and Sandelin, 2020). As promoters activate the transcription of downstream DNA sequences, their evolution should be intrinsically connected to the functionalization of new genes. Comparative genomics has revealed that evolutionarily young genes acquired their transcriptional competency through (1) the utilization of duplicated ancestral promoters, (2) hijacking of pre-existing genes, promoter-like elements or spurious transcription units or (3) *de novo* emergence through mutations (Kaessmann, 2010; Li, Lenhard and Luscombe, 2018; Van Oss and Carvunis, 2019; Zhang *et al.*, 2019). However, the promoters of such evolutionarily young genes are not so “young”, as they had been fixed in the genome through natural selection over a certain evolutionary period. Therefore, little knowledge is available regarding how newly originated coding sequences are transcribed and start evolving after their birth.

Experimental evolution is another approach to scrutinize such gene evolutionary processes, as it enables the analysis of “truly young” genes by mimicking the process of new gene origination in the native genomic environment (Garland, 2009). In plants, exogenously introduced coding sequences that mimic the originated genes through horizontal or endosymbiotic gene transfer (HGT/EGT) events have provided insights about how such newborn coding sequences acquire transcription ability. The escape of plastid DNA to the nucleus suggests that transferred plastid genes become transcriptionally active by trapping neighbouring eukaryotic promoters or by utilizing the prokaryotic plastid promoter sequences (Stegemann and Bock, 2006; Wang *et al.*, 2014). By introducing promoterless coding sequences

into the genome, promoter/gene-trapping screening also simulates gene origination processes, and 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 (Friedrich and Soriano, 1991; Fobert *et al.*, 1994; Topping *et al.*, 1994; Springer, 2000; Mollier *et al.*, 2000; Plesch, Kamann and Mueller-Roeber, 2000; Yamamoto *et al.*, 2003; Sivanandan *et al.*, 2005; Stangeland *et al.*, 2005). However, molecular identities of these cryptic promoters have long been unsolved.

Recently, Kudo *et al.* demonstrated that such unexpected transcriptional activation in gene-/promoter-trapping experiments occurred via at least two different mechanisms in the plant genome: (1) cryptic promoter capturing, in which exogenous DNA was transcribed by trapping a preexisting promoter-like chromatin configuration that is not associating with annotated genes; and (2) promoter *de novo* origination, in which promoter-like epigenetic landscapes were newly formed via chromatin remodeling triggered by the insertion of a coding sequence (Kudo *et al.*, 2020). It should be noted that these two mechanisms could endow transcriptional activity to the incoming coding sequences without disturbing the preexisting nuclear gene network. In examining whether these cryptic promoters could be a source of transcriptional activation in massive gene transfer, we should know how often the cryptic promoter activation occurs in the whole nuclear genome.

In this Chapter 3, we applied a massively parallel reporter assay (Akhtar *et al.*, 2013; Inoue and Ahituv, 2015) to the conventional gene-/promoter-trapping experiments and carried out a genome-wide “transgene location vs. expression” scan. We introduced thousands of promoterless coding sequences of firefly luciferase (LUC) genes as a model of transferred genes into the genome of *Arabidopsis thaliana* cultured cells, and examined the manners by which transcriptionally inert transgenes become activated in the foreign genome environment. We found that a small portion of the transcriptional activation of transgenes was explained by the conventional gene-/promoter-trapping mechanism, but the majority of promoterless *LUC* inserts were transcriptionally activated in a quite different manner, i.e., integration-dependent stochastic transcriptional activation. This transcriptional activation occurred stochastically at about 30% of each insertion event, independently of the integration locus relative to the preexisting genes, inherent transcribed regions, or heterochromatic regions. We discuss the likely mechanism of this transgene activation phenomenon and refer to its possible contribution to the initial

transcriptional activation process of HGT/EGT during plant genome evolution.

Materials and Methods

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 3.S1a and Methods 3.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 method (see Chapter 2) (Hata *et al.*, 2020). 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 3.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 3.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 3.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 3.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 3.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 STAR (Dobin *et al.*, 2013) and StringTie (Pertea *et al.*, 2015), 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).

Results

General view of the transgene expression over the entire genome

To understand the rules that govern the transcriptional activation of alien incoming genes, we introduced thousands of promoterless luciferase (*LUC*) genes into *A. thaliana* T87 suspension-cultured cells via *Agrobacterium*-mediated transformation (Figure 3.S1a). 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) (Figure 3.S1a–c). 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. Based on this scheme, we determined individual insertion loci and corresponding transcription levels in 4,504 *LUC* genes (Figure 3.1, and 3.S1a–c, and 3.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 3.1a). The relative abundances of the *LUC* genes inserted in the intergenic, genic, and promoter regions were roughly proportional to the relative lengths of these genomic regions (Figure 3.1b and 3.S3). On the fine distribution map of the inserts, genic promoter regions (~200 bp) were more prone to be inserted than the other regions by about threefold (Figure 3.1b, and 3.S3 and 3.S4), in accordance with a relatively open chromatin configuration of the promoter region. Despite such slight biases, the *LUC*-mapped loci covered entire chromosomal regions (Figure 3.1a and b), and thus were suitable for the genome-wide scanning of transgene transcriptional activation events.

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 3.1d). 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 (Figure 3.S5 and 3.S6).

Identification of two distinct mechanisms of transgene transcriptional activation

In the simplest-case scenario, promoterless *LUC* transcription is a result of the trapping of endogenous transcription units. To test this conventional model, we classified the 4,504 *LUC* loci into five insertion types in relation to the annotated genes: (i) sense and (ii) antisense orientation

within the gene-coding regions, (iii) sense and (iv) antisense orientation in the promoter regions, and (v) intergenic regions. According to this classification, 25–30% of the *LUC* genes in each insertion type were transcribed, except for the genic-sense insertion type; about 50% of them were transcribed in the genic-sense fraction (Figure 3.1c). Why are the genic-sense inserts more prone to be transcribed?

As shown in Figure 3.1d, 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 that of the genic sense type exceptionally high, at the level of 10^5. The comparison of the distribution profiles of the five insertion types revealed that the genic-sense type had a superposed fraction (light-blue fraction in Figure 3.1d) at higher transcription levels (10^5–10^7). Without this superposed fraction, the distribution curves of the five *LUC* insertion types were remarkably similar (Figure 3.1d). To explain this result, we next examined the *LUC* insertion sites relative to the annotated genic transcription start sites (TSSs) and *LUC* transcription levels (Figure 3.1e). We found that the *LUC* inserts with higher transcription levels (10^5–10^7) were more abundant at 0.2–2.4 kb downstream of the annotated TSSs (Figure 3.1e). Without this superposed fraction in this region, the expressed inserts appeared to be similarly distributed both within and outside of the annotated transcribed regions (Figure S3.7). In *A. thaliana*, the median lengths of the 5′ untranslated regions (UTRs) and mRNAs are ~70 and ~1,900 bp, respectively (as calculated from the TAIR10 database, https://www.arabidopsis.org/index.jsp); thus, the region 0.2–2.4 kb downstream from the annotated TSS roughly corresponds to the intrinsic protein-coding regions. Based on these observations, the *LUC* inserts of the genic-sense type appeared to be transcribed at least in part by the conventional gene-trapping mechanism, in addition to the transgene transcription mechanism that similarly occurred over the entire genome.

If our above assumption is the case, the contribution of the conventional gene-trapping to the whole transcriptional activation of the incoming coding sequences is small; rather, the majority of transcriptional activation occurred by the distinct mechanism, even within the genic-sense insertion type (Figure 3.1d–f). The mean transcription level of this transcriptional activation was 10^4, which was one magnitude lower than that of the conventional transcriptional fusions (Figure 3.1d and e). To confirm that this whole expression profile was not a sequencing artifact, we performed similar analyses using more reliable datasets (i.e., the *LUC* inserts whose sequencing reads were more highly abundant than the background level) with elevated read number threshold. Irrespective of the threshold read numbers, two distinct fractions corresponding to the

gene-trapping type (light-blue fraction in Figure 3.S8a and b) and the other type (light-red fraction in Figure 3.S8a and b) were clearly detected, as in Figure 3.1d. In addition, these distribution profiles were confirmed by three biologically independent samples (Figure 3.S8c–e). Based on these analyses, we concluded that the low-level transcriptional activation of transgenes that occurred over the entire chromosomal regions was not a sequencing artifact.

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

Pervasive transcription throughout the genome characterizes eukaryotic organisms. We asked whether the genome-wide transcription of the *LUC* genes could be explained by the 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 transgenic and WT cells (Figure 3.2a). Unexpectedly, only 7.8% of the *LUC* genes were transcribed in the inherently transcribed genomic regions (type (iii) in Figure 3.2a), whereas 22.3% of the *LUC* genes were transcribed in the transcriptionally inert regions (type (i) in Figure 3.2a). As for the 7.8% of the *LUC* genes (type (iii) in Figure 3.2a), we compared the transcription levels between the transgenic and WT cells, but no correlation was found (Figure 3.2b, $r = 0.21$). Two conclusions were drawn from this analysis: (1) transcriptional activation of the *LUC* inserts occurs independently of the inherent transcriptional status of the genomic region where the *LUC* was inserted; and (2) the transcriptional activities of the *LUC* inserts do not reflect the inherent transcriptional activities of the given genomic regions.

Transcriptional activation of promoterless *LUC* genes was not affected by the inherent heterochromatic status

We wondered whether *LUC* transgenes could overcome the silencing effects of the 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 heterochromatic regions (Bühler and Moazed, 2007; Grewal and Jia, 2007; Shu *et al.*, 2012). 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 regions. In the transgenic cells, only 120 *LUC* genes were inserted into the H3K9me2-containing heterochromatic regions (see the legend of Figure 3.3), indicating that the integration frequency in this region was one-seventh of the rest of the

genome. However, in the H3K9me2-containing region, 28% of the *LUC* inserts were transcriptionally activated and their activation profiles were similar to the other regions (Figure 3.3a). The transcription levels of these *LUC* genes did not show any correlation with the degree of H3K9me2 modification (Figure 3.3b). Furthermore, two transcribed *LUC* genes were located 63 kb and 682 kb from the centromeres (Figure 3.S9), and these regions were covered by pericentromeric heterochromatin (Bernatavichute *et al.*, 2008). Taken together, we concluded that the transcriptional activation of the *LUC* inserts occurred at a rate of about 30% irrespective of the inherent heterochromatic status.

Integration-dependent stochastic activation of transgene transcription

As described above, transgene transcriptional activation was observed for 30% of the *LUC* inserts, which raised a question: What does this 30% mean? To account for this question, we hypothesized two models: (i) the transcriptional activation occurred at 30% of the entire *A. thaliana* chromosomal regions; or (ii) stochastically at 30% of each insertion event. To test which model is suitable for this transcriptional activation, we analyzed the transcriptional behavior of *LUC* genes that were integrated into close neighboring locations (Figure 3.4a). 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 3.4b). If the transgene transcriptional activation depends on the chromosomal locus, the transcriptional fates of neighboring *LUC* inserts are expected to be similar (Figure 3.4c). Hence, in this scenario, only Fates A and C would be observed for the *LUC* pairs (Figure 3.4c). Moreover, the expected ratio between Fates A and C would be 30:70 (Figure 3.4c), assuming an average transcriptional activation rate of 30%. Conversely, as shown in Figure 3.4d, 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. In this model, the distribution ratio among Fates A, B, and C would be 9, 42, and 49, respectively (Figure 3.4d). 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 3.4e, upper panel). This distribution fits the integration-dependent stochastic

transcriptional activation model (Figure 3.4d), rather than the chromosomal-locus-dependent model (Figure 3.4c). 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 3.4f). 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 3.4e, 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 3.4f). It should be emphasized that the individual *LUC* inserts used for this integration-site neighborhood analysis stemmed from different, independently transformed cells that passed through ~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 Chapter 3, we performed a genome-wide screening of promoter-trapping events covering both expressed and unexpressed inserts for the first time, using a non-selective reporter. Collectively, the data revealed a new type of transgene transcriptional activation of the plant genome, which occurs stochastically at about 30% of each DNA integration event but not depending on the chromosomal loci. This transcriptional activation occurred in the transgenic cells that experienced only ~10 times cell divisions after the transgene integration, indicating that it is an immediate response of the plant genome to the incoming coding sequences. To date, we could not find any specific motifs that were enriched at the 5′ proximal regions of the transcribed *LUC* genes, which was quite a different situation from the annotated gene promoters (Figure 3.S10). How can we explain the mechanism of this new type of transcriptional activation that is stochastic and independent of the DNA sequences surrounding the transgene insertion sites?

It is generally accepted that T-DNA is integrated into the host genome following the double-stranded DNA breaks, which are repaired predominantly by the non-homologous DNA end-joining (Magori and Citovsky, 2011; Kleinboelting *et al.*, 2015). This repair process remodels the chromatin and leaves so-called DNA damage scars in the chromatin epigenetic structure (Soria, Polo and Almouzni, 2012; Dabin, Fortuny and Polo, 2016). This chromatin remodeling may account, at least in part, for the integration-dependent stochastic transcriptional activation.

From this viewpoint, it is intriguing to compare the chromatin structures before and after the *LUC* integration, but this analysis remains technically challenging. In the present study, the established transgenic cell pools were highly heterogeneous; they contained thousands of distinct transgenic cell lines, and each cell line consisted of only ~1,000 cells. There is no practical methodology to analyze epigenetic configurations of each transgenic line from such a heterogeneous cell population (discussed in Chapter 6).

TSS determination is another approach to investigate the molecular mechanism of this transcriptional activation. It would provide useful information to depict the transcription initiation mechanism and cis-regulatory elements in this new type of plant genome response. However, the TSS analysis has the same problems as found in chromatin analysis; the molecular abundances of each *LUC* mRNA are quite low and the mRNA sample consists of heterogeneous mRNAs from thousands of different transgenic lines. Therefore, the analysis needs technical breakthroughs, which is described in Chapter 4.

In the present study, we characterized a novel plant genome response to the incoming coding sequences, i.e., integration-dependent stochastic transcriptional activation. Contrary to the conventional gene-/promoter-trapping scenario in the HGT/EGT process, this foreign gene activation mechanism seems less harmful to the host nuclear gene networks because this mechanism does not cause disruption of the preexisting nuclear genes. Therefore, this finding provides a new angle for examining the gene activation mechanism in the massive gene transfer events between phylogenetically distant organisms. To evaluate the biological contribution of this novel genome response to plant genome evolution, further information is needed on how activated transcription via this mechanism continues and behaves over generations (see Chapter 5), and how selective pressure on the activated transcriptions affects their fates. Experimental studies along these lines could open the way to an understanding of how the initial molecular response of the eukaryotic genome is linked to the phenotypic evolution.

References of Chapter 3

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(4), 914–927.

Andersson, R. and Sandelin, A. (2020) Determinants of enhancer and promoter activities of regulatory elements. *Nat Rev Genet,* 21(2), 71–87.

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(9), e3156.

Bühler, M. and Moazed, D. (2007) Transcription and RNAi in heterochromatic gene silencing. *Nat Struct Mol Biol,* 14(11), 1041–1048.

Cardoso-Moreira, M. and Long, M. (2012) The origin and evolution of new genes. *Methods Mol Biol,* 856, 161–186.

Dabin, J., Fortuny, A. and Polo, S. E. (2016) Epigenome Maintenance in Response to DNA Damage. *Mol Cell,* 62(5), 712–727.

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(1), 15–21.

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(4), 567–577.

Friedrich, G. and Soriano, P. (1991) Promoter traps in embryonic stem cells: a genetic screen to identify and mutate developmental genes in mice. *Genes Dev,* 5(9), 1513–1523.

Garland, T. (2009) Experimental Evolution: Concepts, Methods, and Applications of Selection Experiments. *In:* Rose, M.R. (ed.). University of California Press: Berkeley.

Grewal, S. I. and Jia, S. (2007) Heterochromatin revisited. *Nat Rev Genet,* 8(1), 35–46.

Haberle, V. and Stark, A. (2018) Eukaryotic core promoters and the functional basis of transcription initiation. *Nat Rev Mol Cell Biol,* 19(10), 621–637.

Hata, T., Mukae, K., Satoh, S., Matsuo, M. and Obokata, J. (2021) Preculture in an enriched nutrient medium greatly enhances the *Agrobacterium*-mediated transformation efficiency in *Arabidopsis* T87 cultured cells. *Plant Biotechnology.* 38, 179–182.

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(6), 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(22), 15385–15393.

Inoue, F. and Ahituv, N. (2015) Decoding enhancers using massively parallel reporter assays. *Genomics,* 106(3), 159–164.

Kaessmann, H. (2010) Origins, evolution, and phenotypic impact of new genes. *Genome Res,* 20(10), 1313–1326.

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(11), 1651–64.

Kudo, H., Matsuo, M., Satoh, S., Hachisu, R., Nakamura, M., Yamamoto, Y., Yoshiharu, Hata, T., Kimura, H., Matsui, M. and Junichi, O. (2020) Cryptic promoter activation occurs by at least two different mechanisms in the *Arabidopsis* genome. unpublished data, *bioRxiv* [posted 2020 Nov 28]. Available from: https://www.biorxiv.org/content/10.1101/2020.11.28.399337v1 doi: 10.1101/2020.11.28.399337

Langmead, B. and Salzberg, S. L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods,* 9(4), 357–359.

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(5), 676–688.

Magori, S. and Citovsky, V. (2011) Epigenetic control of Agrobacterium T-DNA integration. *Biochim Biophys Acta,* 1809(8), 388–394.

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(1678), 20140332.

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(11), 1076–1083.

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(2), 242–250.

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(3), 290–295.

Plesch, G., Kamann, E. and Mueller-Roeber, B. (2000) Cloning of regulatory sequences mediating guard-cell-specific gene expression. *Gene,* 249(1-2), 83–9.

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(6), 1018–1025.

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

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(3), 202–208.

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(6), 722–734.

Springer, P. S. (2000) Gene traps: tools for plant development and genomics. *Plant Cell,* 12(7), 1007–1020.

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(419), 2495–2505.

Stegemann, S. and Bock, R. (2006) Experimental reconstruction of functional gene transfer from the tobacco plastid genome to the nucleus. *Plant Cell,* 18(11), 2869–2878.

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(4), e1002055.

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(6), 895–903.

Van Oss, S. B. and Carvunis, A. R. (2019) De novo gene birth. *PLoS Genet,* 15(5), e1008160.

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(6), 1327–1334.

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(2), 273–283.

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(4), 679–690.

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(9), R137.

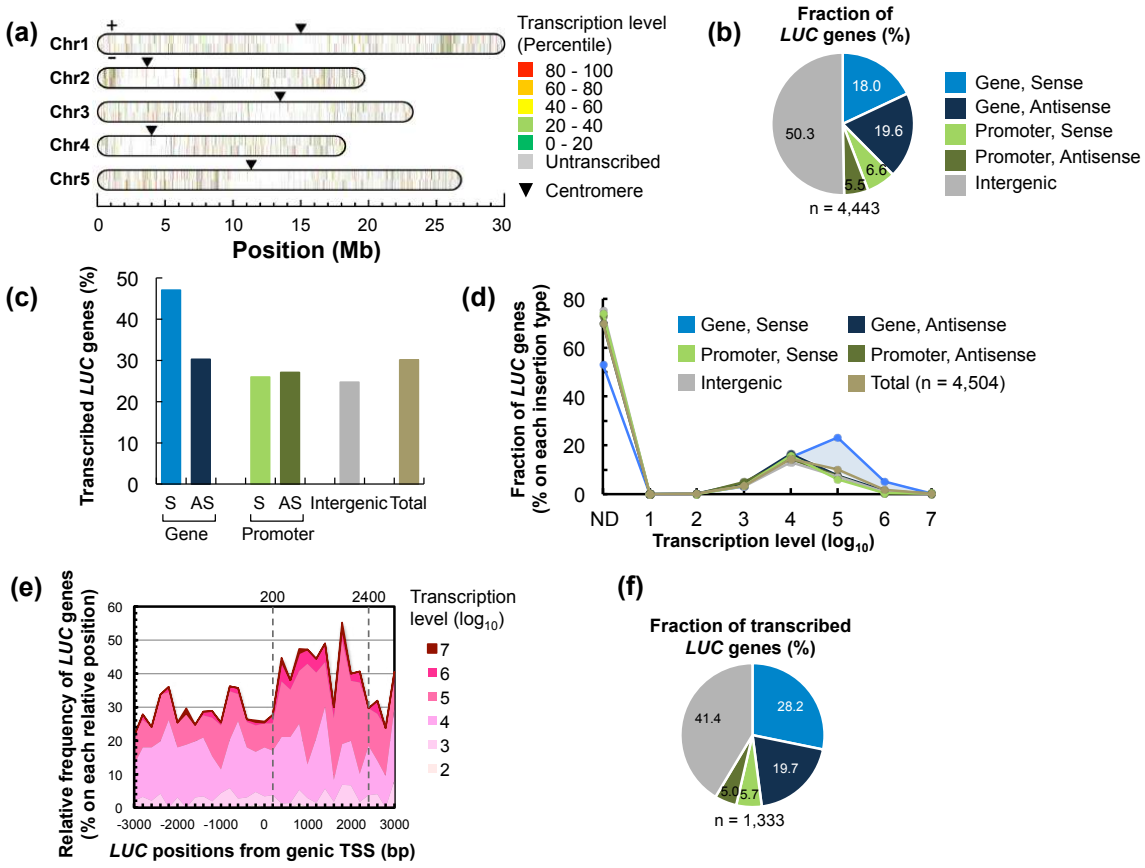


Figure 3.1. Massively parallel promoter-trapping analysis of the *Arabidopsis thaliana* genome.
(**a**) Genomic positions of the inserted promoterless *LUC* genes and respective transcription levels. The bars represent the 4,504 mapped *LUC* genes regarding their orientation towards the upper (+) or 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. The *LUC* genes that cannot be classified their insertion type uniquely were omitted. (**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. (**e**) Abundance of the *LUC* genes with the indicated transcription levels in relation to the distance from the genic TSS in each window (200 bp). (**f**) Classification of the transcribed *LUC* genes according to their insertion types, as in (b).

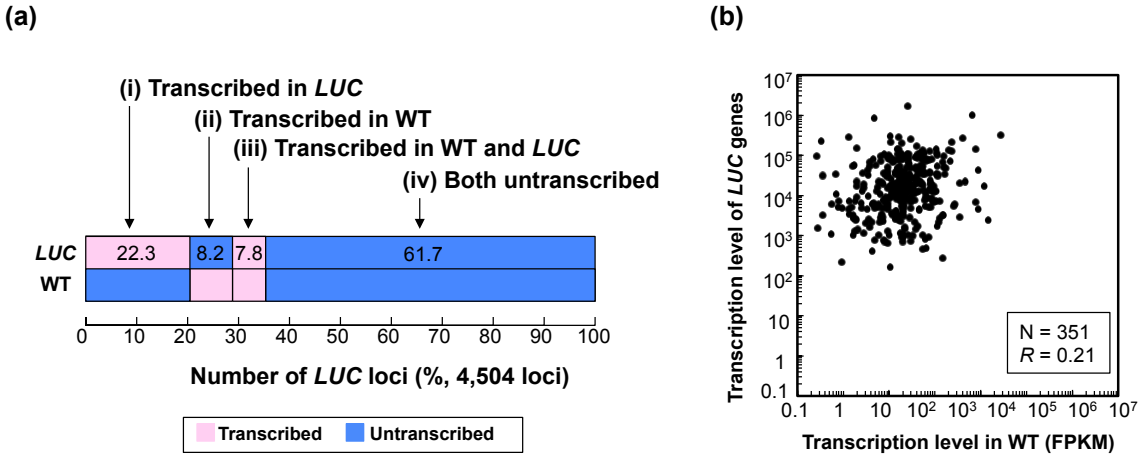


Figure 3.2. Transcription states of the *LUC* loci in WT and transgenic cells.
(**a**) The 4,504 *LUC* loci were clustered into four groups according to the combination of on/off transcription states in WT and transgenic cells. The local transcription landscape in WT cells was determined based on the 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.

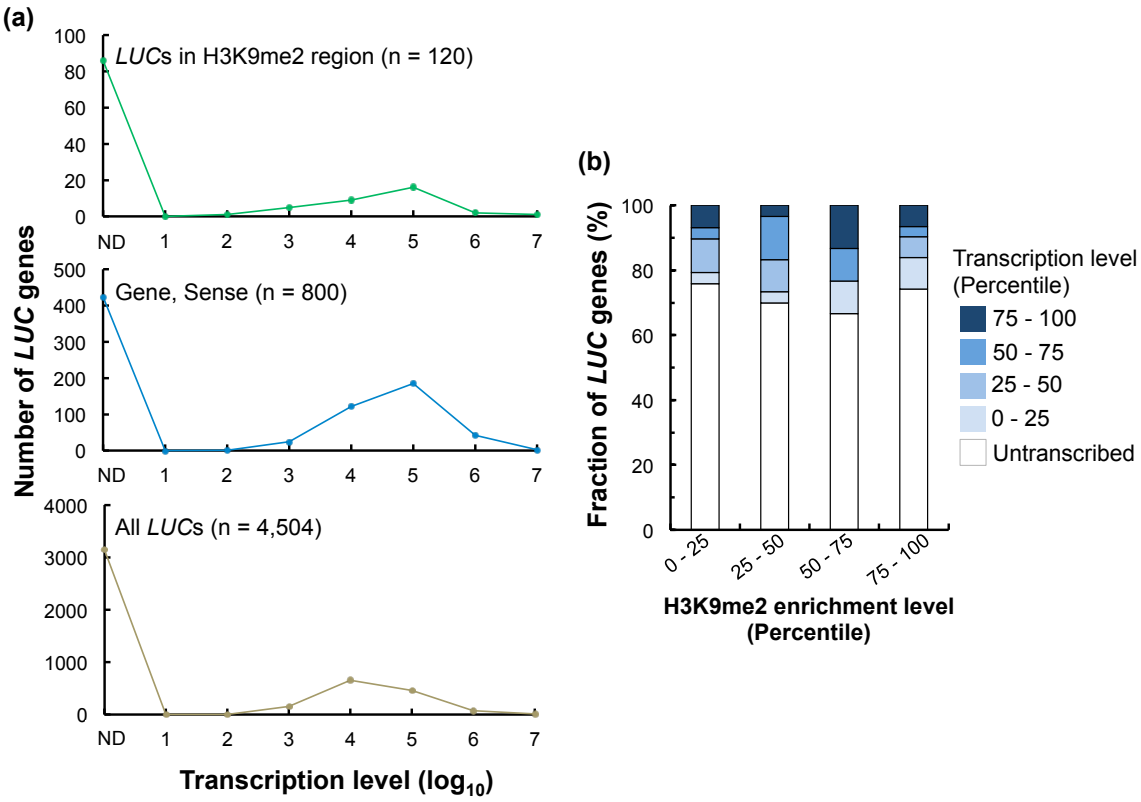
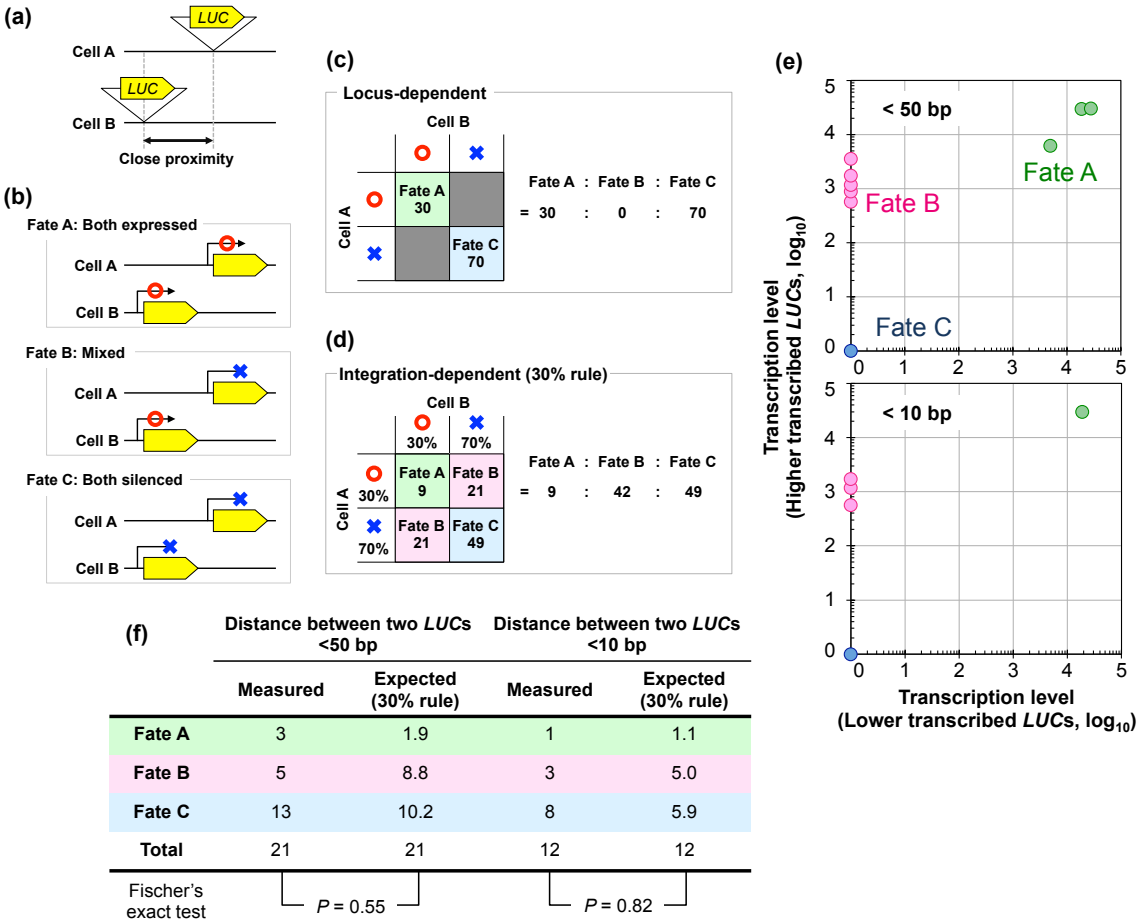


Figure 3.3. Transcription states of the *LUC* genes in the heterochromatic regions.
(**a**) The upper panel shows the transcription profile of the *LUC* genes in the heterochromatic regions. The middle and bottom panels are derived from Figure 3.1d and represent the transcription profiles of the genic-sense 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 3.4. 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.

Supplemental information of Chapter 3

Methods 3.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 3.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 3.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 3.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 3.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 of supporting information

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.

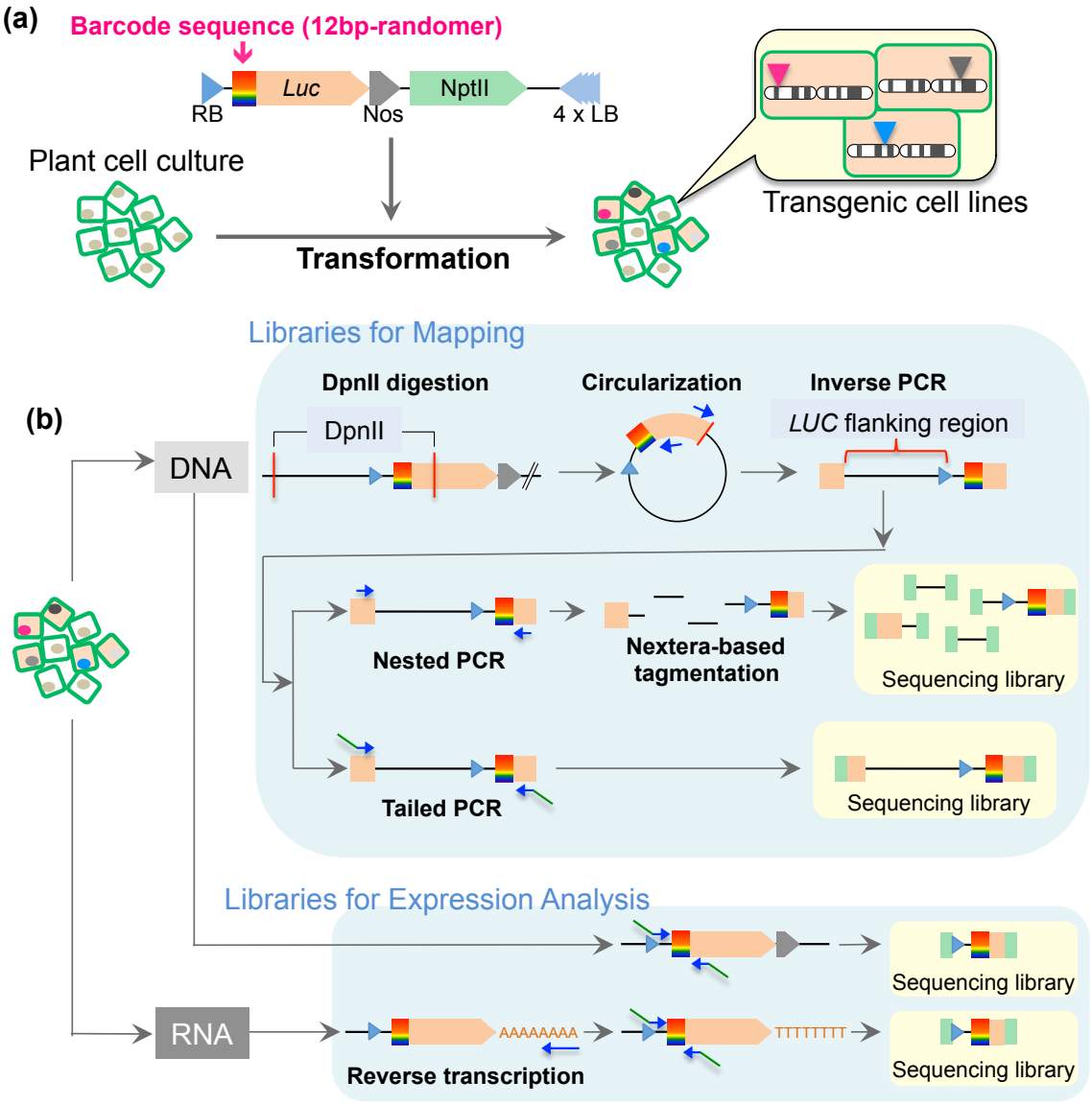


Figure 3.S1. Precise workflow of the promoter analysis that was performed using the TRIP system.
(**a**) Transformation of multiplexed barcoded vectors into *Arabidopsis* T87 suspension-cultured cells. (**b**) Preparation of sequencing libraries for the mapping and expression analyses. To prepare the 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.

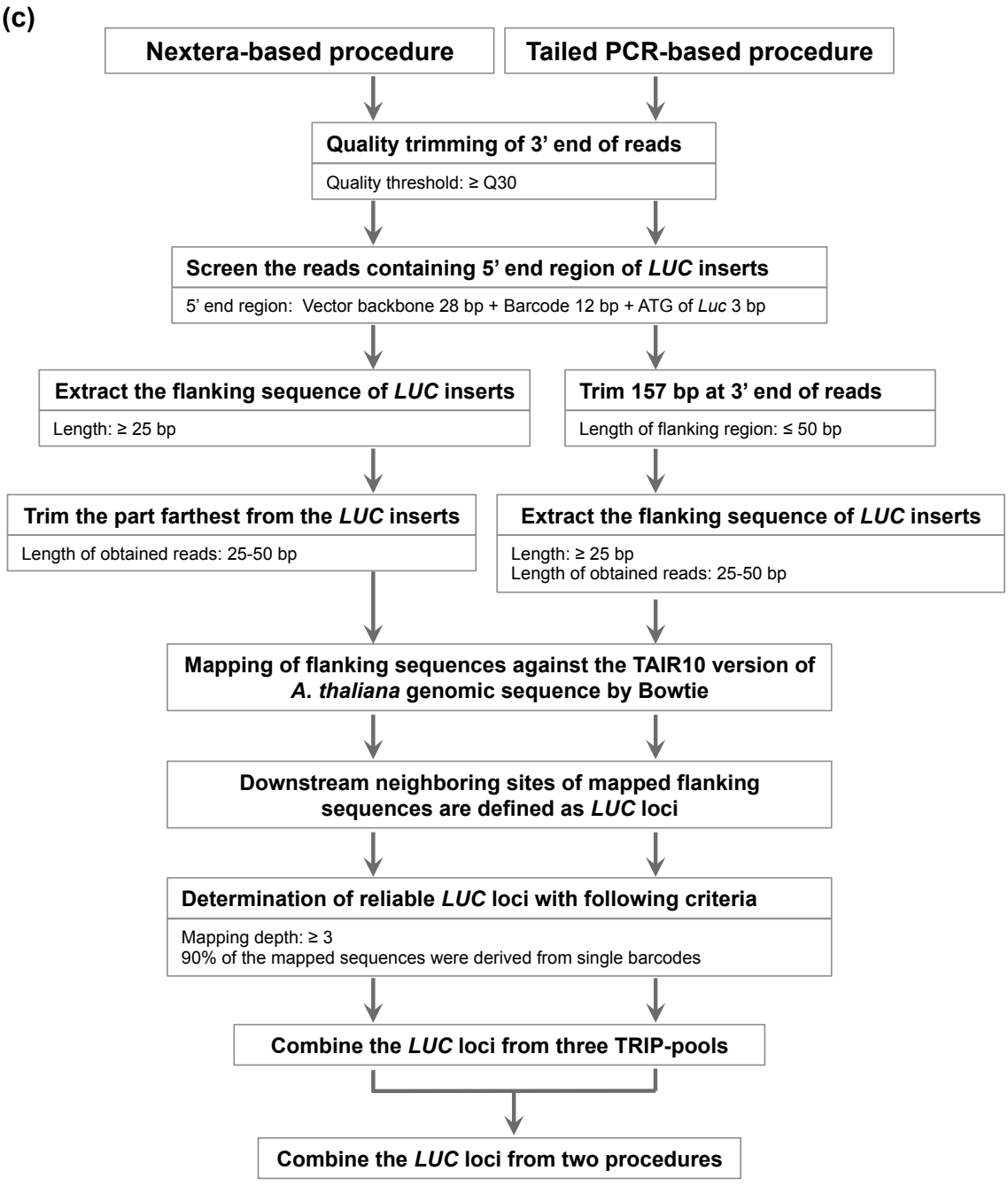


Figure 3.S1. Precise workflow of the promoter analysis that was performed using the TRIP system.
(**c**) Workflow of the data-analysis pipeline that was used for the mapping of *LUC* genes. The flanking sequences of the *LUC* genes were extracted from the Nextera-based mapping library and tailed-PCR-based mapping library using slightly different methods. The *LUC* loci obtained were combined in the final step.

(d)

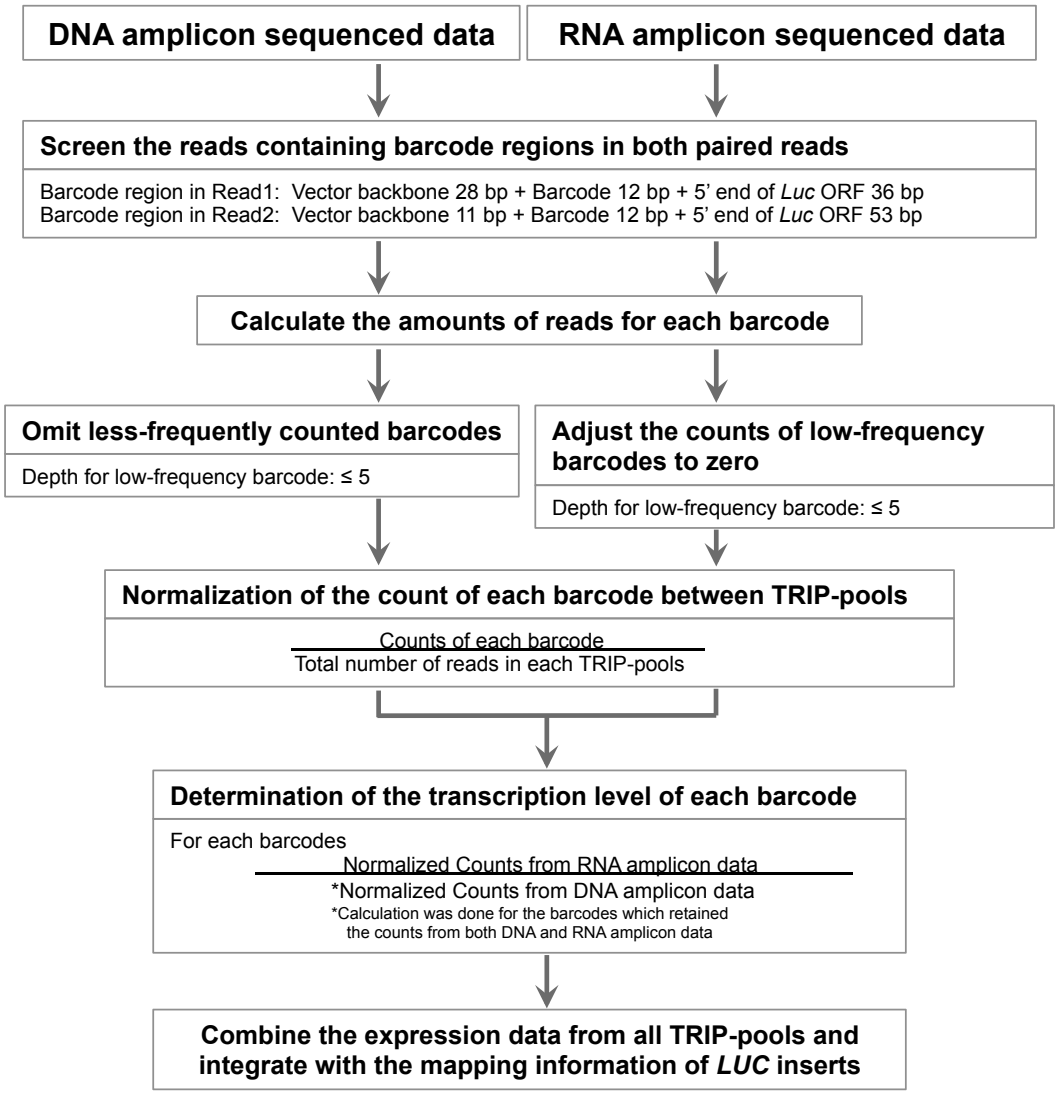


Figure 3.S1. Precise workflow of the promoter analysis that was performed using the TRIP system.
(**d**) Flow diagram used for the determination of *LUC* transcription levels. The transcription level data obtained for individual barcodes were associated with the respective mapped *LUC* genes and used in subsequent analyses.

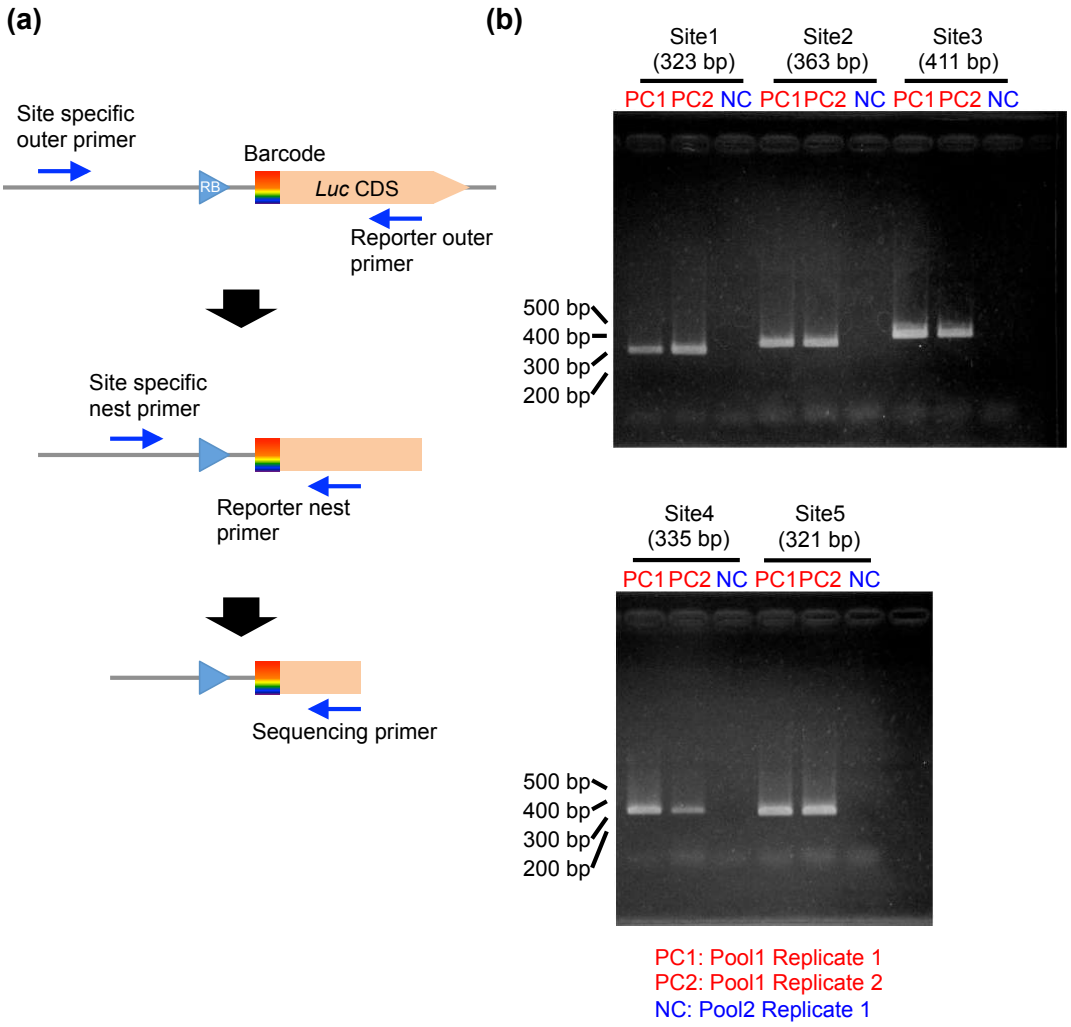


Figure 3.S2. Validation of the *LUC* mapped loci and barcode sequences via PCR amplification in five representative samples.
(**a**) Schematic diagram of the nested PCR that was performed using insertion-site-specific and *LUC*-specific primers. (**b**) Five *LUC* genes were chosen from the TRIP-Pool1 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.

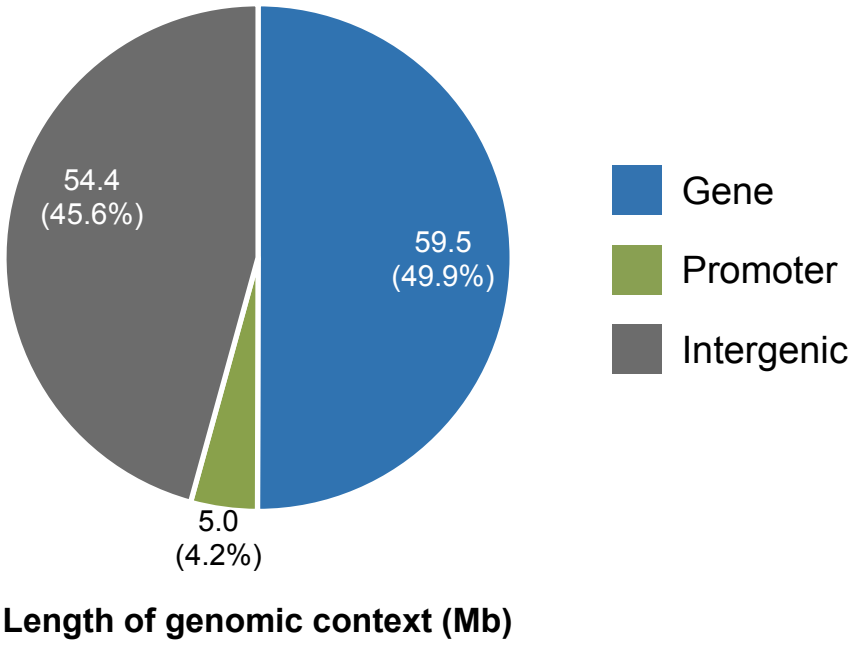


Figure 3.S3. Length of each genomic context.
The total length of the respective genomic contexts and their percentage in the whole genome are shown. The 200 bp segments 5′-proximal to the genic region (CDS plus UTR regions according to TAIR10) were defined as promoter regions, and the remaining sequences 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).

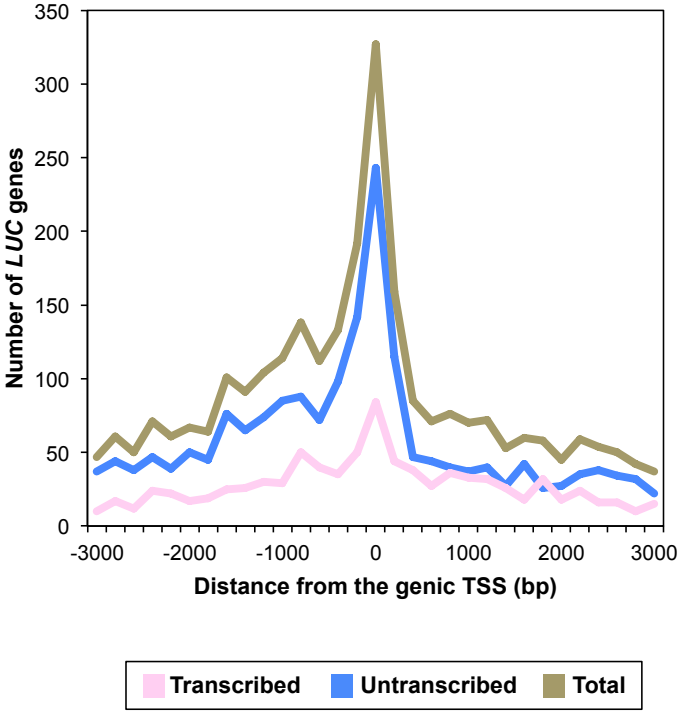


Figure 3.S4. Abundances of *LUC* genes relative to the nearest genic TSS.
Number of LUC genes in relation to the distances from the genic TSS was counted in 200 bp window size.

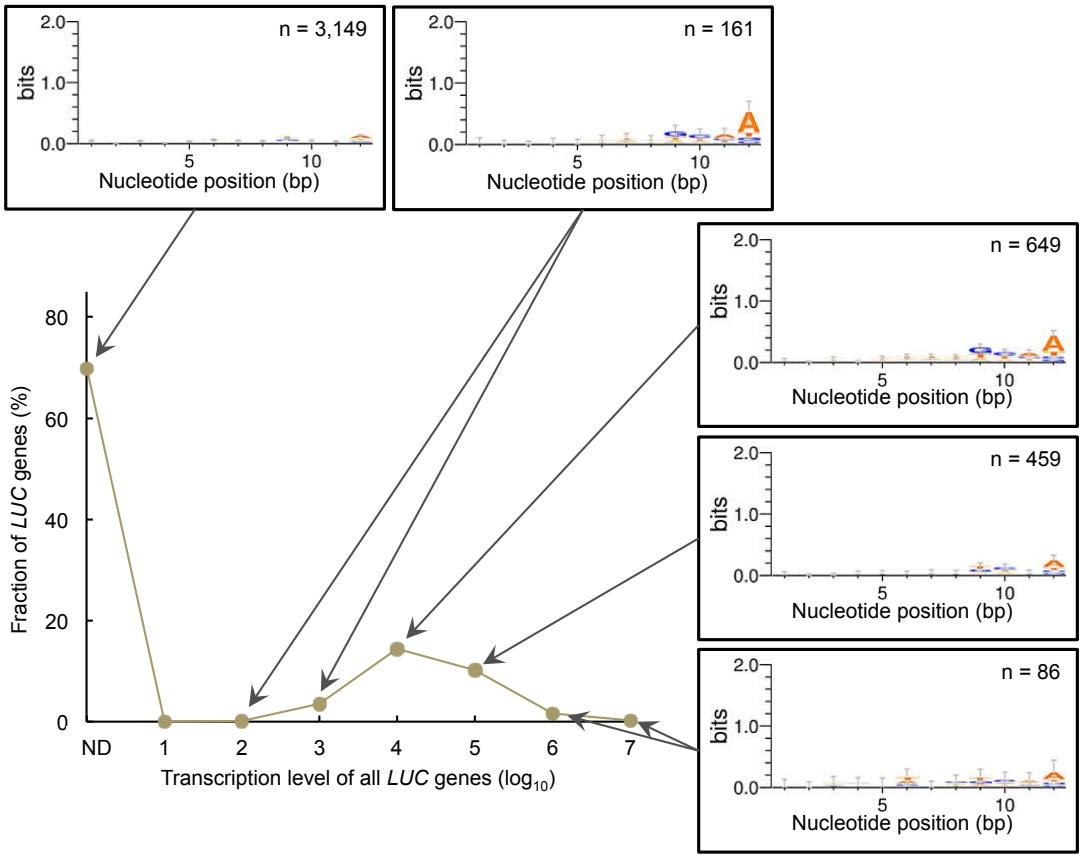


Figure 3.S5. Assessment of the effect of barcode sequences on the LUC transcription levels.
Frequently observed barcode motifs in the LUC insert of indicated transcription levels were analyzed using WebLogo3 (Crooks *et al.*, 2004). The transcription levels of all the LUC genes are shown as in Figure 3.1d. 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.

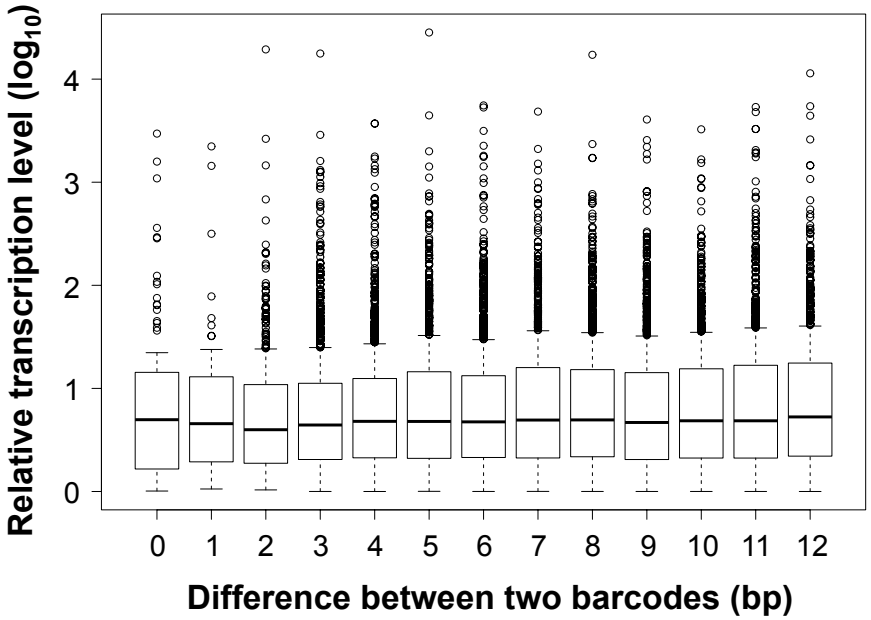


Figure 3.S6. Similarity/dissimilarity of the transcription levels of the randomly selected LUC pairs against the sequence identity of the 12-base barcode.
A pair of LUC genes was randomly selected from the 4,504 mapped LUC genes, and the similarity/dissimilarity of their transcription levels is shown 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.

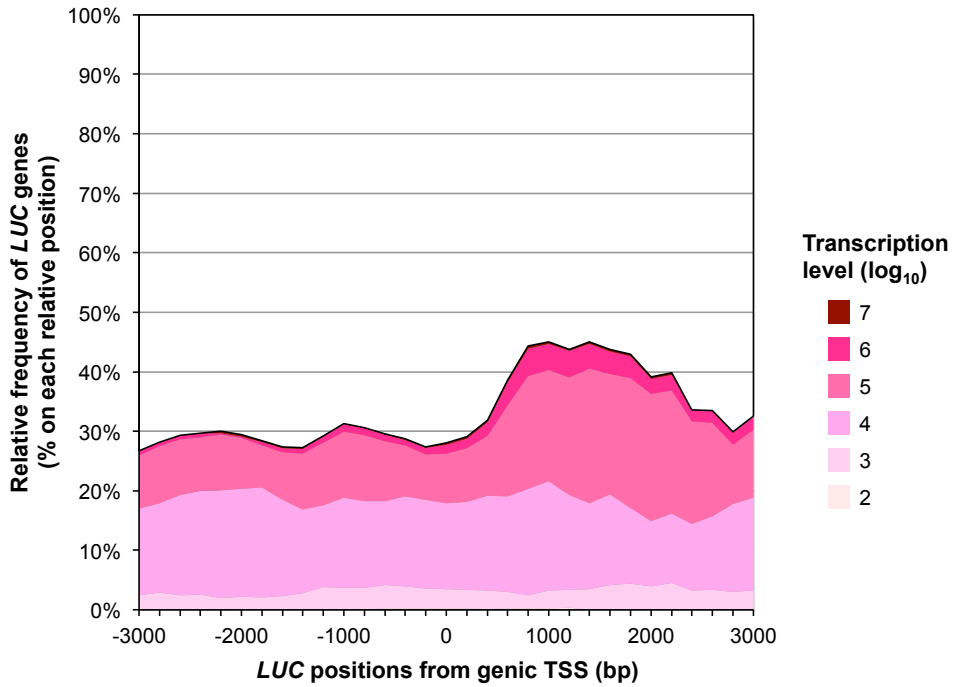


Figure 3.S7. Frequency of transcribed *LUC* genes relative to the annotated genic TSS.
Abundance of the *LUC* genes with the indicated transcription levels in relation to the distance from the genic TSS, as shown in Figure 3.1e. The plot was smoothed by calculating the five-point moving average of integration frequency in each window (200 bp).

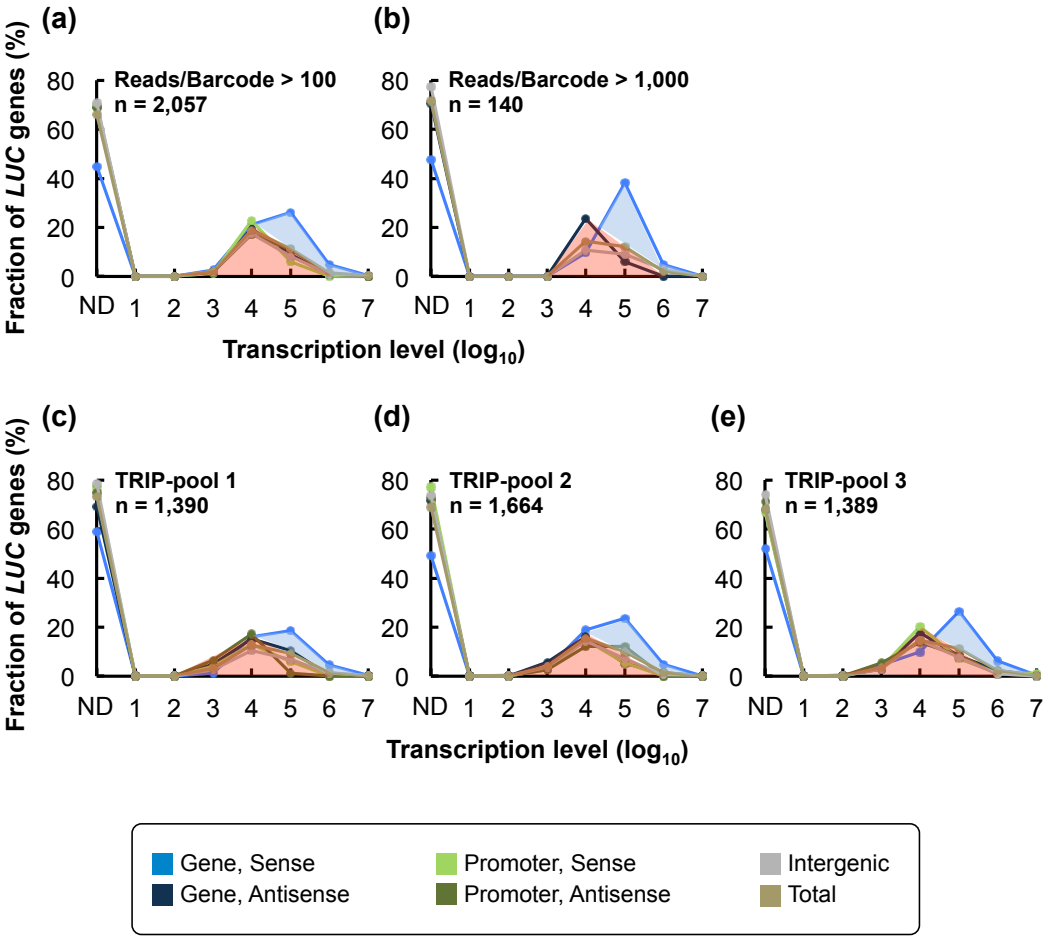


Figure 3.S8. Expression profiles of LUC genes with high-number reads from the amplicon-sequencing data and of LUC genes from biological replicates.
(**a** and **b**) For each barcode, when the number of reads from DNA amplicon sequencing was up to (a) 100 or (b) 1,000, the barcode was omitted from the 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 3.1d. 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. The fraction of the transcribed *LUC* genes attributed by two distinct mechanisms are indicated by light-blue and light-red areas.

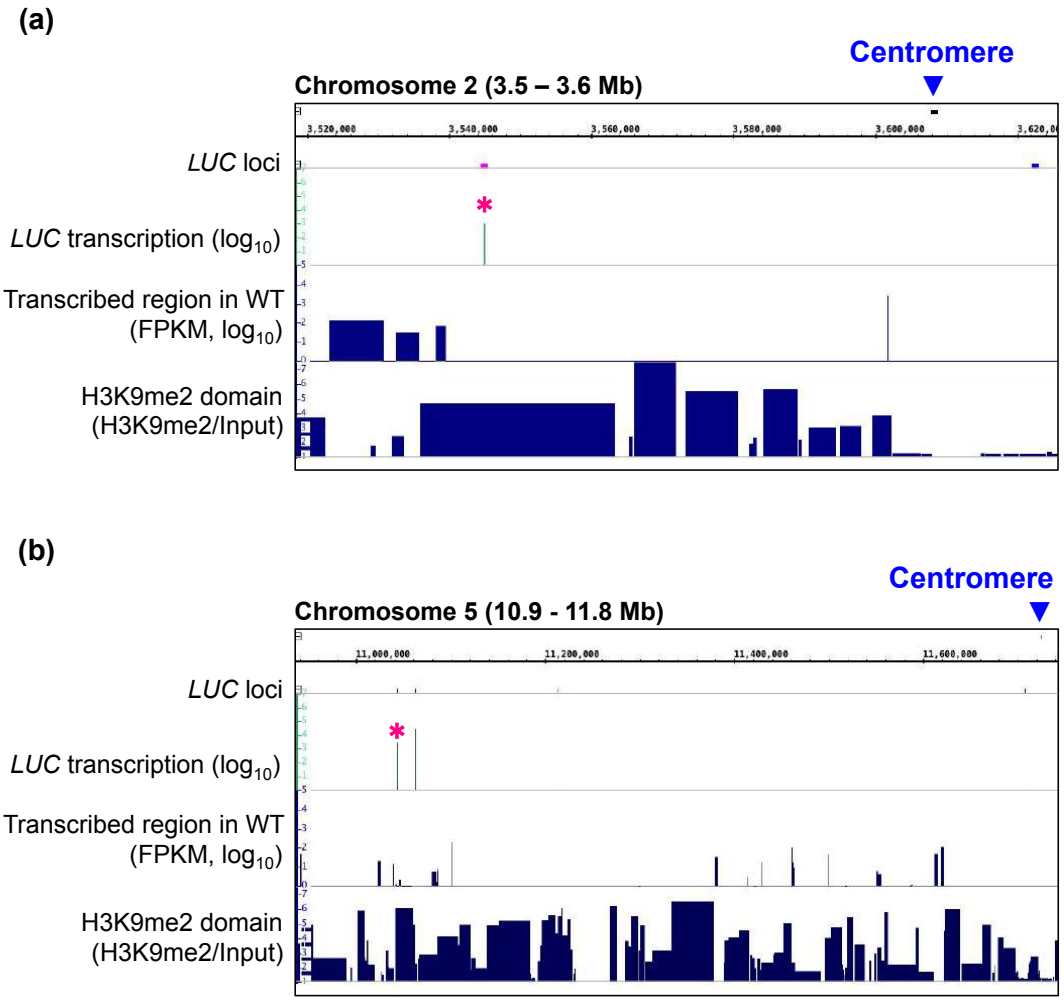


Figure 3.S9. Two examples of transcribed LUC genes in the H3K9me2-marked regions located around the centromere.
(**a** and **b**) Transcribed LUC genes (asterisk) were found 63 kb and 682 kb away from the centromeres of chromosomes 2 (a) and 5 (b), respectively. The respective H3K9me2 levels of these loci were 80 (a) and 91 (b) percentiles, respectively. In WT T87 cells, transcripts were very scarce in these heterochromatic regions.

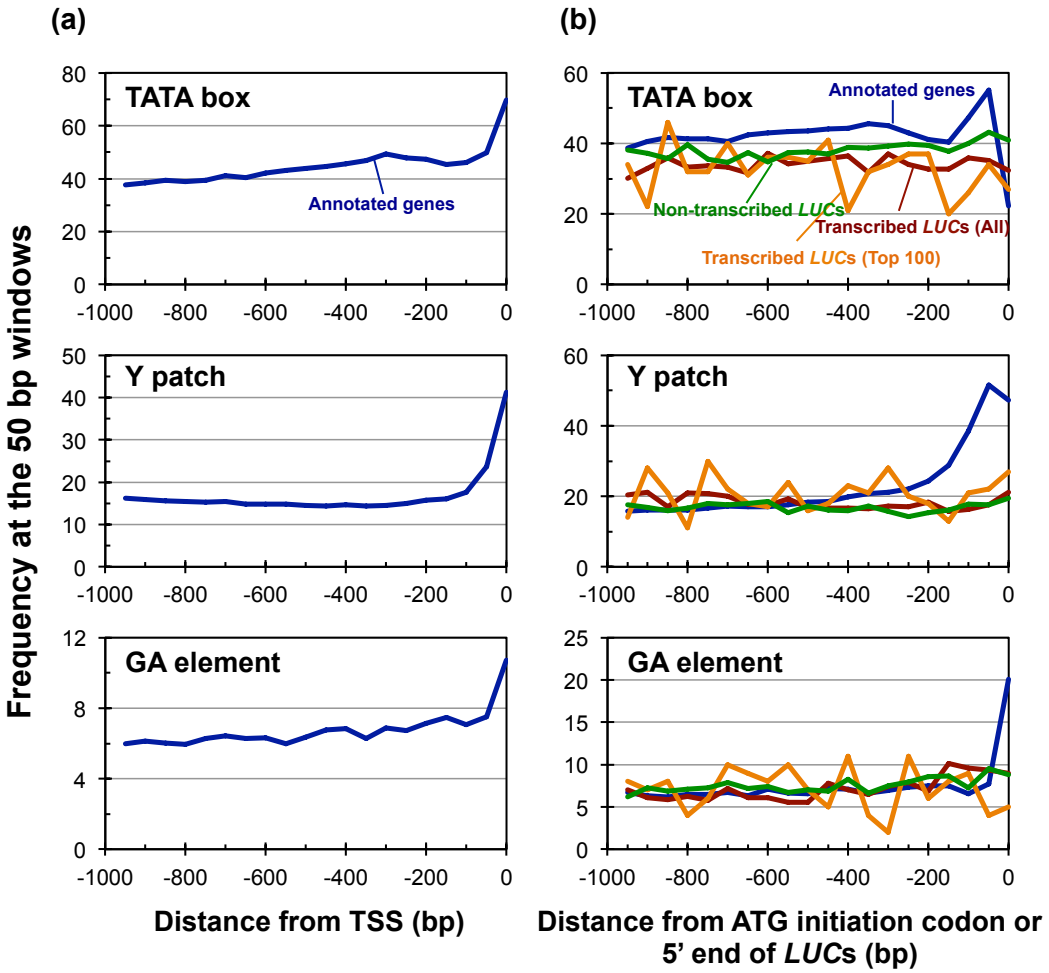


Figure 3.S10. Distribution of *cis*-regulatory elements in the upstream region of *LUC* integration sites.
(**a** and **b**) The frequency of TATA-box, Y patch, and GA elements in the upstream region of the (a) TSS of annotated genes, or (b) of ATG initiation codons of annotated genes and 5′ ends of *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 3.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 3.S2).
TRIP_LUC_iPCR_R2.2	CATTTCGAAGTATTCCGCGTACGTG	Reporter nest primer (see Figure 3.S2).
C1_CGGAAAGACCAA_AS_F1	TCCTCAATGAGTCTGGTGACTTC	Site1 specific outer primer (see Figure 3.S2).
C1_CGGAAAGACCAA_AS_F2	CTCATTGCCCTCAGGTTGGT	Site1 specific nest primer (see Figure 3.S2).
C2_GCACAAAGTCTA_S_F1	TCACTGCTCAATGCGATCTCC	Site2 specific outer primer (see Figure 3.S2).
C2_GCACAAAGTCTA_S_F2	TTAGTGTCGCAACAACGAACCG	Site2 specific nest primer (see Figure 3.S2).
C3_CTAGGGGACTCA_AS_F1	TTCGATCCTTCAAAGCGCATCAC	Site3 specific outer primer (see Figure 3.S2).
C3_CTAGGGGACTCA_AS_F2	CAAGGAGCTTGTCTGGAGAGAG	Site3 specific nest primer (see Figure 3.S2).
N1_TGATGATGTCCA_S_F1	GACTACAAATCATTCATCAACCACG	Site4 specific outer primer (see Figure 3.S2).
N1_TGATGATGTCCA_S_F2	TAGTTGATTCCTCTCGTTCGGC	Site4 specific nest primer (see Figure 3.S2).
T1_TTAGTTGGTCAA_AS_F1	CCAATCTGACACAAAATAGGTCTCT	Site5 specific outer primer (see Figure 3.S2).
T1_TTAGTTGGTCAA_AS_F2	TTAAAGAGGAGTCACGATCATCGGT	Site5 specific nest primer (see Figure 3.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	

Chapter 4:

Kozak sequence acts as a negative regulator for *de novo* transcription initiation of newborn coding sequences in the plant genome

Summary of Chapter 4

The manner in which newborn coding sequences and their transcriptional competency emerge during the process of gene evolution remains unclear. Here, we experimentally simulated eukaryotic gene origination processes by mimicking horizontal gene transfer events in the plant genome. We mapped the precise position of the transcription start sites (TSSs) of hundreds of newly introduced promoterless firefly luciferase (LUC) coding sequences in the genome of *Arabidopsis thaliana* cultured cells. The systematic characterization of the *LUC*-TSSs revealed that 80% of them occurred under the influence of endogenous promoters, while the remainder underwent *de novo* activation in the intergenic regions, starting from pyrimidine-purine dinucleotides. These *de novo* TSSs obeyed unexpected rules; they predominantly occurred ~100 bp upstream of the *LUC* inserts and did not overlap with Kozak-containing putative open reading frames (ORFs). These features were the output of the immediate responses to the sequence insertions, rather than a bias in the screening of the *LUC* gene function. Regarding the wild-type genic TSSs, they appeared to have evolved to lack any ORFs in their vicinities. Therefore, the repulsion by the *de novo* TSSs of Kozak-containing ORFs described above might be the first selection gate for the occurrence and evolution of TSSs in the plant genome. Based on these results, we characterized the *de novo* type of TSS identified in the plant genome and discuss its significance in genome evolution.

Introduction

In Chapter 3, based on the artificial evolutionary approach, I described where, how, how often exogenously introduced coding sequences become transcriptionally active in the plant genome (Chapter 3), and found quite novel transcriptional activation phenomenon: *de novo* transcription. This transcriptional activation phenomenon occurs independently of chromosomal loci (even in the intergenic untranscribed regions), and does stochastically at each integration event (Chapter 3). As the *de novo* transcription did not require fusion with pre-existing genes/transcripts, which would have harmful effects on the host gene network. Therefore, this transcriptional activation 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, in Chapter 4, we determined the precise position of TSS of *de novo* transcription. *De novo* TSS exhibits clear characteristics: they occurred *de novo* about 100 bp upstream of the inserted coding sequences with specific avoidance of pre-existing putative ORFs containing a Kozak motif. We speculated that these features might reflect a first selection gate for the occurrence and evolution of *de novo* TSSs in the genome, regardless of the functionality of the newborn transcripts. Based on these results, we characterized the *de novo* TSSs detected in the plant genome and discuss their significance in genome evolution.

Materials and Methods

Plant material and growth condition

Arabidopsis thaliana T87 cultured cells (Axelos *et al.*, 1992) were maintained in mJPL3 medium (Ogawa *et al.*, 2008) at 22°C with shaking under continuous-light conditions (50–70 μE m^{-2} s^{-1}). One-week-old cultures were harvested using a 10 μm nylon mesh, washed with H_2O twice and subjected to DNA, RNA and chromatin isolation, respectively. We set up two biological replicates for all further experiments, which were processed independently in each experiment.

T87 wild-type (WT) TSS-seq library preparation

All primers used in this study are listed in Table 4.S1. Total RNA was isolated from WT T87 cells using an RNeasy Plant Mini Kit (QIAGEN) followed by DNase I treatment. Next, polyadenylated

RNA (poly (A) RNA) was enriched using a Dynabeads mRNA Purification Kit (Invitrogen) according to the manufacturer's protocols. Poly (A) RNA (2.0 μg) was reverse transcribed using 1000 pmol of random hexamer primers tailed with an Illumina Rd1 adapter. Cap-trapping and subsequent adapter ligation (Illumina Rd2 adapter) steps were performed according to the published methods (Takahashi *et al.*, 2012; Murata *et al.*, 2014). Double-stranded cap-trapped cDNAs were amplified using a Nextera XT index primer (Illumina), then size selected at 200–400 bp using AMPure beads (BeckmanCoulter). Next-generation sequencing (NGS) was performed on an Illumina Mi-Seq platform using a 76 bp paired-end protocol.

T87 WT TSS-seq data processing

Low-quality reads (Q30 <80%) were discarded using FASTX_Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). The first nucleotide of the forward reads was added by the library preparation step, and the second nucleotide was attributed to non-templated addition by reverse transcriptase. Therefore, these two nucleotides were trimmed from both ends and were used for TSS validation after mapping according to Yamamoto *et al.* (Yamamoto *et al.*, 2009). Processed paired reads were mapped to the TAIR10 release of the *A. thaliana* genome assembly (https://www.arabidopsis.org/) using STAR (version: 2.5.4b) (Dobin *et al.*, 2013) with the following parameters: *STAR –outFilterMultimapNmax 1 –alignEndsType EndToEnd –alignIntronMax 6000* (Marquez *et al.,* 2012) *–twopassMode Basic.* Concordantly and uniquely mapped forward reads were extracted according to their SAM Flags (Li *et al.*, 2009); 99 (sense to reference) and 83 (antisense to reference). Precise TSSs were called according to their cap signature (Yamamoto *et al.*, 2009).

T87 WT chromatin immunoprecipitation sequencing (ChIP-seq) library preparation

Chromatin isolation and subsequent ChIP of WT T87 cells were performed according to the published method (Materials and Methods in Chapter 3) (Satoh *et al.*, 2020) with modifications, as follows. Fixed cells (0.2 g) were used for chromatin isolation. ChIP was performed with 10–20 ng of solubilized chromatin, Dynabeads Protein-G magnetic beads (Invitrogen) and antibodies: 2.4 μg of an anti-H2A.Z rabbit polyclonal antibody (Kudo *et al.*, 2020) and 1.0 μg of an anti-H3K36me3 rabbit polyclonal antibody (Abcam: ab9050) were used in this experiment. Successful enrichment of ChIPed DNA was validated by quantitative PCR (qPCR) according to Deal *et al.* (Deal *et al.*, 2007) for H2A.Z, and to Yang *et al.* (Yang, Howard and Dean, 2014) for H3K36me3. ChIP-seq libraries were prepared using a DNA SMART ChIP-seq Kit (Clontech) with

1.0 ng of ChIPed DNA and input DNA (DNA extracted from sheared chromatin), respectively. Libraries were size selected at 200–400 bp using AMPure beads. NGS was performed using a 51 bp single-ended protocol on an Illumina HiSeq 2000 platform.

T87 WT methyl-CpG binding domain protein-enriched genome sequencing (MBD-seq) library preparation

DNA was extracted from WT T87 cells using a DNeasy Plant Mini Kit (QIAGEN). DNA (2.0 μg) was sheared to obtain 50–500 bp fragments (median size, 200 bp) by sonication (TOMY, UD-201), and purified using a QIAquick PCR Purification Kit (QIAGEN). Sheared DNA (500 ng) was used for methylated DNA enrichment, followed by NGS library preparation using an EpiXplore Meth-Seq DNA Enrichment Kit (Clontech). Methylated DNA enrichment was verified by qPCR according to Erdmann *et al.* (Erdmann *et al.*, 2014). Enriched DNA (5.0 ng) was used for NGS library preparation. Libraries were size selected at 200–400 bp using AMPure beads. Sequencing was performed using a 51 bp single-ended protocol on an Illumina HiSeq 2000 platform.

T87 WT ChIP-seq and MBD-seq data processing

ChIP-seq data for H3K9me2 were retrieved from *DDBJ Sequence Read Archive* under accession DRA009315. Low-quality reads (Q20 <80%) were discarded using FASTX_Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). The first three nucleotides added during the library preparation step were trimmed. Processed reads were mapped to the *A. thaliana* genome (TAIR10) using Bowtie2 (version: 2.2.5) (Langmead and Salzberg, 2012) allowing for one mismatch. Uniquely mapped reads were adopted, and duplicated reads were removed using Picard tools (version: 2.16.0) (http://broadinstitute.github.io.picard/).

***LUC*-TSS-seq library preparation**

Transgenic T87 cells harbouring promoterless *LUC* genes were established previously (Chapter 3) (Satoh *et al.*, 2020). For three biological replicates of transformed cells, we prepared two technical replicates, respectively. RNA preparation, Cap-trapping and subsequent adapter ligation were performed as described for the WT TSS-seq library preparation with modifications, as follows (Figure 4.1a). Poly (A) RNA (2.0 μg) was reverse transcribed using a 0.2 μM *LUC*-specific primer tailed with an SgfI site. After Cap-trapping, the adapter oligo containing the SgfI site was ligated to the 3′ end of the cDNA. Subsequently, double-stranded cDNA (1–5 ng)

was completely digested by SgfI. Because SgfI sites appear at an exceptionally low frequency in the *A. thaliana* genome (~2 sites/Mb), we could avoid undesirable digestion at endogenous SgfI sites almost completely. Digested cDNAs were then circularized by T4 DNA ligase, and 0.5–1 ng of circularized cDNA was used for inverse PCR to enrich *LUC* cDNA using a *LUC*-specific primer set. Subsequently, a sequencing library was prepared by two rounds of PCR; the first round was performed to add Illumina adapters, and the second was carried out using Nextera XT index primers. Libraries were sequenced on an Illumina MiSeq platform. Possible biases made during the library preparation and sequencing steps were described in the Methods 4.S1.

LUC-TSS-seq data processing

Forward and reverse reads (TSS side and *LUC* side, respectively) were independently processed before mapping for the sake of removing cloning artefacts, trimming unmappable sequences derived from library design, and determining precise TSSs and their barcode sequences (Methods 4.S1 and Figure 4.S1). Subsequently, processed paired reads were mapped onto the *A. thaliana* genome (TAIR10) using STAR (version: 2.5.4b) (Dobin *et al.*, 2013) with the following parameters: *STAR --outFilterMultimapNmax 1 –alignEndsType EndToEnd –alignIntronMax 6000* (Marquez *et al.*, 2012) *–outFilterMismatchNoverLmax 0.06 twopassMode Basic.* Concordantly and uniquely mapped read pairs were collected according to their SAM Flag pairs (Li *et al.*, 2009); the forward and reverse read sets were 99 and 147, or 83 and 163, respectively. Precise TSSs were called according to their cap signature (Yamamoto *et al.*, 2009). Subsequently, we eliminated *LUC*-TSS artefacts caused by PCR and sequencing errors using the procedures described in Methods 4.S1 and Figure 4.S2.

LUC-TSS classification

The distances between individual *LUC*-TSSs and their nearest WT-TSS in the same strand were calculated using bedtools (version: v2.17.0) (Quinlan and Hall, 2010). Using the distribution curve of *LUC*-TSSs against the distance described above, 1,000 times bootstrap repetition of linear approximation using the “segmented” R package (https://CRAN.R-project.org/package=segmented) revealed the presence of an inflection point at ±15 bp from the nearest WT-TSS. According to the inflection point, *LUC*-TSSs were divided into two groups: within or outside of ±15 bp from the nearest WT-TSS. *LUC*-TSSs were then classified according to the combination of TSS and *LUC* positions while considering their orientations (sense or antisense) relative to the *A. thaliana* genome annotations, as well as the

initiation type of the *LUC*-TSSs grouped as described above. For genome annotation, we used the TAIR10 annotation with the exception of the 5′-untranslated region (5′-UTR); these regions were expanded to 200 bp upstream of the annotated position. The annotated regions, with the exception of protein-coding genes (i.e., transposable elements), were defined as “Others”.

TSS characterization

Nucleotide frequency was calculated in a 5 bp window around ±50 bp of *LUC*-TSSs and WT-TSSs, respectively. The sequence logo was generated by the “RWebLogo” R package (version: 1.0.3) (https://CRAN.R-project.org/package=RWebLogo). A metagene plot of epigenetic status was generated by deeptools (version: 3.2.1) (Ramírez *et al.*, 2014) using TAIR10 annotation and *LUC*-TSS positions, respectively. A motif enrichment analysis was performed using Centrimo with reported motif databases (Bailey and Machanick, 2012; O'Malley *et al.*, 2016). Initiation codon (ATG) frequency was calculated in a 100 bp window around *de novo* TSSs and *LUC*-ORFs. The real lengths of the regions located between individual *de novo* TSSs and *LUC*-ORFs varied according to individual sites. Therefore, their individual lengths were normalized to 100 bp when calculating ATG frequency. The distribution of putative ORFs was analysed around ±0.2 kb of intergenic *de novo* TSSs, 5′-UTR of endogenous genes and randomly extracted intergenic regions, respectively. The 5′-UTR of endogenous protein-coding genes was defined as the region located between the annotated initiation codon and their strongest TSS, as determined by the TSS-seq analysis of WT cells. 5′-UTRs with splice sites were excluded from the analysis. Randomly extracted intergenic regions were prepared via the random extraction of 100 bp fragments from the intergenic region over 10,000 times. The heat map and meta-plot of ORF distribution were generated by deeptools (version: 3.2.1) (Ramírez *et al.*, 2014).

Results

TSS determination for the newly inserted promoterless *LUC* genes

As a model of HGT/EGT events, we previously introduced promoterless luciferase (*LUC*) genes into the genome of *A. thaliana* T87 cells, and established cell pools containing thousands of distinct transgenic cell lines (Chapter 3) (Satoh *et al.,* 2020). Each *LUC* insert was indexed by distinct short random sequences (“barcode”), which enabled us to identify individual transgenic

lines *in silico* without establishing isogenic lines. Notably, the cells experienced only 5–10 vegetative divisions without luciferase-based screening; thus, we assumed that they had retained the characteristic features of newborn genes.

To scrutinize the manner in which newborn promoters occur in the plant genome, we analysed transcription start sites (TSSs) and insertion loci of the promoterless *LUC* genes. For this sake, we modified the conventional TSS determination method (Takahashi *et al.*, 2012; Murata *et al.*, 2014) for compatibility with inverse PCR for the selective analysis of the *LUC* transcripts. As shown in Figure 4.1a, we added the recognition sites of a rare-cutter enzyme at both ends of full-length cDNAs, to circularize them. *LUC* cDNAs were then selectively amplified by inverse PCR and subjected to paired-end deep sequencing. To obtain a precise map of *LUC*-TSSs and their corresponding insertion loci with single-nucleotide resolution, we carefully eliminated sequence artefacts derived from non-specifically amplified endogenous cDNAs and erroneous reads generated during the library preparation and sequencing steps (Figures 4.S1 and 4.S2, and Methods 4.S1).

Figure 4.1b shows an example of the *LUC*-TSSs identified here, indicating that four independent *LUC* genes were inserted into the same gene body (AT1G69530), with their corresponding TSSs overlapping endogenous TSSs (Figure 4.1b). In total, we identified 550 *LUC* inserts and 858 corresponding TSSs across the *A. thaliana* genome (Figure 4.1c). Among the 550 *LUC* inserts, 74% were associated with a single TSS and the remainder were associated with two or more TSSs (Figure 4.1d). The *LUC* inserts were unbiasedly distributed over the *A. thaliana* genome (Chapter 3) (Satoh *et al.*, 2020), whereas the *LUC* loci identified in this TSS analysis were over-represented in the genic regions (Figure 4.1e). This bias might reflect the fact that the inserts in the genic regions have relatively higher transcription levels and that their cDNAs were more easily obtained than were those located in intergenic regions. Nevertheless, we should note that one-fourth of the *LUC* inserts identified here were transcriptionally activated in the intergenic regions (Figure 4.1e) and were treated as candidate *de novo*-activated transcripts.

LUC-TSSs were categorized into two types

To elucidate the mechanism via which promoterless *LUC* genes acquired their transcriptional competency, we next examined if the identified *LUC*-TSSs were associated with inherent TSSs. To prepare reference TSS datasets of WT cells, we performed genome-wide TSS-seq. We

obtained 636,507 loci of highly reliable WT-TSS data, which covered 65.9% (18,064/27,416) of the annotated *A. thaliana* protein-coding genes. Compared with WT-TSSs, 64.6% (554/858) of the *LUC*-TSSs matched WT-TSSs with one-nucleotide resolution (Figure 4.2a). It was plausible to conclude that these *LUC*-TSSs were the result of transcriptional fusions with the endogenous transcripts. However, it was unclear whether the remaining *LUC*-TSSs were all *de novo* activated. To address this question, we tested the distribution of *LUC*-TSSs against the distance from the nearest WT-TSSs. Unexpectedly, the plot showed one clear inflection point at ±15 bp (Figure 4.2b). This result led us to hypothesize that a region of ±15 bp of WT-TSSs was under the influence of endogenous promoter activities. Based on these findings, we classified the *LUC*-TSSs into two categories; those located within ±15 bp of WT-TSSs and those located outside these regions. According to this categorization, out of 858 *LUC*-TSSs, we found that 654 (76%) were transcribed by pre-existing promoter activities, whereas the remainder (204, 24%) were candidate *de novo* TSSs that were unaffected by WT promoters (Figure 4.2c).

Systematic classification of *LUC*-TSSs revealed the transcriptional activation mechanism of newborn genes

To clarify the features of *LUC*-TSSs in greater detail, we further classified them based on the combination of (i) *LUC* loci relative to the WT genes, (ii) TSS loci relative to the WT genes and (iii) types of *LUC*-TSS initiation (Figure 4.2c), to give 72 TSS types (Figure 4.3a). Among these 72 types, we identified 17 types in this study (Figure 4.3b, and Figure 4.S3). This classification revealed that ~80% of the *LUC*-TSSs identified in this study were accounted for by transcriptional activation via the trapping of endogenous genes or transcription units (Figure 4.3b, and Figure 4.S3). We found that transposable elements were also sources of transcriptional activation (Figure 4.S3).

As our interest lay in the mechanism via which new promoters emerge in the plant genome, hereafter we focused on the *de novo*-activated TSSs in the intergenic regions ("Intergenic *de novo*", A-α-2 type in Figure 4.3a). To compare the features of *de novo*-activated TSSs with those of pre-existing ones, we chose two additional types of *LUC*-TSSs: "Endogenous fusion" (C_1-β_1-1 type in Figure 4.3a), in which *LUC* genes were inserted in the pre-existing protein-coding genes and their TSSs overlapped with inherent WT-TSSs; and "Intergenic fusion" (A-α-1 type in Figure 4.3a), in which *LUC* genes were found in the intergenic region, but their TSSs overlapped with endogenous intergenic transcripts. In addition, we selected the "Intragenic

de novo" type (C_1-γ_1-2 type in Figure 4.3a) to examine the differences in *de novo* TSSs between genic and intergenic regions. These four types accounted for 80% of the total *LUC*-TSSs identified here (Figure 4.3a).

Newly activated TSSs have RNA polymerase II initiator and TATA-like motifs

Generally, transcription initiates preferentially at purine nucleotides (A/G) that are preceded by pyrimidine nucleotides (C/T) in the eukaryotic genome (Haberle and Stark, 2018; Andersson and Sandelin, 2020; Yamamoto *et al.*, 2009). We confirmed that the *A. thaliana* protein-coding genes utilized the same initiation dinucleotide motif based on the TSS-seq of WT cells (Figure 4.4a, left and middle panels). We found that *LUC*-TSSs also initiated at a Py-Pu dinucleotide motif, even in the *de novo*-activated cases (Figure 4.4b–e, middle panels). A nucleotide composition analysis revealed the existence of an AT-rich region at ~30 bp upstream of *LUC*-TSSs, which might act as a TATA-box for facilitating PIC recruitment (Figure 4.4a–e, left panels). In addition to the AT-rich region described above, we were unable to find any characteristic motifs associated with the *de novo* TSSs.

Promoter-like epigenetic status is not necessary for *de novo* TSS occurrence

Epigenetic status, including histone modification, histone variants, and DNA methylation, plays an important role in eukaryotic gene expression regulation (Gibney and Nolan, 2010). Therefore, we wondered whether the inherent epigenetic status is responsible for *LUC*-TSS activation. We first prepared a genome-wide map of four epigenetic marks in WT T87 cells, i.e., variant of histone H2A (H2A.Z) and lysine (K) tri-methylation of histone H3 (H3K36me3) as active transcription marks and lysine di-methylation of histone H3 (H3K9me2) and methylated cytosine (mC) as repressive marks, in the *A. thaliana* genome (Lauria and Rossi, 2011). In WT cells, we observed typical distributions of these four epigenetic marks around the TSS of endogenous protein-coding genes; H2A.Z exhibited peaks just downstream of TSSs, and H3K36me3, H3K9me2 and mC were distributed broadly along gene bodies (Figure 4.4a, right panel). The epigenetic landscapes of the "Endogenous fusion" type around its TSSs were similar to those of WT-TSSs (Figure 4.4a and b, right panels), because this type utilized the WT-TSS. In the "Intragenic *de novo*" type, slight enrichments of H2A.Z and H3K36me3 were found around the TSSs (Figure 4.4c, right panel). However, these apparent enrichments were attributed to those located upstream of WT-TSSs, because WT- and *LUC*-TSSs were located in the close proximity of this insertion type (Figure 4.S4). We also found promoter-specific epigenetic patterns in the

“Intergenic fusion” type, indicating that unannotated WT transcription was trapped in this case (Figure 4.4d, right panel). In contrast with these observations, no significant epigenetic patterns were detected around “Intergenic *de novo*” TSS loci (Figure 4.4e, right panel). Therefore, we concluded that a promoter-like epigenetic status was not necessary for the activation of *de novo* TSSs.

De novo TSSs originated ~100 bp upstream of newborn coding sequences

Pervasive and spurious transcription is a characteristic of the eukaryotic genome and is one of the resources used for the transcriptional activities of new genes (Zhang *et al.*, 2019). Our next question pertained to whether the *de novo* TSSs were activated by trapping cryptic transcripts that were not detected in our transcriptomics analysis of WT cells. To address this question, we attempted to determine the genomic distances between *LUC* insertion sites and the corresponding TSSs (TSS-to-*LUC* distances) for each TSS type. If the pre-existing WT-TSSs were utilized for *LUC*-TSSs after the insertion of *LUC* genes, the TSS-to-*LUC* distances should vary according to their insertion sites relative to the WT-TSSs. Expectedly, the TSS-to-*LUC* distances in these cases were broadly distributed (Figure 4.5a). Next, we examined the *de novo* TSSs. Surprisingly, “Intergenic *de novo*” TSSs initiated predominantly in the close vicinity of *LUC* insertion sites (median distance, 108 bp) (Figure 4.5a), with a relatively small coefficient of variation (CV = 0.60) compared with the “Intergenic fusion” type (CV = 1.08). This short and sharp distribution of TSS-to-*LUC* distances in the case of *de novo* TSSs was not explained by the size of the 5′ upstream intergenic regions of the inserts, because their sizes exhibited a large variation (Figure 4.5b, and Figure 4.S5). We confirmed these distribution profiles in three different biological samples (Figure 4.S6). Taken together, the unique features of *LUC*-to-*de novo* TSS distances suggest that they were not caused by the trapping of pre-existing cryptic transcripts at certain genomic loci; rather, the *de novo* TSSs were really caused by the *de novo* insertion of *LUC* coding sequences in their close proximity.

De novo TSSs do not occur in the pre-existing Kozak-containing ORFs

In this study, *LUC* transcripts were translatable because they had a 5′-cap, a coding sequence and a 3′-polyadenylated tail. We wondered whether a relationship existed between this property and the *de novo* transcriptional activation. We observed that the initiation codon (ATG-triplets) frequency was low around *de novo* TSS loci compared with the distal regions (Figure 4.6a, and Figure 4.S7). This characteristic was similar to the 5′-UTR of endogenous genes (Kim *et al.*,

2007), which suggests that the *de novo* TSS regions might serve as the 5′-UTR of *LUC* messages. However, the determined *LUC* inserts did not have a minimum Kozak motif (A/GNNAUGG) (Nakagawa *et al.*, 2008), as purine residue (A/G) was not enriched at the –3 position from the initiation codon of *LUC*-ORF (Figure 4.6b, and Figure 4.S8a). In addition, the pre-existing putative ORFs around *de novo* TSS regions did not contribute to the translatability of the *LUC* messages; such putative ORFs provided an in-frame Kozak-ATG to the downstream *LUC*-ORFs in only 6.9% of cases (9/129) (Figure 4.S8b). These results indicate that our *LUC*-TSS population was not enriched for translatability of the *LUC* messages. This was a reasonable conclusion because transgenic cells had not been screened for luciferase activity. However, we found that Kozak-containing ORFs exhibited an unusual distribution around *de novo* TSSs: these two entities were mutually exclusive (Figure 4.6c and d). As shown in Figure 6c, *de novo* TSSs did not occur within Kozak-containing ORFs (Figure 4.6c, middle panel, and Figure 4.S8c), while ORFs without Kozak sequences were uniformly distributed around *de novo* TSS loci as well as in randomly sampled intergenic regions (Figure 4.6d, left and middle panels). These distribution patterns were commonly observed among three distinct biological replicates (Figure 4.S8d). Interestingly, the repulsion between TSSs and ORFs was more evident in WT genes, with few ORFs found around TSSs and 5′-UTRs regardless of the Kozak motif (Figure 4.6c and d, right panels). Therefore, the anti-Kozak rule of the *de novo* TSSs might be an initial stage of the repulsion between the TSSs and ORFs. These findings imply that the anti-Kozak rule might be an outcome of the immediate responses to sequence insertion, with subsequent natural selection steps eliminating the ATG-triplets interposed in the 5′-UTR through evolutionary timescales.

Discussion

A long-standing question in biology concerns the principles of evolutionary innovation. The origination of new genes is a central driver of evolution and has attracted the interest of researchers. Comparative genomics has been an effective tool in this research area, as it has provided various insights into the gene evolutionary process (Kaessmann, 2010; Cardoso-Moreira and Long, 2012; McLysaght and Guerzoni, 2015; Van Oss and Carvunis, 2019). However, the time resolution of comparative genomics has intrinsic limitations and is not suitable for dissecting the ordered events of the gene origination process in a relatively short

period. In this regard, our artificial evolutionary experiment, which mimicked the HGT/EGT process, has advantages in the study of a much nearer time point to gene birth. By attempting to perform an elaborate classification of the gene insertion types relative to the annotated gene loci (Figure 4.3, and Figure 4.S3), we succeeded in isolating the genuine *de novo*-type transcription of the inserts and in discriminating it from the other types that occurred under the influence of pre-existing promoters.

De novo transcription had the following characteristics: (1) its TSS was located at a Py-Pu dinucleotide located ~100 bp upstream of the *LUC* insert; (2) it tended to have an AT-rich region located ~30 bp upstream of the TSS; (3) inherent promoter-like epigenetic profiles were not needed; and (4) its TSS avoided overlap with pre-existing Kozak-containing ORFs. These analyses were performed using transgenic cells that experienced only 5–10 vegetative cell divisions, and were not screened for luciferase activity (Chapter 3) (Satoh *et al.*, 2020). Therefore, these characteristics were intrinsic properties of noticeably young promoters that were observed right after their birth, before their exposure to evolutionary selective pressures.

Based on the sequence characteristics of *de novo* TSSs mentioned above, as well as the 5′-capped and 3′-polyadenylated nature of the RNA samples (Figure 4.1a), it is probable that the *de novo* transcription that we detected in this study was mediated by RNA polymerase II (pol II) (Haberle and Stark, 2018; Andersson and Sandelin, 2020). An AT-rich region was not always detected upstream of the *de novo* TSS (Figure 4.4); hence, it does not seem to be necessary for *de novo* transcription, but likely facilitates chromatin opening (Zuo and Li, 2011). The relatively low GC content of the *A. thaliana* genome (36%) (Barakat, Matassi and Bernardi, 1998) might increase the occurrence of *de novo* TSSs.

Expression levels of the individual *LUC*-mRNAs could give us further insights into the transcriptional regulation of the respective *LUC* genes. However, the experimental system in this study could not provide reliable data about the expression level of each *LUC*-mRNA due to the experimental limitations (Methods 4.S1). Overcoming this experimental drawback needs further technical improvements.

As *de novo* TSSs occur without inherent promoter-like epigenetic profiles (Figure 4.4e), a transcription-supporting chromatin configuration in these cases is supposed to be formed after sequence insertion. We found analogous cases in transgenic plants, in which promoterless *LUC* genes became transcriptionally activated concomitant with chromatin remodelling around the

LUC insertion loci (Figures 5.4 and 5.5 in Chapter 5) (Hata *et al.*, 2020; Kudo *et al.*, 2020). From the massive analysis of transgenic cultured cells, we also found that transcriptional activation occurred stochastically at 30% of the insertion events across the genome and was independent of chromosomal loci, suggesting that this transcriptional activation reflects the stochastic nature of chromatin remodelling (Figure 3.4 in Chapter 3) (Satoh *et al.*, 2020). Taken together, these findings suggest that gene insertion events stochastically activate local chromatin remodelling to form a transcription-competent chromatin configuration. If this is the case, how is the inserted *LUC* ORF sequence involved in this phenomenon?

De novo TSSs occurred ~100 bp upstream of *LUC* ORFs (Figure 4.5a), suggesting that *LUC* ORFs are involved in the positioning of the PIC. This putative positioning mechanism is buttressed by our previous observation. When core promoter regions were triplicated in front of the *LUC* ORF, the most proximal core promoter unit was predominantly utilized in transgenic plants (Kudo, Matsuo, and Satoh *et al.*, 2020). Therefore, the coding sequence is likely to act as a *cis*-determinant element of the pol II PIC recruitment. The mechanism underlying this PIC positioning warrants further analysis.

Another intriguing finding of this study was the mutual repulsion between the *de novo* TSSs and Kozak-containing ORFs (Figure 4.6c). The simplest explanation for this repulsion is that Kozak-containing ORFs are covered by transcription-repressive chromatin marks, as is known for many annotated genes (Neri *et al.*, 2017; Nielsen *et al.*, 2019). Notably, this repressive effect was not observed for ORFs without a Kozak motif (Figure 4.6d). Considering that the Kozak motif is generally thought to function on mRNA molecules, the repulsion detected here suggests that the epigenetic configuration of the genomic ORF is retro-regulated by the mRNA translatability. Does this feedback mechanism operate within the nucleus, or is it linked to cytoplasmic activities, as are the mRNA surveillance mechanisms (Chang, Imam and Wilkinson, 2007; Smith and Baker, 2015)? This question deserves further investigation.

Based on the collective findings reported above, we propose a model to explain the very initial step of the gene origination process in the plant genome, which is an overlooked time-period under the comparative genomics approach (Figure 4.7). First, when brand-new coding sequences are originated/introduced by genome shuffling or the EGT/HGT process, initial transcriptional activation occurs stochastically anywhere in the genome (Figure 3.1 in Chapter 3 and Figure 4.1c) (Satoh *et al.*, 2020). The newly occurred TSSs avoid pre-existing

Kozak-containing ORFs to avoid interference with the pre-existing genetic information (Figure 4.6c). These processes within a biochemical timescale determine the initial configuration of the pol II promoters, in which the initial recruitment steps of the transcriptional machinery warrant further investigation (Step 2 in Figure 4.7). After the initial activation, *de novo* TSSs are subjected to subsequent natural selection on genetic and evolutionary timescales as observed in the evolutionary trajectory of young genes (Li, Lenhard and Luscombe, 2018; Werner *et al.*, 2018; Zhang *et al.*, 2019; Durand *et al.*, 2019).

In conclusion, our artificial evolutionary experiment allowed the detailed scrutiny of the origination process of functional genes in a biochemical timescale. We describe unique properties of *de novo* TSSs for the first time, which served as the basis of a gene origination model in the plant genome. Because the current study was performed using cultured cells, the genetic behaviour of *de novo* transcription requires further examination regarding heredity (see Chapter 5) and functional adaptation with/without selective pressures.

References of Chapter 4

Andersson, R. and Sandelin, A. (2020) Determinants of enhancer and promoter activities of regulatory elements. *Nat Rev Genet,* 21(2), 71–87.

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.

Bailey, T. L. and Machanick, P. (2012) Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Res,* 40(17), e128.

Barakat, A., Matassi, G. and Bernardi, G. (1998) Distribution of genes in the genome of Arabidopsis thaliana and its implications for the genome organization of plants. *Proc Natl Acad Sci U S A,* 95(17), 10044–10049.

Cardoso-Moreira, M. and Long, M. (2012) The origin and evolution of new genes. *Methods Mol Biol,* 856, 161–186.

Chang, Y. F., Imam, J. S. and Wilkinson, M. F. (2007) The nonsense-mediated decay RNA

surveillance pathway. *Annu Rev Biochem,* 76, 51–74.

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(1), 74–83.

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(1), 15–21.

Durand, É., Gagnon-Arsenault, I., Hallin, J., Hatin, I., Dubé, A. K., Nielly-Thibault, L., Namy, O. and Landry, C. R. (2019) Turnover of ribosome-associated transcripts from de novo ORFs produces gene-like characteristics available for de novo gene emergence in wild yeast populations. *Genome Res,* 29(6), 932–943.

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), 1–8.

Gibney, E. R. and Nolan, C. M. (2010) Epigenetics and gene expression. *Heredity (Edinb),* 105(1), 4–13.

Haberle, V. and Stark, A. (2018) Eukaryotic core promoters and the functional basis of transcription initiation. *Nat Rev Mol Cell Biol,* 19(10), 621–637.

Hata, T., Takada, N., Hayakawa, C., Kazama, M., Uchikoba, T., Tachikawa, M., Matsuo, M., Satoh, S. and Obokata, J. (2020) *De novo* activated transcription of inserted foreign coding sequences is inheritable in the plant genome. *PLOS ONE,* forthcoming, doi.org/10.1371/journal.pone.0252674

Kaessmann, H. (2010) Origins, evolution, and phenotypic impact of new genes. *Genome Res,* 20(10), 1313–1326.

Kim, B. H., Cai, X., Vaughn, J. N. and von Arnim, A. G. (2007) On the functions of the h subunit of eukaryotic initiation factor 3 in late stages of translation initiation. *Genome Biol,* 8(4), R60.

Kudo, H., Matsuo, M., Satoh, S., Hachisu, R., Nakamura, M., Yamamoto, Y., Yoshiharu, Hata, T., Kimura, H., Matsui, M. and Junichi, O. (2020) Cryptic promoter activation occurs by

at least two different mechanisms in the *Arabidopsis* genome. unpublished data, *bioRxiv* [posted 2020 Nov 28]. Available from: https://www.biorxiv.org/content/10.1101/2020.11.28.399337v1 doi: 10.1101/2020.11.28.399337

Langmead, B. and Salzberg, S. L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods,* 9(4), 357–359.

Lauria, M. and Rossi, V. (2011) Epigenetic control of gene regulation in plants. *Biochim Biophys Acta,* 1809(8), 369–378.

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(5), 676–688.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and Subgroup, G. P. D. P. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25(16), 2078–2079.

Marquez, Y., Brown, J. W., Simpson, C., Barta, A. and Kalyna, M. (2012) Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. *Genome Res,* 22(6), 1184–1195.

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(1678), 20140332.

Murata, M., Nishiyori-Sueki, H., Kojima-Ishiyama, M., Carninci, P., Hayashizaki, Y. and Itoh, M. (2014) Detecting expressed genes using CAGE. *Methods Mol Biol,* 1164, 67–85.

Nakagawa, S., Niimura, Y., Gojobori, T., Tanaka, H. and Miura, K. (2008) Diversity of preferred nucleotide sequences around the translation initiation codon in eukaryote genomes. *Nucleic Acids Res,* 36(3), 861–871.

Neri, F., Rapelli, S., Krepelova, A., Incarnato, D., Parlato, C., Basile, G., Maldotti, M., Anselmi, F. and Oliviero, S. (2017) Intragenic DNA methylation prevents spurious transcription initiation. *Nature,* 543(7643), 72–77.

Nielsen, M., Ard, R., Leng, X., Ivanov, M., Kindgren, P., Pelechano, V. and Marquardt, S.

(2019) Transcription-driven chromatin repression of Intragenic transcription start sites. *PLoS Genet,* 15(2), e1007969.

O'Malley, R. C., Huang, S. C., Song, L., Lewsey, M. G., Bartlett, A., Nery, J. R., Galli, M., Gallavotti, A. and Ecker, J. R. (2016) Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. *Cell,* 165(5), 1280–1292.

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(2), 242–250.

Quinlan, A. R. and Hall, I. M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics,* 26(6), 841–842.

Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A. and Manke, T. (2014) deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res,* 42(Web Server issue), W187–191.

Satoh, S., Hata, T., Takada, N., Tachikawa, M., Mitsuhiro, M., Kushnir, S. and Obokata, J. (2020) Plant genome response to incoming coding sequences: stochastic transcriptional activation independent of integration loci. unpublished data, *bioRxiv* 401992 [posted 2020 Nov 28; revised 2021 Feb 4]. Available from: https://www.biorxiv.org/content/10.1101/2020.11.28.401992v2 doi: 10.1101/2020.11.28.401992

Smith, J. E. and Baker, K. E. (2015) Nonsense-mediated RNA decay--a switch and dial for regulating gene expression. *Bioessays,* 37(6), 612–623.

Takahashi, H., Lassmann, T., Murata, M. and Carninci, P. (2012) 5' end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. *Nat Protoc,* 7(3), 542–561.

Van Oss, S. B. and Carvunis, A. R. (2019) De novo gene birth. *PLoS Genet,* 15(5), e1008160.

Werner, M. S., Sieriebriennikov, B., Prabh, N., Loschko, T., Lanz, C. and Sommer, R. J. (2018) Young genes have distinct gene structure, epigenetic profiles, and transcriptional regulation. *Genome Res,* 28(11), 1675–1687.

Yamamoto, Y. Y., Yoshitsugu, T., Sakurai, T., Seki, M., Shinozaki, K. and Obokata, J. (2009) Heterogeneity of Arabidopsis core promoters revealed by high-density TSS analysis. *Plant J,* 60(2), 350–362.

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(15), 1793–1797.

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(4), 679–690.

Zuo, Y. C. and Li, Q. Z. (2011) Identification of TATA and TATA-less promoters in plant genomes by integrating diversity measure, GC-Skew and DNA geometric flexibility. *Genomics,* 97(2), 112–120.

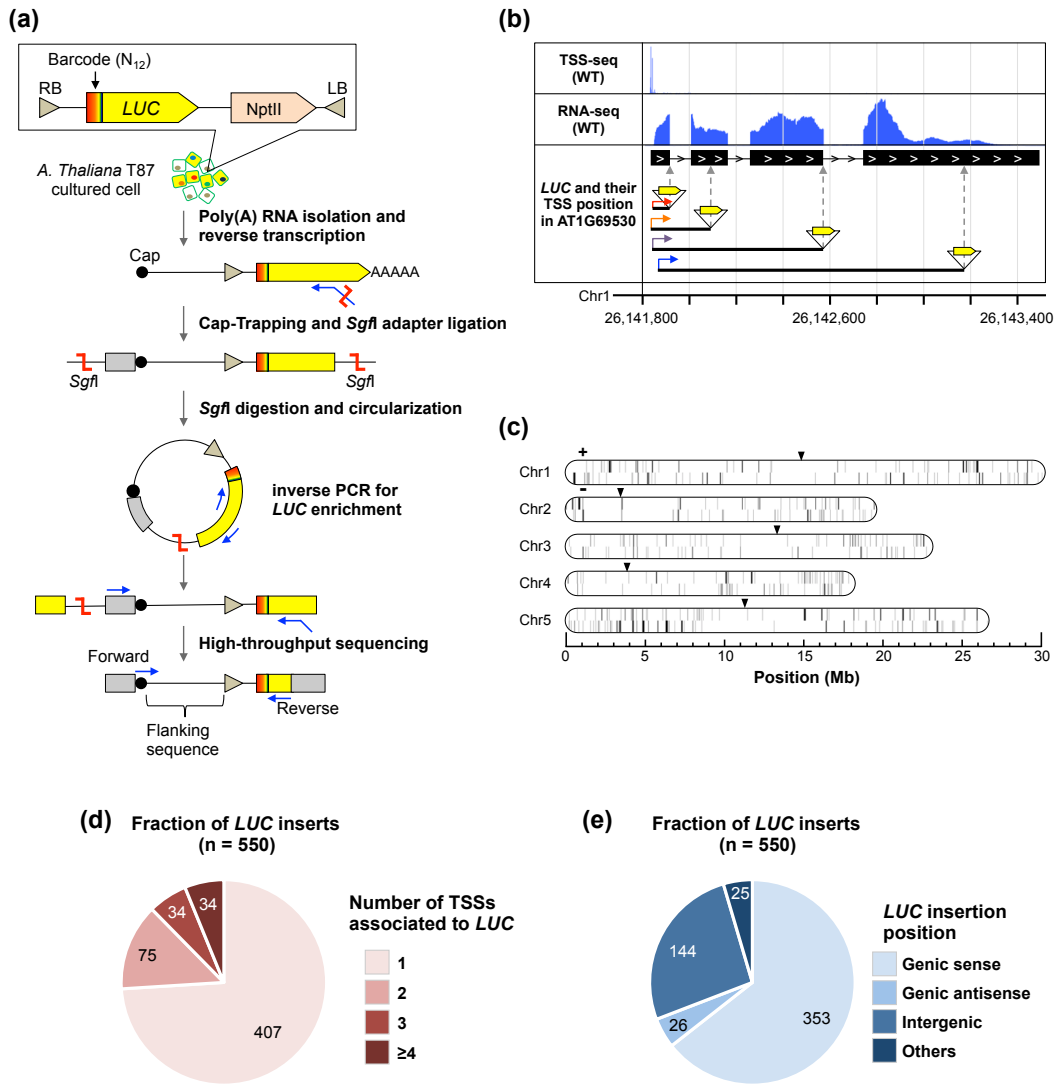


Figure 4.1. **Determination of the TSSs of promoterless *LUC* genes at single-nucleotide resolution.** (**a**) Experimental design of the parallel determination of promoterless *LUC* insertion sites and their corresponding TSSs. cDNAs reaching the 5′ end of *LUC* RNAs were prepared by the Cap-trapper method followed by inverse PCR. Amplified cDNAs were subjected to paired-end sequencing. For details, see the Methods. (**b**) Example of determined *LUC*-TSSs in the genome viewer. The coloured arrows indicate the determined *LUC*-TSSs. (**c**) Chromosomal map of all determined *LUC*-TSSs. The ticks indicate the genomic loci of 858 *LUC*-TSSs with sense (+) and antisense (–) orientations on *Arabidopsis thaliana* chromosomes. The black triangles indicate centromeres. (**d**) Relative abundance of the *LUC* inserts associated with the indicated number of TSSs. (**e**) Relative abundance of the *LUC* inserts with respect to the insertion types. Genic, protein-coding gene; Others, TAIR10-annotated region excluding protein-coding genes; Intergenic, unannotated region in TAIR10.

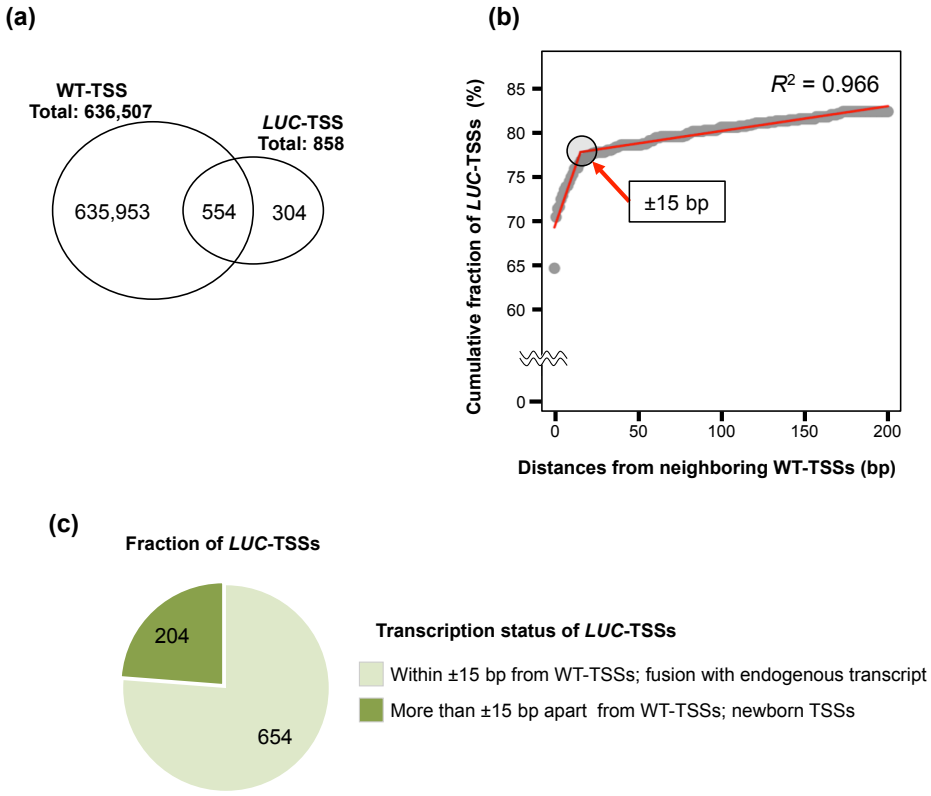


Figure 4.2. Categorization of *LUC*-TSSs with respect to WT-TSSs. (**a**) Venn diagram summarizing the overlap between the positions of WT-TSSs and *LUC*-TSSs at single-nucleotide resolution. (**b**) The grey dots show a cumulative fraction of *LUC*-TSSs according to the distances from their nearest WT-TSSs. The red line indicates the linear approximation of the grey dot plots, and the estimated inflection point is indicated by a black circle. The adjusted R^2 was 0.966. (**c**) Number of *LUC*-TSSs categorized according to (b).

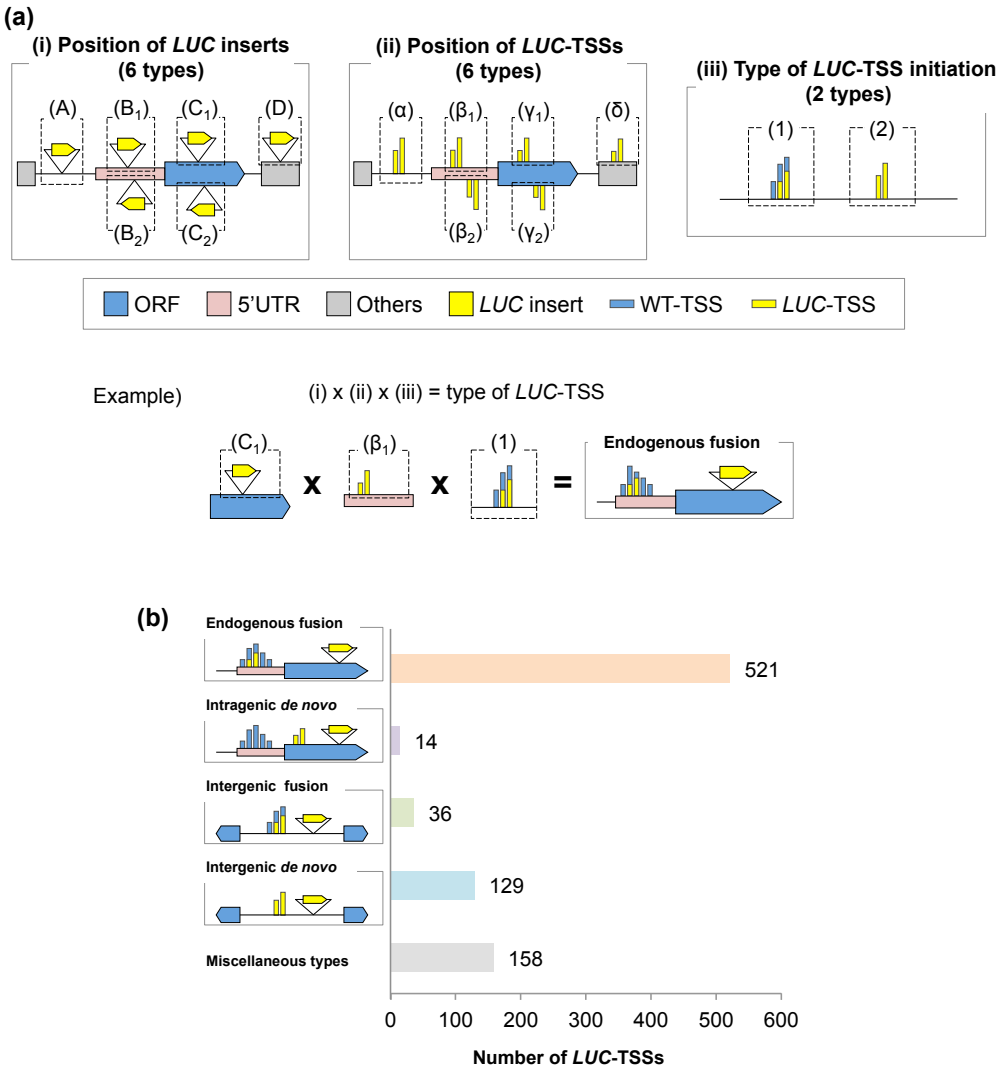


Figure 4.3. Classification of *LUC*-TSSs according to the combination of their genomic loci and types of TSS initiation. (**a**) *LUC*-TSSs were classified according to the combination of the position of (i) the *LUC* insert, (ii) the *LUC*-TSS relative to *Arabidopsis thaliana* annotated genes and (iii) the types of *LUC*-TSS initiation, as categorized in Figure 4.2c. Example showing the classification scheme of the "Endogenous fusion" type, in which the *LUC* gene was inserted in an endogenous ORF and the TSS initiated from the 5′-UTR of the ORF with an overlapping WT-TSS. (**b**) Number of *LUC*-TSSs of the representative insertion types, as described in the text. The contents of miscellaneous types are shown in Figure 4.S3.

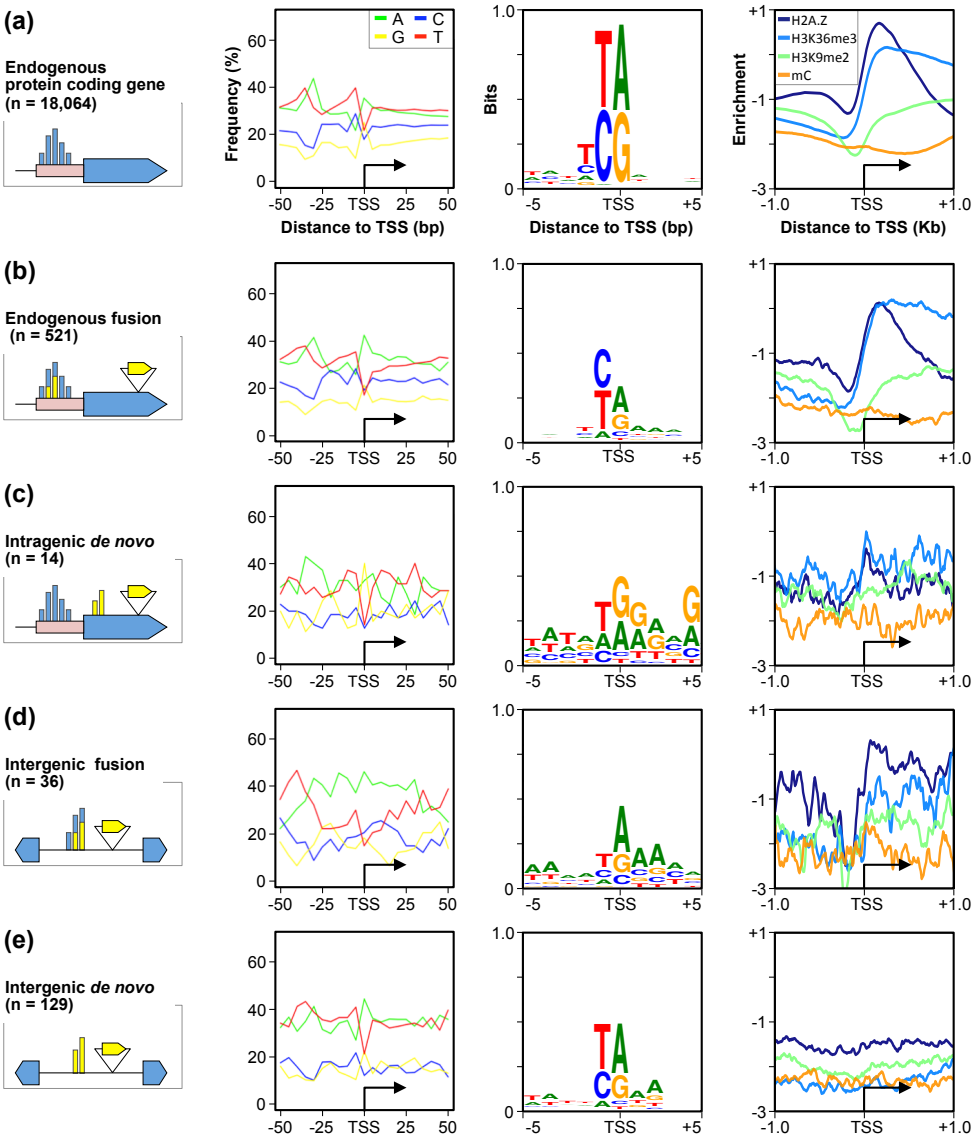


Figure 4.4 Sequence and epigenetic characteristics of the *LUC*-TSSs. (**Left panels**) Nucleotide frequency at 5 nt resolution centred on the TSSs of (**a**) endogenous protein-coding genes (n = 18,064) and *LUC*-TSSs classified as (**b**) "Endogenous fusion" type (n = 521), (**c**) "Intragenic *de novo*" type (n = 14), (**d**) "Intergenic fusion" type (n = 36) and (**e**) "Intergenic *de novo*" type (n = 129). The black arrows indicate the TSS. (**Middle panels**) Sequence logo around ±5 bp of the TSSs of (a) endogenous genes and (b–e) *LUC* genes. (**Right panels**) Distribution profiles of H2A.Z, H3K36me3, H3K9me2 and methylated cytosine (mC) in WT cells, within ±1.0 kb of the TSSs of (a) endogenous genes and (b–e) *LUC* genes.

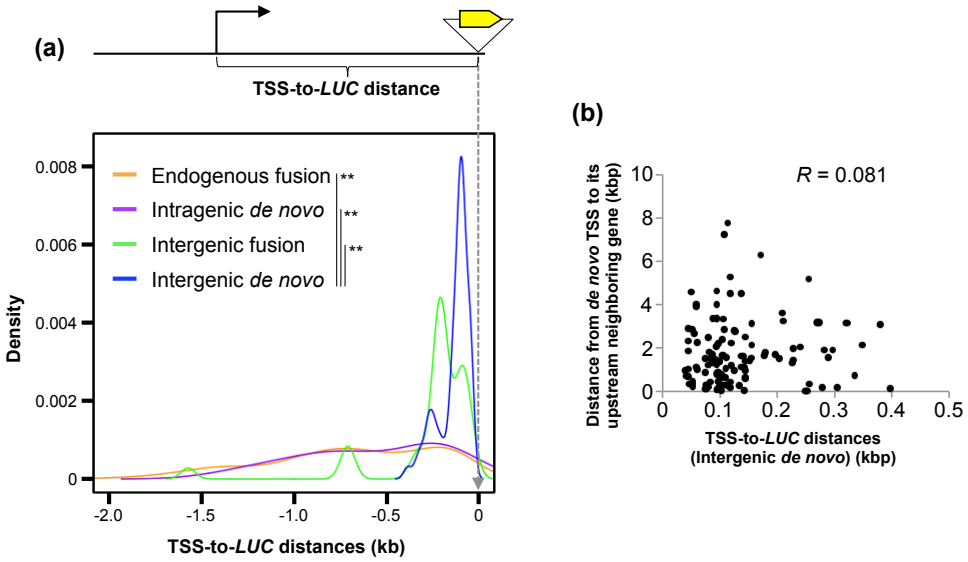


Figure 4.5. *De novo* TSSs occur in the 5′ proximity of the *LUC* inserts. (**a**) Density plot showing the distribution of the distance between the TSS and *LUC* insertion site (TSS-to-*LUC* distance) in each *LUC*-TSS type. The median TSS-to-*LUC* distance was (in bp): Endogenous fusion, 666; Intragenic *de novo*, 573; Intergenic fusion, 212.5; and Intergenic *de novo*, 108. ***P*-value < 0.01 (Wilcoxon rank-sum test). (**b**) Scatter plot showing the correlation between the TSS-to-*LUC* distance ("Intergenic *de novo*" type) and the TSS-to-upstream neighbouring gene distance. *R*, Pearson's product-moment correlation test.

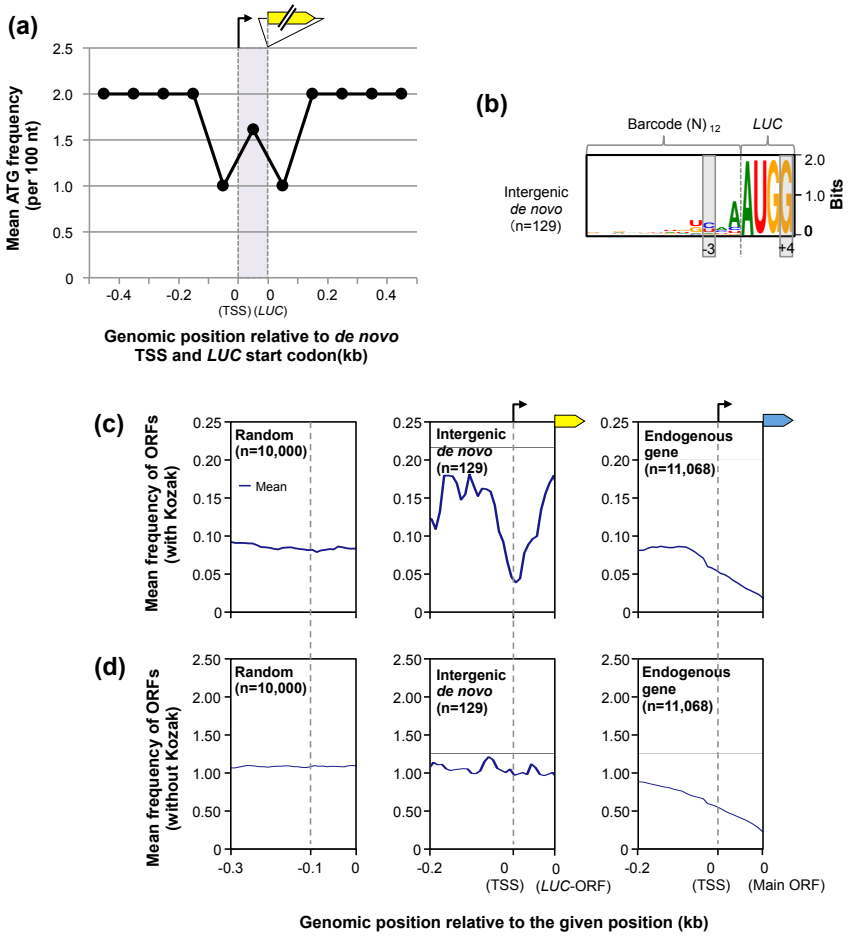


Figure 4.6. *De novo* TSSs avoid pre-existing Kozak-containing ORFs. (**a**) Mean frequency of the initiation codon (ATG) per 100 bp around *de novo* TSS regions. The ATG frequency in the *de novo* TSS regions was normalized per 100 bp. (**b**) Sequence logo of the barcode region on the Intergenic *de novo*-type *LUC* inserts (n = 129). The conserved positions of a minimum Kozak motif (A/GNNAUGG) are indicated by the grey boxes. (**c** and **d**) Meta-plot of the distribution profiles of ORFs (**c**) with or (**d**) without a Kozak motif within 0.3 kb of randomly sampled intergenic regions (left panels), the region from 0.2 kb upstream of the Intergenic *de novo* TSS to its *LUC*-ORF (middle panels) and the region from 0.2 kb upstream of the TSS of endogenous protein-coding genes to their main ORF (right panels). The frequencies of ORFs located within the region from the *de novo* TSS to the *LUC*-ORF and from the genic TSS to the main ORF were normalized per 0.1 kb. *Arabidopsis thaliana* genes with introns in the 5′-UTR were excluded from the analysis. The grey dotted lines indicate the TSS positions.

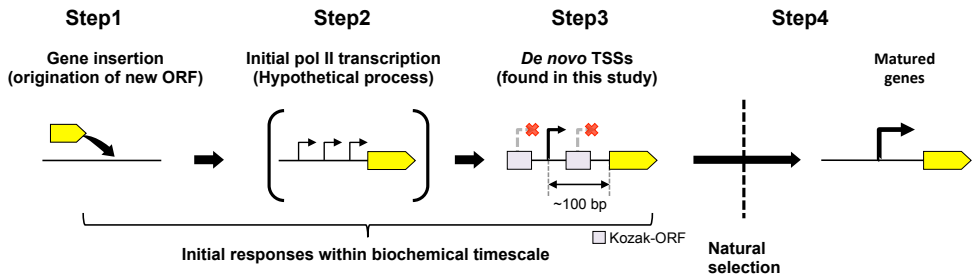


Figure 4.7. Model of the evolutionary processes of new genes. Brand-new coding sequences are originated/introduced by genome shuffling or the EGT/HGT process. *De novo* TSSs occur in response to the origination of a new coding sequence, with satisfying an anti-Kozak rule. *De novo* TSSs are originated within biochemical timescale, independently of the functionality of the messages. After *de novo* TSS occurrence, the neighbouring putative ORFs are eliminated via function-based natural selection in the evolutionary timescale.

Supplemental information of Chapter 4

Methods 4.S1

Forward read (TSS side) processing before mapping (Figure 4.S1)

Read trimming

Sequences were trimmed to 75 nt from the 3' end of individual reads in order to remove low-quality sequences.

Cap identification and trimming

The first two nucleotides were trimmed since they are added in the library preparation step and therefore unmappable to the genome. Note that the second nucleotide was corresponding to the 5' cap position, so their sequence information was used for TSS validation as cap-signature (Yamamoto *et al.*, 2009).

Check for overlapping sequences to reverse reads

If the genomic position of *LUC*-TSS is very close to the *LUC* insert, forward read sequences would overlap their corresponding reverse read. For extracting properly mappable genomic sequences, we detected such genome-*LUC* chimeric junctions in individual forward reads by BLASTn (Camacho *et al.*, 2009) (version: 2.4.0+) in megablast task (Morgulis *et al.*, 2008) using *LUC* insert sequences obtained from the corresponding reverse read as a subject. Aligned sequences in forward reads were trimmed when they fulfill following cases; (1) the alignment started 5' end of subject (from reverse read), (2) the alignment reached 3' end of the query (forward read), (3) the alignment has no gaps, and (4) the alignment allowed up to 3 mismatches. If the above conditions were not satisfied, read trimming was not performed.

Extracting flanking sequences to map

After checking and removing overlapping sequences to reverse reads, flanking genomic sequences were extracted for paired-end mapping.

Reverse read (*LUC* side) processing before mapping (Figure 4.S1)

Read trimming

Sequences were trimmed to 100 nt from the 3' end of the read in order to remove low-quality sequences. If the reverse read was shorter than 100 nt, the last two nucleotides were trimmed since they are derived from sequence library preparation steps.

Identification of genome-*LUC* chimeric junction

In order to identify precise junction point between the genome and *LUC* insert, and also to identify each barcode sequence, each reverse read was aligned with ideal *LUC* insert sequence (5'-TTATGTTTTTGGCGTCTTCCATNNNNNNNNNNNNCTGTAAGCTGATAACGTCGAGGCCT TGA-3'; N corresponds to barcode) as a BLASTn (Morgulis *et al.*, 2008) subject sequence. Aligned sequences in reverse reads were trimmed when they fulfill following cases; (1) the alignment started 5' end of both subject (ideal sequence) and query (reverse read), (2) the alignment length was at least 45 nt, (3) the alignment has no gaps, and (4) the alignment allowed up to 15 mismatches (this means 3 mismatches allowed excluding barcode sequences). If the above conditions were not satisfied, those reads (and also corresponding forward reads) were discarded as contaminated artifacts.

Extracting flanking sequences and barcode sequences

After checking and removing *LUC* insert sequences, flanking sequences were extracted for paired-end mapping. In addition, each barcode sequence was extracted for *LUC* insert validation.

Genuine TSS-to-*LUC* tag calling (Figure 4.S2)

In order to eliminate aberrant *LUC*-TSS candidates caused by the mutations that occurred during PCR and sequencing steps, we made four histograms (see below) of mapping depth. Each mapping depth was calculated according to the read-tags specified by the TSS position, *LUC* insertion position, and individual barcode sequence (hereafter called TSS, *LUC*, Barcode).

Histogram (A): read depth distribution (Figure 4.S2a)

We firstly made a cumulative histogram by read-tags (TSS, *LUC*, and Barcode) for each experimental replicates. The histogram showed that 50 - 60% of tags appeared only one time in the results, which indicated that a significant amount of tags were erroneously generated due to the errors on either variable; TSS, *LUC*, or Barcode.

Histogram (B): for Barcode validation (Figure 4.S2b)

In order to elucidate erroneous barcode sequences from the results, we calculated the read occupancy of each Barcode in the associated locus (TSS and *LUC*). The histogram showed that low frequent Barcode species occupied 50 - 80 % of the reads that were mapped onto the same TSS-to-*LUC* loci.

Histogram (C): for Barcode and TSS validation (Figure 4.S2c)

Each *LUC* insert can have multiple TSSs. Hence, for each *LUC* insert, the genomic position of associated TSSs can be multiple, but their barcode sequence should be the same. In order to validate TSSs for individual *LUC* inserts, we calculated the read occupancy of each Barcode species in the associated *LUC* locus with ignoring their TSS locus. The histogram showed that among the reads that were mapped onto the same *LUC* locus, lower frequent Barcode occupied 60 - 75 % of them.

Histogram (D): for *LUC* validation (Figure 4.S2d)

Erroneous sequencing data can cause the miss-identification of the *LUC*-genome junction during the read processing step, which will shift the mapping result of the *LUC* locus with several nucleotides. Thus, in order to validate *LUC* insertion sites, we calculated the read occupancy of each *LUC* in the associated TSS with Barcode. The histogram showed that 10 - 20% of reads were mapped on different *LUC* locus among respective TSS-barcode sets.

Based on the above histograms, we collected tags when they fulfill following cases; (1) at least two counts (based on the histogram (A)), and (2) reads of which occupancies in histograms (B), (C), and (D) are more than 70%. If any above conditions were not satisfied, those reads were discarded regarded as erroneous ones.

Experimental limitations and possible biases of *LUC*-TSS determination

There are some experimental limitations and possible biases made during the library preparation and sequencing steps in the *LUC*-TSS determination.

(1) Transcriptional strength of each *LUC* mRNA

In the present study, the established transgenic cell pools were highly heterogeneous; they contained thousands of distinct transgenic cell lines, and each cell line consisted of only ~1,000 cells. Moreover, the *LUC* mRNA level of each transgenic line was extremely low compared with the annotated gene transcripts. It was quite difficult to prepare the sequencing library keeping with the initial molecular abundances of such low abundant RNA sample.

The inverse PCR method enables us to enrich *LUC* mRNAs from the heterogeneous sample and moreover to determine TSS and insertion loci of individual *LUC* genes in parallel (Figure 4.1a). The length of *LUC* mRNA varies depending on the location where the *LUC*-TSSs occurred, and sometimes became significantly long (more than 2 kbp), which affected the reaction efficiency of reverse transcription, inverse PCR, and subsequent PCR steps for the library preparation. Moreover, the sequencing library length affects the sequencing efficiency on the illumina MiSeq platform; >1kb library exhibits low sequencing yield because of the limitation of bridge-PCR amplification (personal communication with illumina technical support). Therefore, the sequencing depth of *LUC*-TSSs in this study did not always reflect the initial molecular abundances of *LUC* mRNA, and thus, the present experimental system could not provide reliable data to evaluate the expression level of each *LUC*-mRNA.

(2) Molecular variation of *LUC* mRNA

We found that some sequencing reads contained splice junctions when the *LUC*-mRNAs formed transcriptional fusions with endogenous transcripts (‘Endogenous fusion’ and ‘Intragenic *de novo*’ types). However, in this study, full-length information of *LUC*-mRNAs could not be obtained because of the library preparation method (inverse PCR) and the use of the short-read sequencer.

In addition, the current method also has a limitation in the analysis of *LUC*-mRNAs without flanking genomic sequences. For example, some *LUC*-TSSs could initiate within the *LUC* inserts with either sense or antisense orientations. Such TSSs were mostly overlooked in our

experimental design, because (1) we utilized a sequencing primer that hybridizes to the 5'-end region of *LUC* ORF, and (2) their genomic loci of TSS and *LUC* genes could not be determined uniquely. However, the frequency of such TSSs should be very low: basically, intragenic transcription is epigenetically suppressed (Neri *et al.*, 2017; Nielsen *et al.*, 2019). Indeed, *LUC*-TSSs within the *LUC* inserts were not detected in our experiment. Moreover, *de novo* activated TSSs were less frequent in the intragenic regions than in the upstream of the ORF (Figure 4.S3). Thus, although TSSs could occur within the *LUC* inserts, their frequency is expected to be very low.

(3) Untranscribed populations

The genome-wide characteristics of this artificial evolutionary experiment including untranscribed cell population were described in Chapter 3 (Satoh and Hata *et al.* 2020). The present study utilized an aliquot of the transgenic cell population that was established in this previous study (Satoh and Hata *et al.* 2020). In this study, transcribed *LUC* genes were distributed along the entire *A. thaliana* chromosomes (Figure 4.1c and e). Moreover, "the *LUC* genes detected in the intergenic regions" were proportionally distributed in respect to the length of five *A. thaliana* chromosomes in this study (R = 0.987, Pearson's product-moment correlation test). Thus, the *LUC*-TSSs that we analyzed in this study were thought to be randomly distributed.

References of supporting information

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.

Morgulis, A., Coulouris, G., Raytselis, Y., Madden, T. L., Agarwala, R. and Schäffer, A. A. (2008) Database indexing for production MegaBLAST searches. *Bioinformatics,* 24(16), 1757–1764.

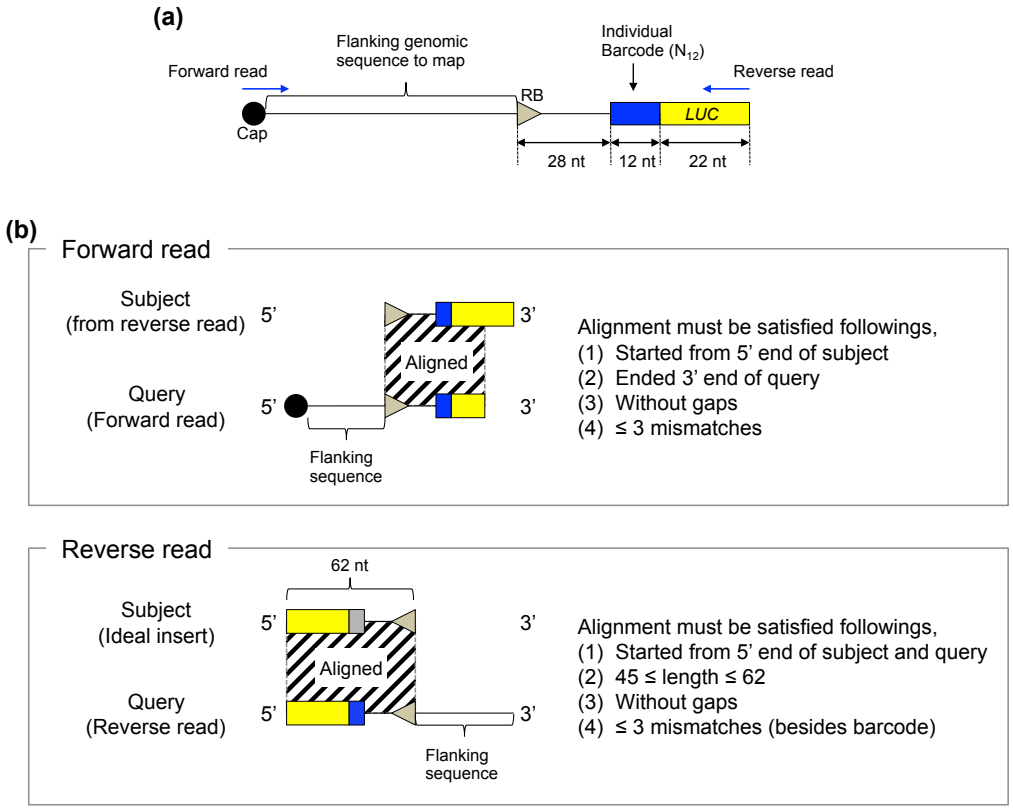


Figure 4.S1. Schematic illustration of paired-end read processing before mapping. (**a**) Sequencing library design. (**b**) Detection scheme of *LUC*-genome junction from forward (Top panel) and reverse (Bottom panel) sequencing reads by Nucleotide BLAST. Shaded areas indicate an ideally aligned region. For detail, see Methods 4.S1.

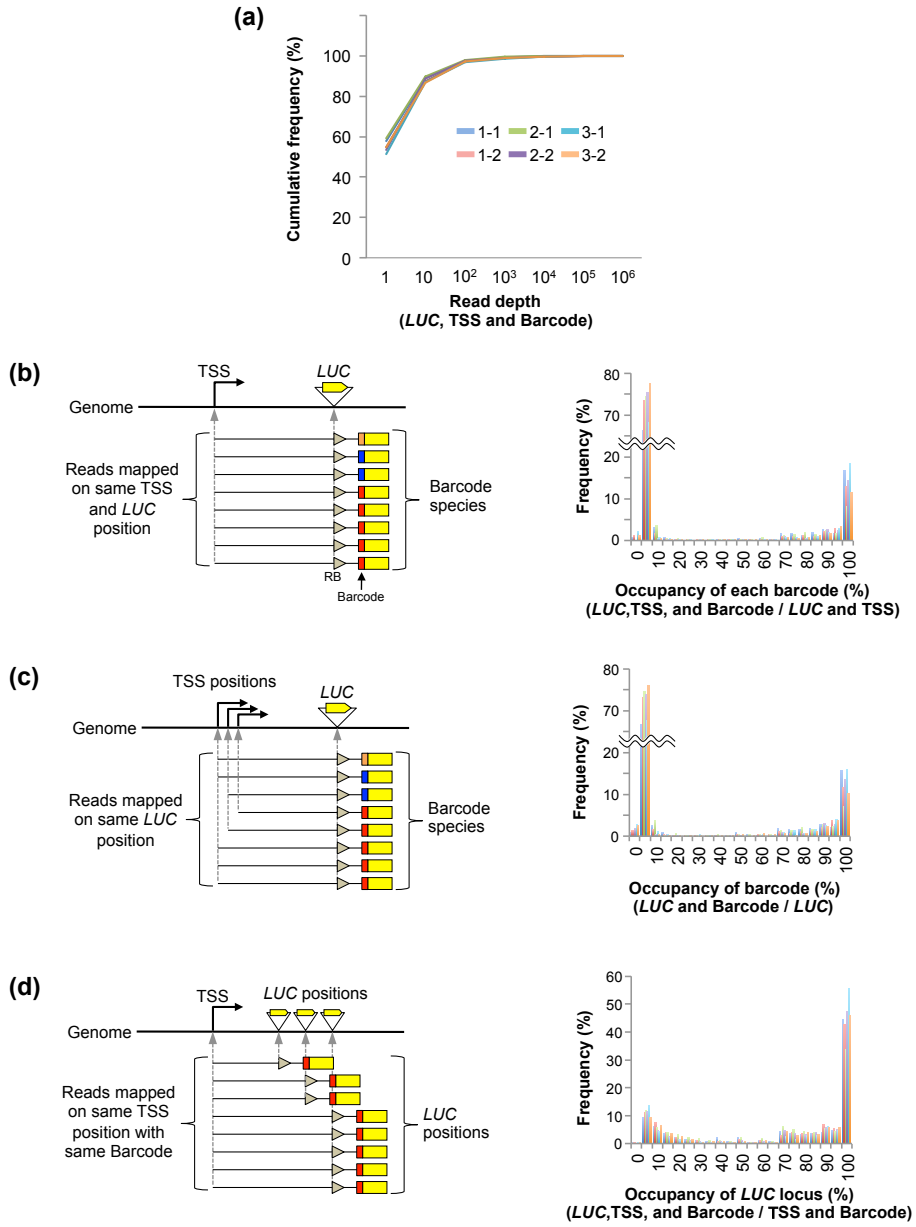


Figure 4.S2. Mapping depth-based determination of genuine *LUC*-TSSs. (**a**) Cumulative frequency of *LUC*-TSS species (specified by TSS, *LUC*, and Barcode) according to their mapping depth in each experimental replicate. (**b**) Barcode validation. (Left) Barcode species and its abundances in the reads that mapped on the same TSS and *LUC* loci were calculated. Different color indicates different barcode sequences. (Right) Barplot shows the frequency of the occupancy of individual Barcode in the reads that mapped on the same TSS and *LUC* loci. (**c**) Barcode and TSS validation. (Left) Barcode species and their abundances in the reads that mapped on the same *LUC* locus were calculated. (Right) Barplot shows the frequency of the occupancy of individual Barcode in the reads that mapped on the same *LUC* loci. (**d**) *LUC* locus validation. (Left) Each frequency of *LUC* loci in the reads that mapped on the same TSS loci with the same Barcode sequences was calculated. (Right) Barplot shows the frequency of the occupancy of the individual *LUC* loci in the reads that mapped on the same TSS loci with the same barcode sequences.

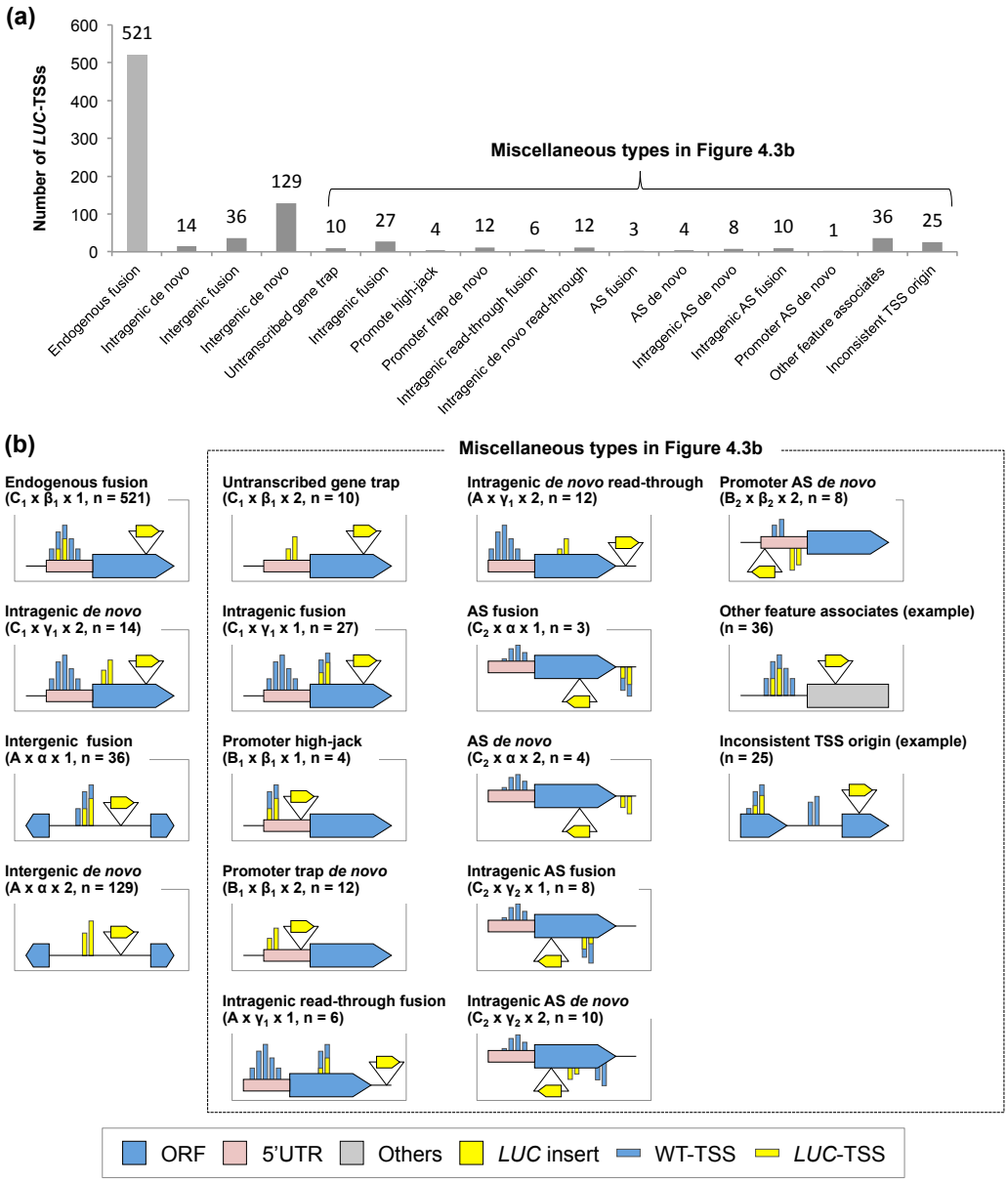


Figure 4.S3. All determined *LUC*-TSS types. (**a**) The number of *LUC*-TSSs classified according to Figure 4.3a. AS: antisense to reference. (**b**) Schematic illustration of each type of *LUC*-TSSs. 'Other feature associates' and 'Inconsistent TSS origin' types represent an example of each type.

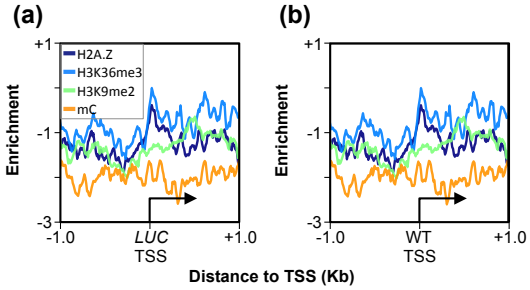


Figure 4.S4. Distribution of epigenetic marks around the TSS of Intragenic *de novo* type. (**a and b**) Distribution profiles of H2A.Z, H3K36me3, H3K9me2, and methylated cytosine (mC) in WT cells, within +/- 1.0 kb of the TSSs of (**a**) Intragenic *de novo* type (from Figure 4.4c), and (**b**) corresponding trapped endogenous genes.

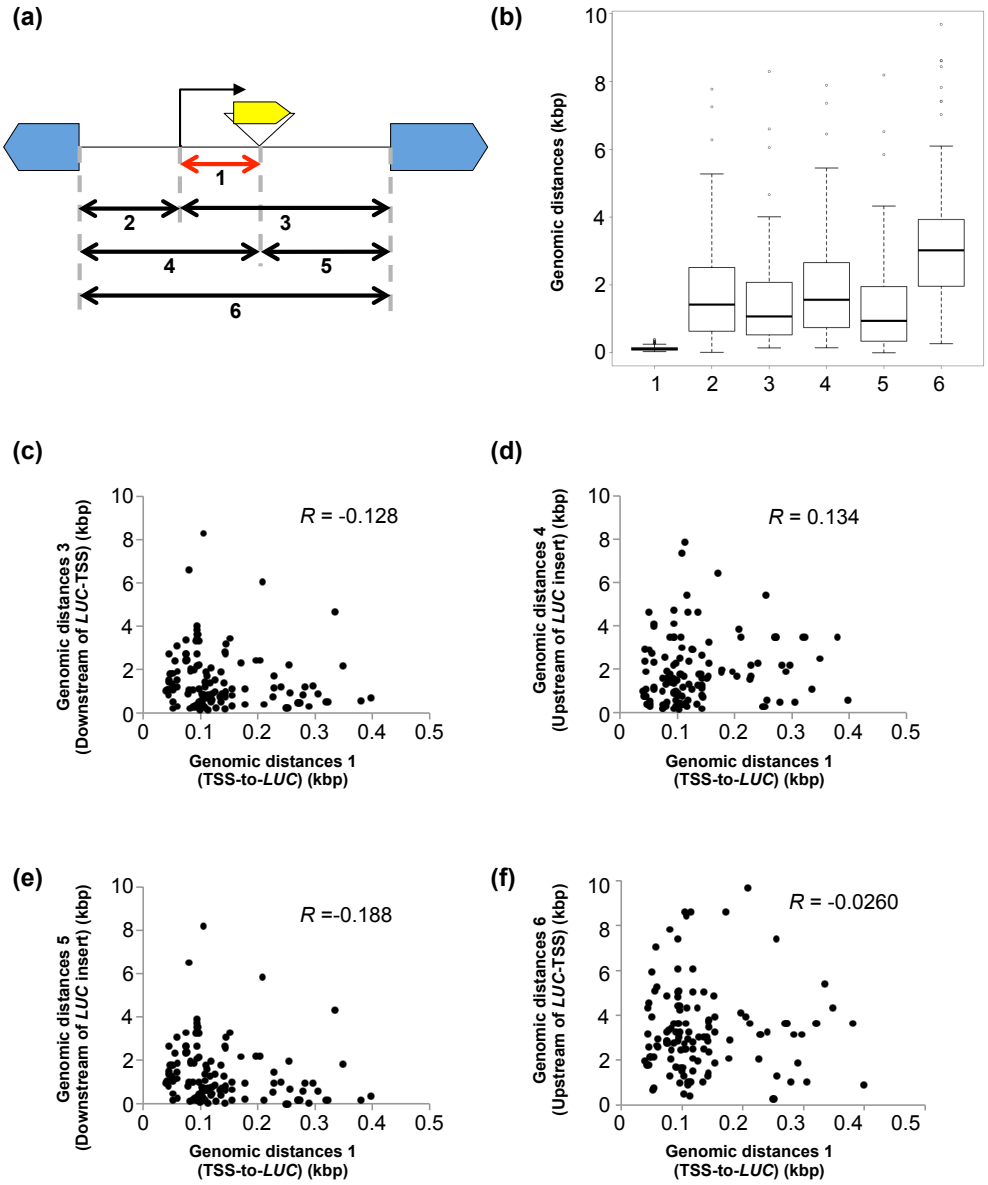


Figure 4.S5. Comparison between the length of *de novo* TSS region and that of the corresponding intergenic region. (**a**) Schematic illustration of distances to be compared; 1: TSS-to-*LUC*, 2: from upstream neighboring gene to TSS, 3: from TSS to downstream neighboring gene, 4: from upstream neighboring gene to *LUC* insertion site, 5: from *LUC* insertion site to downstream genomic feature, and 6: width of the intergenic region. (**b**) Boxplots show the distribution of each distance classified in (a). (**c–f**) Scatterplots show a comparison between genomic distances of TSS-to-*LUC* (Intergenic *de novo* type) and region 3 (c), 4 (d), 5 (e), and 6 (f), of which regions were classified in (a), respectively. *R*, Pearson's product-moment correlation test.

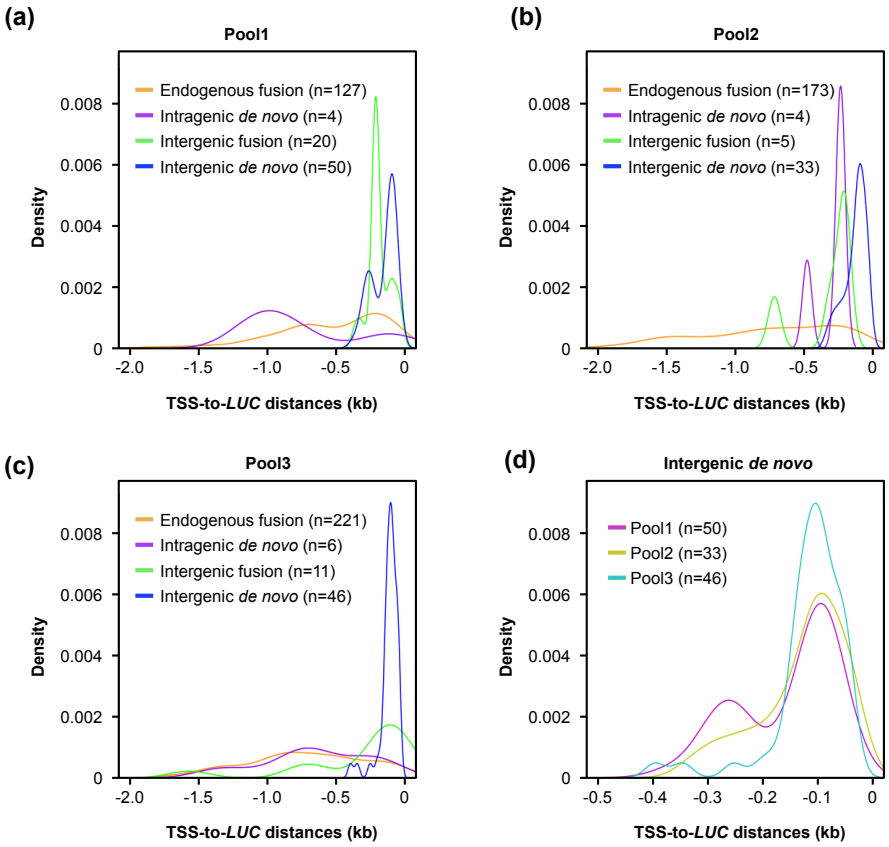


Figure 4.S6. Distribution profiles of TSS-to-*LUC* distances in biological replicates. (**a**–**c**) Density plot shows distribution of distance between TSS and *LUC* insertion site (TSS-to-*LUC* distance) in each *LUC*-TSS type in distinct biological replicates; pool1 (a), pool2 (b), and pool3 (c). (**d**) Distribution profiles of TSS-to-*LUC* distances of intergenic *de novo* types in each biological pool.

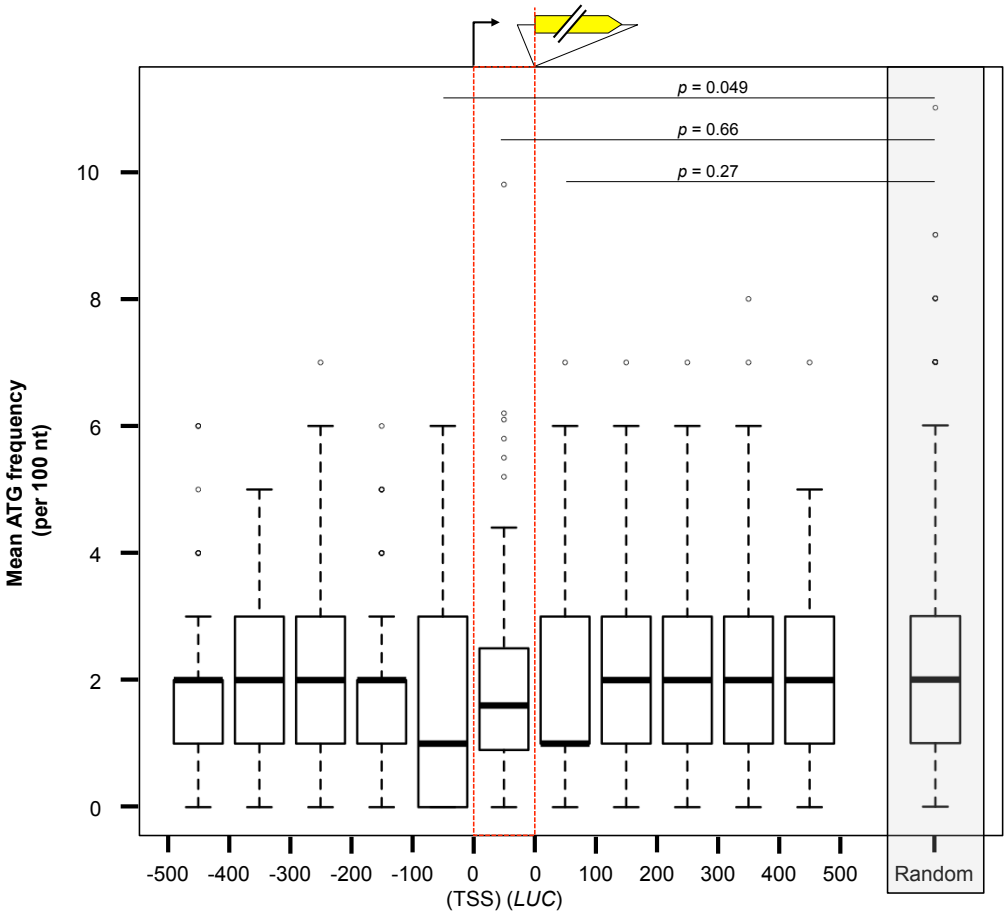


Figure 4.S7. Initiation codon frequency around *de novo* TSS region. Boxplots show the mean frequency of initiation codon (ATG) around *de novo* TSS regions in 100 bp windows. The ATG frequency at *de novo* TSS region was normalized to per 100 bp, and was indicated by red dotted box. Grey shaded box indicates the mean ATG frequency of the 10,000 times bootstrap sampled intergenic region. *P*-values of the Wilcoxon rank sum test are indicated on the plot.

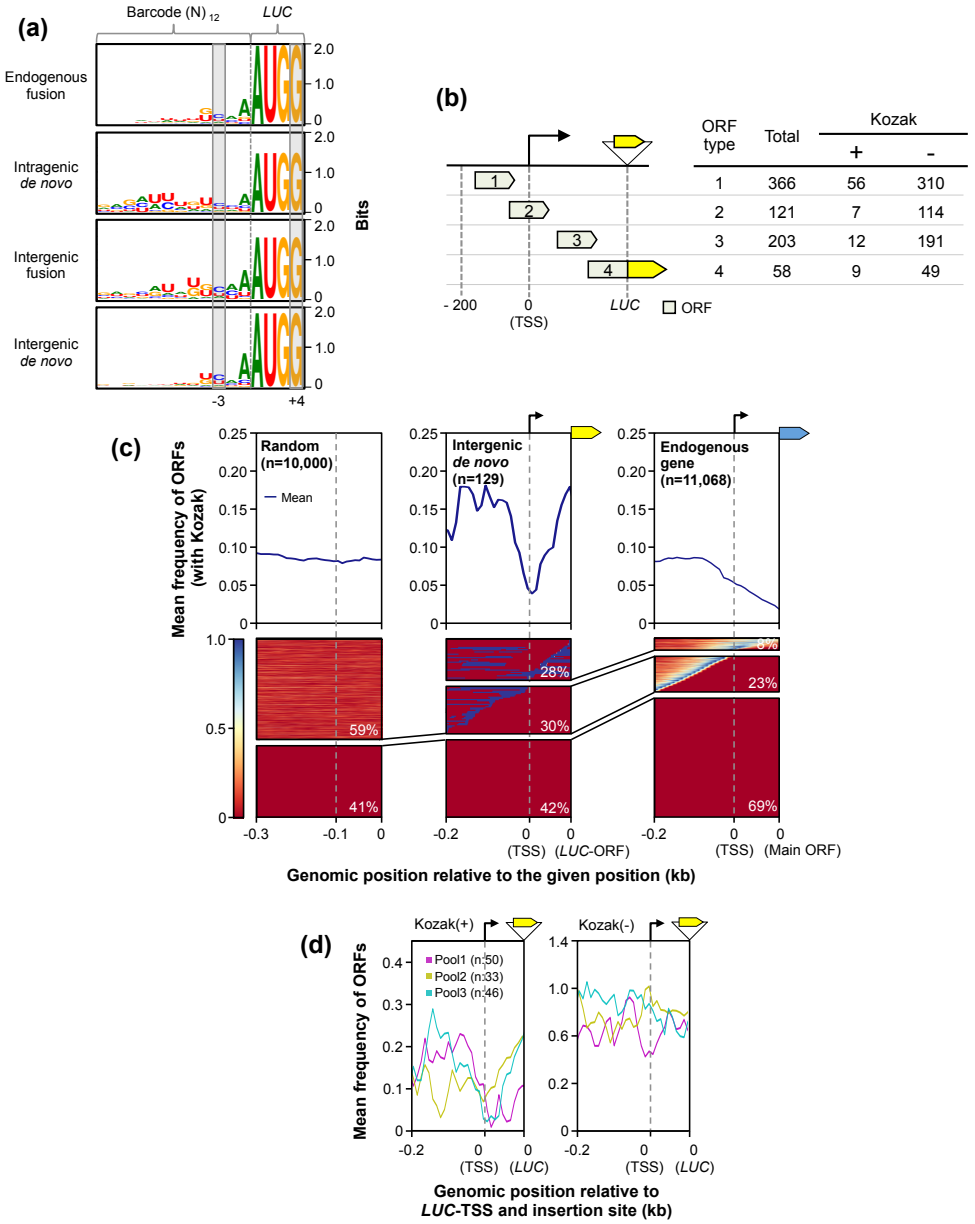


Figure 4.S8. Distribution profiles of Kozak sequence around *LUC* insertion loci. (**a**) Sequence logos of barcode sequences of *LUC* inserts classified according to Figure 3a. Positions of Kozak motifs are indicated by grey boxes. (**b**) Putative ORFs were classified into four types; (1) – 200 bp upstream from *de novo* TSS, (2) over the de novo TSS, (3) between *de novo* TSS to *LUC*-ORF, and (4) fusion with *LUC* ORF. The right table shows the number of putative ORFs found. (**c**) Metaplot and heatmap of distribution profiles of with Kozak-motif within 0.3 kbp of randomly sampled intergenic regions (left panels), the region from 0.2 kbp upstream of the Intergenic *de novo* TSS to its *LUC*-ORF (middle panels), and the region from 0.2 kbp upstream of the TSS of endogenous protein-coding gene to its main ORF (right panels). As for the regions between individual *de novo* TSS and *LUC*-ORF, their real lengths were varied according to individual sites. Therefore, their individual lengths were normalized to 100 bp when calculating ATG frequency. *Arabidopsis* genes with introns in the 5'-UTR were excluded from the analysis. Gray dotted lines indicate TSS positions. (**d**) Metaplot of distribution profiles of pre-existing ORFs with (left panel) or without (right panel) Kozak sequences. *De novo* TSSs in the different biological pools were analyzed separately as in (c).

Table 4.S1. Information of All DNA oligos used in this study.

WT TSS-seq library preparation

Name	Sequence (5' → 3')	Description
Random hexamer tailed with illumina adapter	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNNNNNC	Reverse transcription primer. Tailed illumina adapter (underlined) corresponds to read1 sequencing primer binding sites (Rd1)
linker GN5	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGNNNNN Phos	3' end was phosphorylated. Double stranded linker is prepared by aneal ing with Down oligo. Illumina adapter corresponds to read2 sequencing primer (Rd2) is underlined.
linker N6	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGANNNNNN Phos	3' end was phosphorylated. Double stranded linker is prepared by aneal ing with Down oligo. Illumina adapter corresponds to read2 sequencing primer (Rd2) is underlined.
linker Down	Phos TCTGTCTCTTATACACATCTGACGCTGCCGACGA NH_2 (C6)	5' and 3' end was phosphorylated and aminated. Double stranded linker is prepared by aneal ing with GN5 or N6. Illumina adapter corresponds to read2 sequencing primer (Rd2) is underlined.
2nd strand primer	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG	A primer for synthesis of 2nd strand of cDNA, which aneals with Down oligo of double stranded linker. Illumina adapter corresponds to read2 sequencing primer (Rd2) is underlined.

H2A.Z ChIP qPCR validation (Plant Cell, 2007. 19(1): p. 74-83.)

Name	Sequence (5' → 3')	Description
FLC-480F	TGTAGAGTGGAGGTTCTTTCTG	Primer set for negative control site for H2A.Z enrichment
FLC-480R	TTTTGGGGGTAAACGAGAGT	
FLC+49F	CGACAAGTCACCTTCTCCAA	Primer set for positive control site for H2A.Z enrichment
FLC+49R	CAGAAGATAAAAGGGGGAACA	
FLC+558F	TTGAGTGAAGTTTCAAGCCATC	Primer set for negative control site for H2A.Z enrichment
FLC+558R	TCAGCATTAACCCCATAGCA	
FLC+2353F	TGGAGTTTGGCTTCCTCATAC	Primer set for negative control site for H2A.Z enrichment
FLC+2353R	CGTGTCAAAATTGGTAACATCA	

H3K36me3 ChIP qPCR validation (Curr Biol, 2014. 24(15): p. 1793-7.)

Name	Sequence (5' → 3')	Description
ACTIN_-871_F	CGTAGTTGATATGTATCTTGCTCC	Primer set for negative control site for H3K36me3 enrichment
ACTIN_-773_R	GATTGATCGGTTTTCGTGATATATC	
ACTIN_31_F	GAGCTATATATTCGCACATGTACTCG	Primer set for positive control site for H3K36me3 enrichment
ACTIN_124_R	GATACAGAAGATTCGAGAAGCAGC	
ACTIN_446_F	GTTCGAAATGACTTCGTGTATG	Primer set for positive control site for H3K36me3 enrichment
ACTIN_570_R	CGCTTCATATGTTGATTATTGAG	
ACTIN_2317_F	CTTTAGGATGCTTGTGATGATG	Primer set for negative control site for H3K36me3 enrichment
ACTIN_2463_R	CACCCGATACTTAAATAATTGTCTC	

Methylated DNA fraction enrichment qPCR validation for MBD-seq (G3 (Bethesda), 2014. 5(1): p. 1-8.)

Name	Sequence (5' → 3')	Description
at1g13410_F	AGGTGGACATTGGCGAAGTTGC	Primer set for positive control site for mC enrichment
at1g13410_R	AGCCGGGTTTCTTGGTTCAAGC	
at1g22500_F	ATTGATGCCTGGCTCCGTTCTC	Primer set for negative control site for mC enrichment
at1g22500_R	ACCCGGTACAGGAACGAGATTG	

LUC TSS-seq library preparation

Name	Sequence (5' → 3')	Description
LUC specific RT primer	AAAgcgatcgCGTTTTCACTGCATACGACGATTCTG	LUC specific reverse transcription primer. SgfI site is lowercased.
SgfI linker GN5	AAAgcgatcgTCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGNNNNN Phos	3' end was phosphorylated. Double stranded linker is prepared by aneal ing with Down oligo. Illumina adapter corresponds to read1 sequencing primer (Rd1) was underlined. SgfI site is lowercased.
SgfI linker N6	AAAgcgatcgTCGTCGGCAGCGTCAGATGTGTATAAGAGACAGANNNNNN Phos	3' end was phosphorylated. Double stranded linker is prepared by aneal ing with Down oligo. Illumina adapter corresponds to read1 sequencing primer (Rd1) was underlined. SgfI site is lowercased.
SgfI linker down	Phos TCTGTCTCTTATACACATCTGACGCTGCCGACGAgcgatcgcTTT NH_2 (C6)	5' and 3' end was phosphorylated and aminated. Double stranded linker is prepared by aneal ing with GN5 or N6 oligo. Illumina adapter corresponds to read1 sequencing primer (Rd1) is underlined. SgfI site is lowercased.
2nd strand primer	AAAgcgatcgcTCGTCGGCAGCGTCAGATGTGTATAAGAGACAG	A primer for synthesis of 2nd strand of cDNA, which aneals with Down oligo of double stranded linker. Illumina adapter corresponds to read1 sequencing primer (Rd1) was underlined. SgfI site is lowercased.
LUC_inverse PCR_F	TCTCCAGCGGTTCCATCCTC	Primer set for inverse PCR for LUC enrichment
LUC_inverse PCR_R	CGTTCGGTTGGCAGAAGCTA	
LUC tailed-illumina_adapter_F	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTTATGTTTTTGGCGTCTTCC	Primer set for tailed PCR for library preparation. Illumina adapter corresponds to sequencing primer (Rd1 and Rd2) is underlined.
illumina_adapter_R	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG	

Chapter 5:

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

Summary of Chapter 5

The manner in which inserted foreign coding sequences 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 chromatin 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. A comparison of the results of the T2-plant and cultured cell experiments revealed that the *de novo*-activated transcription concomitant with 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 by which inserted foreign coding sequences become transcriptionally activated and fixed in the plant genome.

Introduction

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, i.e., *de novo* transcription (Chapters 3 and 4) (Satoh *et al.*, 2020; Hata *et al.*, 2021). Could this *de novo* transcriptional activation be a causative mechanism by which newly originated coding sequences acquire their transcriptional competency in the plant genome evolution? Testing this possibility requires the assessment of the genetic behaviour of the *de novo* transcription over generations. The cultured cell-based experiment is not suitable for such scope because the cultured cells continue only the vegetative growth. In this respect, artificial evolutionary experiments with whole plants could provide clues to the above question. The plant body develops with the continuous formation of various tissues and organs from stem cells. Heterogeneity of the transcriptome and epigenome among these different tissues and developmental stages are well characterized in *A. thaliana* plants (Slane *et al.*, 2014; Palovaara *et al.*, 2017; Shulse *et al.*, 2019; Shi *et al.*, 2020). During plant reproductive development, dynamic chromatin remodelling including the localizations of DNA methylation and specific histone species occurs (Ingouff *et al.*, 2010; Jullien *et al.*, 2012; Kawakatsu *et al.*, 2017; Tao *et al.*, 2017; Gehring, 2019). It is unpredictable from the cultured cell-based experiment how such chromatin remodelling could have an influence on the *de novo* transcription in the plant genome over generations.

In Chapter 5, we aimed to establish a model system to elucidate the mechanism by which inserted foreign 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 (Chapter 3) (Satoh *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, 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 newly emerged coding sequences become transcriptionally activated and fixed in the plant genome at their early evolutionary stages.

Materials and Methods

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 short sequences (5'-aggcctcgacgttatcagcttacag-3'), a 12 bp random sequence ('barcode'), a promoterless *LUC*-coding sequence, 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 using a modified pGreenII vector (Materials and Methods in Chapter 3) (Satoh *et al.*, 2020). *Agrobacterium tumefaciens* (GV3101) cells were transformed with the Ti-plasmid libraries. *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. For the promoter-trap experiment, 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 in 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, and subjected to the preparation of the NGS libraries.

Determination of the *LUC* insertion sites

NGS libraries for determining *LUC* insertion loci were prepared according to a published method (Materials and Methods in Chapter 3) (Satoh *et al.*, 2020) with modifications as follows. Genomic DNA (2.0 μg) was digested completely with DpnII, MseI or ApoI, and then purified using the QIAquick PCR purification Kit (QIAGEN). Each digested DNA (600 ng) was independently circularized with T4 DNA ligase. An aliquot of each circularized DNA was subjected to inverse PCR using primer sets that hybridize within the *LUC* ORF. Subsequently, NGS libraries were

prepared by two rounds of PCR; the first round was performed to add Illumina adapters, and the second was carried out using Nextera XT index primers (Illumina). Sequencing was performed using a 301 bp paired-ended protocol on an Illumina MiSeq platform. All primers used in this study are listed in Table 5.S2.

For the determination of each *LUC* insertion site, NGS reads were first processed, before mapping to the genome according to a published method (Materials and Methods in Chapter 4) (Hata *et al.*, 2021), with the following modifications. NGS reads were aligned to the T-DNA vector sequence (5'-tcaaggcctcgacgttatcagcttacagNNNNNNNNNNNNATGGAAGACGCCAAAAACATAAAGAAAGGCCCGGCGCCATTCTATCCTCTAGAG-3'; lowercase, border sequence; N, barcode; underlined, *LUC* fraction) 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.

Determination of the relative transcription level of *LUC* genes

NGS libraries for determining *LUC* transcription level were prepared according to the TRIP method (Materials and Methods in Chapter 3) (Satoh *et al.*, 2020) with modifications as follows. RNA (5.0 μg) was subjected to reverse transcription using an oligo dT_{15} primer and SuperScript III Reverse Transcriptase (Thermo Fisher Scientific). An NGS library (termed RNA library, hereafter) was prepared by amplification of the barcode region of *LUC*-cDNA using primer sets with an Illumina adapter extension, followed by the tailed-PCR using Nextera XT index primers. From an aliquot of DNA used in the *LUC* insertion site determination, barcode regions were amplified and an NGS library (termed DNA library, hereafter) was prepared according to the method described above. Sequencing was performed on the Illumina MiSeq under a 76 bp paired-ended protocol.

To obtain an indicator of the molecular abundances of each *LUC*-mRNA per transgenic cells, barcode sequences were extracted from the sequencing reads and counted. Barcodes with a read number ≤5 in the DNA library were omitted from further analysis. In the RNA library,

barcodes with a read number ≤5 were set as zero. For DNA and RNA libraries, the read number of each barcode was normalized to the total sequencing reads of the corresponding library. The relative transcription level of each *LUC* gene was calculated as follows: the RNA read number of each barcode was divided by the corresponding DNA read number and multiplied by 10,000. Subsequently, individual *LUC* loci and transcription levels were associated 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 5.4d and e, and 5.5b) were prepared and grown under the same condition as that used for T2-plants. ChIP and MBDIP were performed according to a published method (Materials and Methods in Chapters 3 and 4) (Satoh *et al.*, 2020; Kudo *et al.*, 2020; Hata *et al.*, 2021), 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 5.S2) 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.

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 (Chapter 3) (Satoh *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 5.1). Each *LUC* gene was tagged by distinct short random sequences called 'barcodes' (Figure 5.1), 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 5.1). 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 (Chapter 3) (Akhtar *et al.*, 2013; Satoh *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 5.1). 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) (see Chapter 3). 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 5.1 and 5.2a, and Table 5.S1). 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 5.S1). As shown in Figure 5.2a, 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 (Chapter 3) (Kim, Veena and Gelvin, 2007; Satoh *et al.*, 2020). One-third of the 76 *LUC* genes (n = 27) were transcribed (Figure 5.2b). To examine further how 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 5.2c). 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.

In Chapter 3, we found that exogenously inserted promoterless genes became transcriptionally activated in two distinct types: promoter trapping and *de novo* transcriptional activation (Chapters 3 and 4) (Satoh *et al.*, 2020; Hata *et al.*, 2021). 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 5.2d and e). As shown in Figure 5.2d, ~30% of the promoterless *LUC* genes were transcribed similarly in T2-plants and cultured cells (Figure 5.2d). Their relative transcription levels ranged from 10^1 to the 10^7 orders, with a peak at 10^4 (Figure 5.2e). Regarding the three insertion types, the abundances of transcribed *LUC* genes were almost the same in T2-plants and cultured cells, except for the Genic Sense type, in which the transcribed fraction was much greater in the T2-plants (Figure 5.2d). 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 5.2e).

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 5.2f, 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 5.2f). As the Genic Sense and AS types were transcribed presumably by trapping the transcriptional activities of endogenous genes (Chapter 4) (Hata *et al.*, 2021), these features suggested that gene-trapping events are more prone to occur in plants than in 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 growing cultured cells (Figure 5.2d 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

Are there any other similarities/differences between T2-plants and cultured cells? To address this question, we next focused on the correlation of the transcriptional status between the *LUC* genes and the corresponding WT loci. For this, we prepared a dataset of the transcribed regions of 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 WT dataset represents mostly (97.8%) the annotated genic regions, which cover 70.4% (19,308/27,416) of the annotated protein-coding genes. We classified the *LUC* insertion loci into four types by the combination of the transcriptional status of the *LUC* genes and the corresponding WT loci: (i) a *LUC* gene was transcribed in the WT transcribed region; (ii) a *LUC* gene was untranscribed in the WT transcribed region; (iii) a *LUC* gene was transcribed in the WT untranscribed region; and (iv) a *LUC* gene was untranscribed in the WT untranscribed region. The relative abundance of each type in T2-plants and T87 cells (Satoh *et al.*, 2020) was as follows: (i) 14.5% and 7.8%, (ii) 7.9% and 8.2%, (iii) 21.1% and 22.3% and (iv) 56.6% and 61.7%, respectively (Figure 5.3a). Based on these data, we evaluated the transcriptional activation rates in the WT untranscribed regions more precisely. We then redrew Figure 5.3a using the sum of types (iii) and (iv) as 100% (Figure

5.3b). In this data presentation, the transcriptional activation frequency was surprisingly similar between T2-plants and cultured cells (27.1 vs. 26.5 in Figure 5.3b), which supports the contention that *de novo* transcriptional activation in the untranscribed region occurs similarly in plants and cultured cells. Next, we performed the same analysis for the WT transcribed regions (types (i) and (ii)), and showed that the transcriptional activation frequency was higher in the T2-plants than in the cultured cells (66.0 vs. 48.7 in Figure 5.3c). This feature of the transcribed regions (Figure 5.3c) was reminiscent of the feature of the annotated genes (Figure 5.2d), because most of the WT transcribed regions (97.8%) represent annotated protein-coding genes. They both showed that the *LUC* inserts in the genic regions were activated more frequently in the T2-plants than in the cultured cells. The possible explanations for this feature from the viewpoint of plant life cycles and Km-based selection during the T2-plants establishment are referred to in the discussion section.

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 (Figure 3.2b in Chapter 3) (Satoh *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 5.3d). 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.

Chromatin remodelling occurred across the newly activated transcribed regions

Eukaryotic transcription is regulated by the control of the localization of transcription-related chromatin marks (Haberle and Stark, 2018; Andersson and Sandelin, 2020). Therefore, next we focused on the chromatin configuration around *LUC* inserts to examine whether the transcribed T2-plants were regulated by such chromatin marks. First, we screened T2-plants according to

the following criteria: the 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 5.4a 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 5.4a). 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 5.4b). Transcription of inserted promoterless *LUC* genes was verified in both lines by reverse-transcription quantitative-PCR (Figure 5.4c, and Figure 5.S2), whereas no RNA-sequencing reads were mapped on the same strand of each *LUC* insert in the corresponding WT genome, indicating that they were inherently untranscribed regions. For these two lines, we scanned the localization of chromatin marks around the *LUC* insertion loci and compared them with those obtained from the WT genome. In this study, we analysed three transcription-related chromatin 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 5.4d and e, upper and middle panels), which agreed with the general properties of these epigenetic marks (Jones, 2012; Wagner and Carpenter, 2012). However, in the T2-plants, these two chromatin marks were not found within the *LUC* gene bodies (Figure 5.4d and e, upper and middle panels). Although weak signals were observed 200 bp upstream from the *LUC* insert in the T2:161 line (Figure 5.4d, upper and middle panels), they reflected the chromatin 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 chromatin marks (Figure 5.4d and e, lower panels). 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 5.4d and e, lower panels). 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 chromatin 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 5.4c, and Figure 5.S2).

In the T2:161 line, H2A.Z was newly localized 200 bp upstream from the *LUC* insert (Figure 5.4d, lower panel), which suggests that chromatin remodelling 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 5.5a, and Figure 5.S3). Sanger sequencing revealed that this transcript was spliced (Figure 5.5a, and Figure 5.S3). We reanalysed the distribution profiles of H3K36me3 and H2A.Z around the determined TSS (Figure 5.5b). 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 5.5b, 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 5.5b, lower panel).

Overall, the chromatin and TSS analyses revealed that exogenously inserted promoterless genes acquired a brand-new chromatin configuration, and that such chromatin remodelling occurred throughout the newly activated transcription unit. In addition, this chromatin remodelling might have been involved in the transcriptional behaviour of the trapping type of *LUC* transcription (Figure 5.3d): *de novo*-activated transcription events concomitant with the chromatin remodelling might overwrite their inherent transcriptional status.

Discussion

In this Chapter 5, 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 (Chapter 3) (Satoh *et al.*, 2020) showed that *de novo* transcriptional activation is an inheritable phenomenon of the plant

genome (Figures 5.1–5.3). We also demonstrated that chromatin remodelling occurred across the transcribed regions of the inserted coding sequences in the selected two transgenic lines (Figures 5.4 and 5.5), which probably regulated the newly activated transcription of these loci by overwriting the inherent chromatic and 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 5.5a). It is plausible to propose that this was caused by activating (rather than trapping) a cryptic antisense transcript of the given locus (Figure 4.S3 in Chapter 4) (Hata *et al.*, 2021) . 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 (Figure 4.5 in Chapter 4) (Hata *et al.*, 2021) . 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 5.4e).

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 5.1). 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 5.2d and 5.3b), and the transcriptional strength of such activated transcription tended to be bipolarized to lower and higher transcription levels according to the insertion type (Figure 5.2f). 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. We observed a weak insertion preference of *LUC* genes in the accessible chromatin regions (Table 5.S1) by utilizing a Plant Chromatin State Database (Liu *et al.*, 2018)(http://systemsbiology.cau.edu.cn/chromstates). However, we could not evaluate any clear correlation between the transcriptional activation of promoterless *LUC* genes and the chromatin states of the corresponding WT loci, probably because the detected *LUC* population was not sufficiently large for such an analysis. 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).

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 accompanying 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.

References of Chapter 5

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(4), 914–927.

Andersson, R. and Sandelin, A. (2020) Determinants of enhancer and promoter activities of regulatory elements. *Nat Rev Genet,* 21(2), 71–87.

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(8), 474–485.

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.

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(3), 327–336.

Cermakova, K., Smith, E., Veverka, V. and Hodges, H. (2019) Dynamics of transcription-dependent H3K36me3 marking by the SETD2:ISW1:SPT6 ternary complex. *bioRxiv* [posted 2019 May 14]. Available from: https://www.biorxiv.org/content/10.1101/636084v1 doi: 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(6), 735–743.

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(1), 134–145.

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(1), 74–83.

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(1), 15–21.

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), 1–8.

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(2), 149–157.

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(3), 464–477.

Gehring, M. (2019) Epigenetic dynamics during flowering plant reproduction: evidence for reprogramming? *New Phytol,* 224(1), 91–96.

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(1), 2828.

Haberle, V. and Stark, A. (2018) Eukaryotic core promoters and the functional basis of transcription initiation. *Nat Rev Mol Cell Biol,* 19(10), 621–637.

Hata, T., Satoh, S., Takada, N., Matsuo, M. and Obokata, J. (2021) Kozak sequence acts as a negative regulator for de novo transcription initiation of newborn coding sequences in the plant genome. *Mol Biol Evol*. msab069

Ingouff, M., Rademacher, S., Holec, S., Soljić, L., Xin, N., Readshaw, A., Foo, S. H., Lahouze, B., Sprunck, S. and Berger, F. (2010) Zygotic resetting of the HISTONE 3 variant repertoire participates in epigenetic reprogramming in Arabidopsis. *Curr Biol,* 20(23), 2137–2143.

Jones, P. A. (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet,* 13(7), 484–492.

Jullien, P. E., Susaki, D., Yelagandula, R., Higashiyama, T. and Berger, F. (2012) DNA methylation dynamics during sexual reproduction in Arabidopsis thaliana. *Curr Biol,* 22(19), 1825–1830.

Kawakatsu, T., Nery, J. R., Castanon, R. and Ecker, J. R. (2017) Dynamic DNA methylation reconfiguration during seed development and germination. *Genome Biol,* 18(1), 171.

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(5), 779–791.

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(1), 165–174.

Kudo, H., Matsuo, M., Satoh, S., Hachisu, R., Nakamura, M., Yamamoto, Y., Yoshiharu, Hata, T., Kimura, H., Matsui, M. and Junichi, O. (2020) Cryptic promoter activation occurs by at least two different mechanisms in the *Arabidopsis* genome. *bioRxiv* [posted 2020 Nov 28]. Available from: https://www.biorxiv.org/content/10.1101/2020.11.28.399337v1 doi: 10.1101/2020.11.28.399337

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(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(22), 12715–12722.

Lei, B. and Frederic, B. (2020) H2A Variants in Arabidopsis: Versatile Regulators of Genome Activity. *Plant Communications,* 1, 100015.

Liu, Y., Tian, T., Zhang, K., You, Q., Yan, H., Zhao, N., Yi, X., Xu, W. and Su, Z. (2018) PCSD: a plant chromatin state database. *Nucleic Acids Res,* 46(D1), D1157–D1167.

Maruyama, K. and Sugano, S. (1994) Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. *Gene,* 138(1-2), 171–174.

Meinke, D. W. (2020) Genome-wide identification of EMBRYO-DEFECTIVE (EMB) genes required for growth and development in Arabidopsis. *New Phytol,* 226(2), 306–325.

Palovaara, J., Saiga, S., Wendrich, J. R., van 't Wout Hofland, N., van Schayck, J. P., Hater, F., Mutte, S., Sjollema, J., Boekschoten, M., Hooiveld, G. J. and Weijers, D. (2017) Transcriptome dynamics revealed by a gene expression atlas of the early Arabidopsis embryo. *Nat Plants,* 3(11), 894–904.

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(3), 290–295.

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(7), 528–534.

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(1), pdb.prot5559.

Satoh, S., Hata, T., Takada, N., Tachikawa, M., Mitsuhiro, M., Kushnir, S. and Obokata, J. (2020) Plant genome response to incoming coding sequences: stochastic transcriptional activation independent of integration loci. *bioRxiv* 401992 [posted 2020 Nov 28; revised 2021 Feb 4]. Available from: https://www.biorxiv.org/content/10.1101/2020.11.28.401992v2 doi: 10.1101/2020.11.28.401992

Shi, D., Jouannet, V., Agustí, J., Kaul, V., Levitsky, V., Sanchez, P., Mironova, V. V. and Greb, T. (2020) Tissue-specific transcriptome profiling of the Arabidopsis inflorescence stem reveals local cellular signatures. *Plant Cell*. doi: 10.1093/plcell/koaa019.

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(7), e1006875.

Shulse, C. N., Cole, B. J., Ciobanu, D., Lin, J., Yoshinaga, Y., Gouran, M., Turco, G. M., Zhu, Y., O'Malley, R. C., Brady, S. M. and Dickel, D. E. (2019) High-Throughput Single-Cell Transcriptome Profiling of Plant Cell Types. *Cell Rep,* 27(7), 2241–2247.e4.

Slane, D., Kong, J., Berendzen, K. W., Kilian, J., Henschen, A., Kolb, M., Schmid, M., Harter, K., Mayer, U., De Smet, I., Bayer, M. and Jürgens, G. (2014) Cell type-specific transcriptome analysis in the early Arabidopsis thaliana embryo. *Development,* 141(24), 4831–4840.

Springer, P. S. (2000) Gene traps: tools for plant development and genomics. *Plant Cell,* 12(7), 1007–1020.

Tao, Z., Shen, L., Gu, X., Wang, Y., Yu, H. and He, Y. (2017) Embryonic epigenetic

reprogramming by a pioneer transcription factor in plants. *Nature,* 551(7678), 124–128.

Teissandier, A. and Bourc'his, D. (2017) Gene body DNA methylation conspires with H3K36me3 to preclude aberrant transcription. *EMBO J,* 36(11), 1471–1473.

Wagner, E. J. and Carpenter, P. B. (2012) Understanding the language of Lys36 methylation at histone H3. *Nat Rev Mol Cell Biol,* 13(2), 115–126.

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(15), 1793–1797.

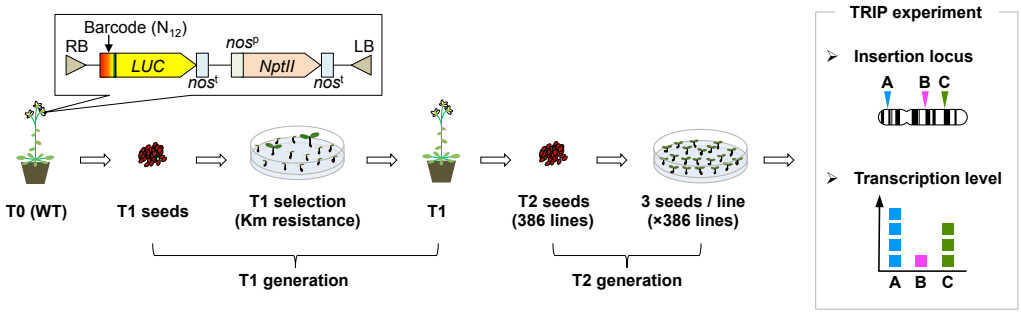


Figure 5.1. Experimental design of the promoter-trap experiment in *A. thaliana* plants.
Schematic illustration of the TRIP experiment performed in the T2 generation of *A. thaliana* transgenic lines. T-DNA including a barcode, a promoterless *LUC* gene and an expression cassette with a Km-resistance 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.

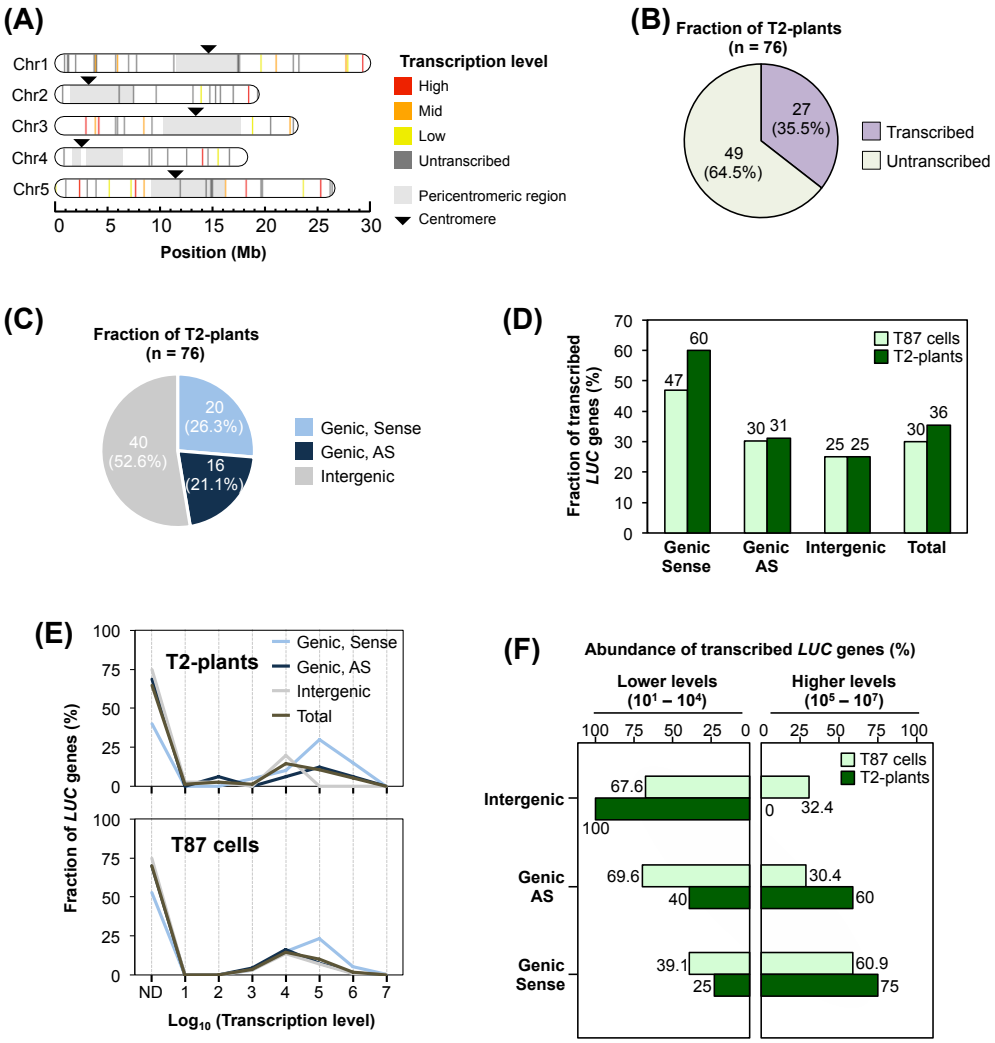


Figure 5.2. An artificial evolutionary experiment revealed the genetic behaviours of the activated transcription of coding sequences inserted in *A. thaliana* plants.
(**a**) The insertion loci and transcription levels of determined T2-plants (n = 76) were mapped on the *A. thaliana* chromosomes. The 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 Materials and Methods. (**d**) Fraction of transcribed *LUC* genes among T2-plants (n = 76) and T87 cultured cells (n = 4,443) (Satoh *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 *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%.

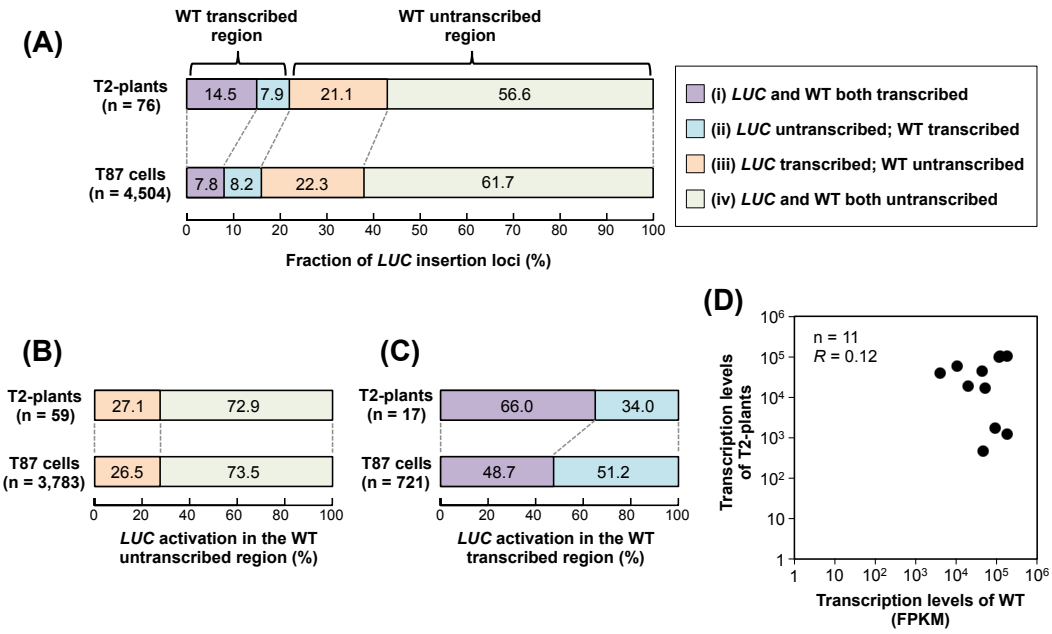


Figure 5.3. Transcription status of *LUC* insertion loci in T2-plants and WT plants.
(**a**) *LUC* insertion loci were classified by the combination of the transcription status of the *LUC* gene and the corresponding WT locus as follows; (i) *LUC* gene and the corresponding WT locus were both transcribed; (ii) *LUC* gene was untranscribed, but WT locus was transcribed; (iii) *LUC* gene was transcribed, but WT locus was untranscribed; and (iv) *LUC* gene and WT locus were both untranscribed. Upper panel, T2-plants; lower panel, T87 cells (Satoh *et al.*, 2020). (**b**) Relative fractions of types (iii) and (iv) normalized to the sum of their abundances in (a), as 100%. (**c**) Relative fractions of types (i) and (ii) normalized to the sum of their abundances in (a), as 100%. (**d**) The transcription levels of T2-plants and corresponding inherent transcripts in WT plants were compared. *R*, Spearman's correlation test.

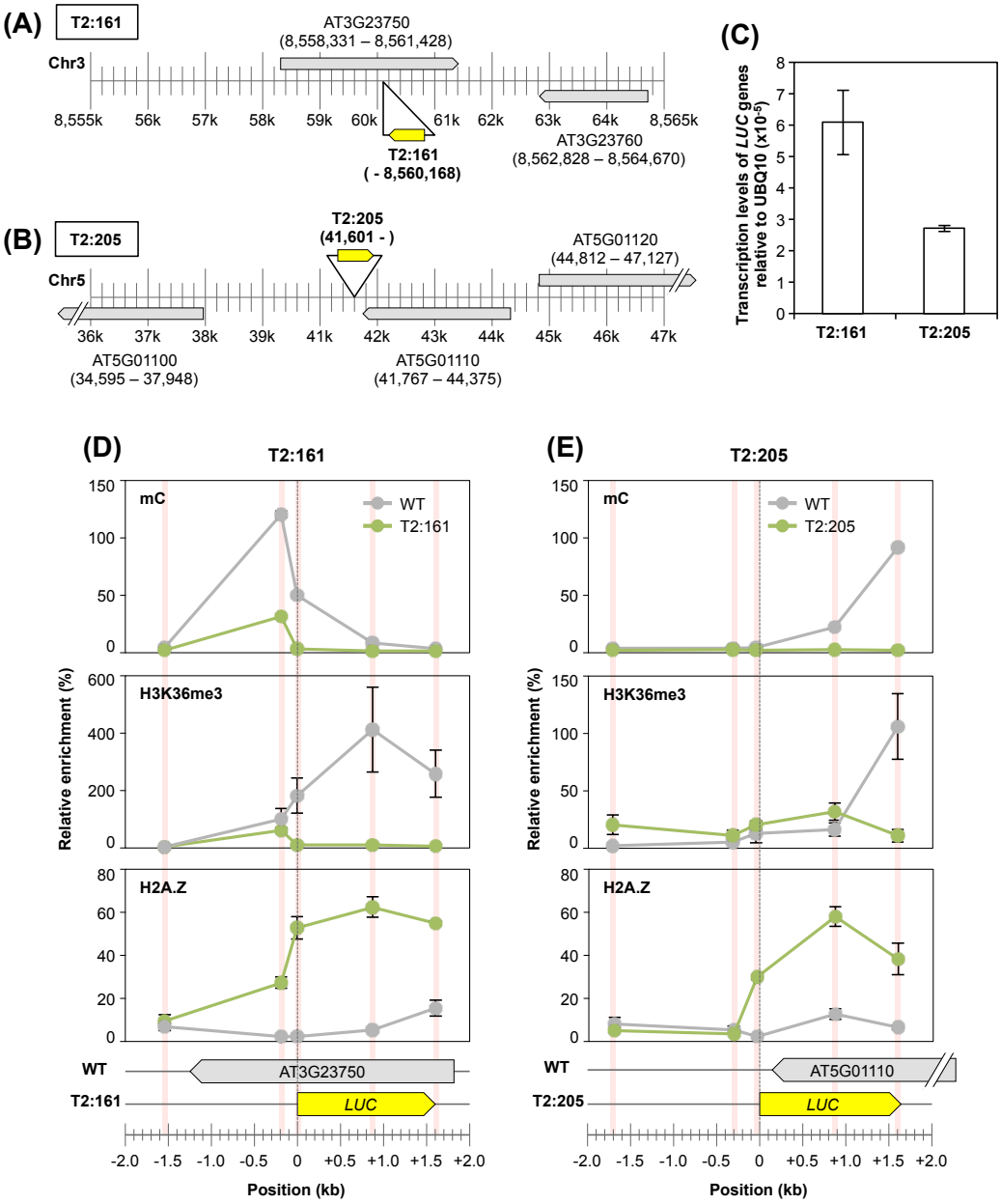


Figure 5.4. Localization analysis of chromatin marks in selected T2-plants.
(**a** and **b**) Locus details of the (a) T2:161 and (b) T2:205 lines. The genomic loci of *LUC* inserts are represented as the individual position of the RB–genome junction. (**c**) Transcription levels of the T2:161 and T2:205 lines relative to the endogenous *UBQ10* gene (AT4G05320). (**d** and **e**) Localization patterns of three chromatin 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. Individual localization signals were normalized to the enrichment of the control locus of each chromatin mark (see Materials and Methods) 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.

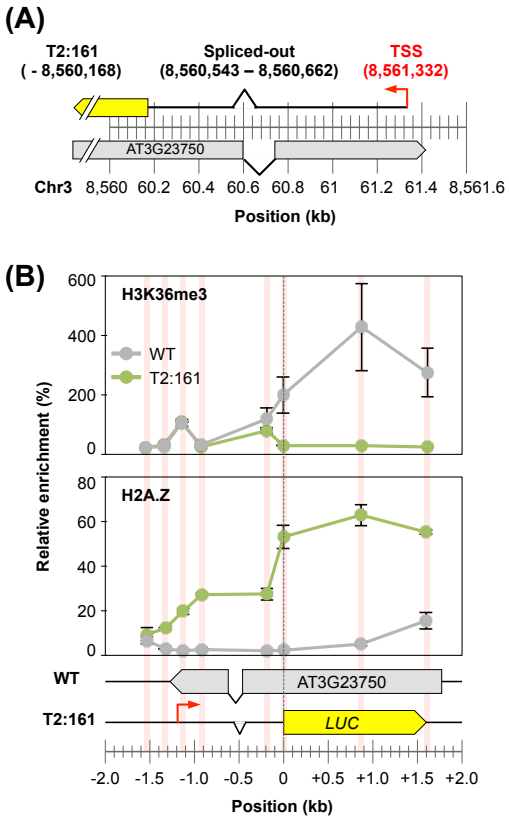


Figure 5.5. Localization of chromatin marks around the transcription start site of the T2:161 line. (**a**) Transcription start site of the T2:161 line, as determined using a template-switching-based method. (**b**) Localization analysis of H2A.Z and H3K36me3 in T2:161 and WT plants, as in Figure 5.4d.

Supporting Information of Chapter 5

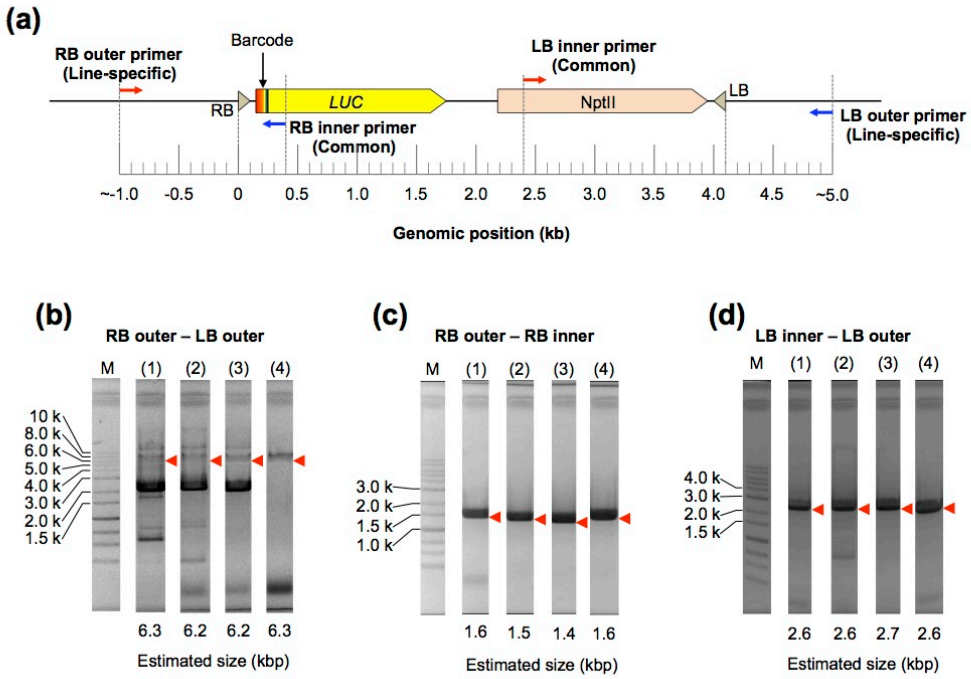


Figure 5.S1. Validation of *LUC* insertion loci. (**a**) Schematic illustration of PCR-based validation of *LUC* insertion locus. In each selected transgenic line, RB outer and LB outer primers were designed about ±1.0 kb from RB and LB, respectively. RB inner and LB inner primers were in common with 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.

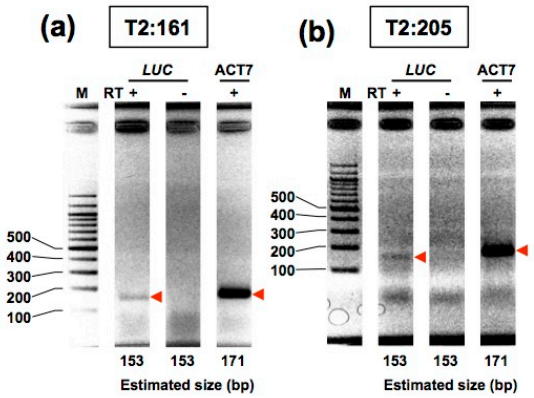


Figure 5.S2. Expression analysis of T2-plants. (**a** and **b**) Expressions of (**a**) T2:161 line and (**b**) T2:205 line were validated by RT-PCR followed by gel electrophoresis. The bands corresponding to the expected sizes were indicated by red triangles. M: Molecular size marker, RT: Reverse transcription, and ACT7: AT5G09810.

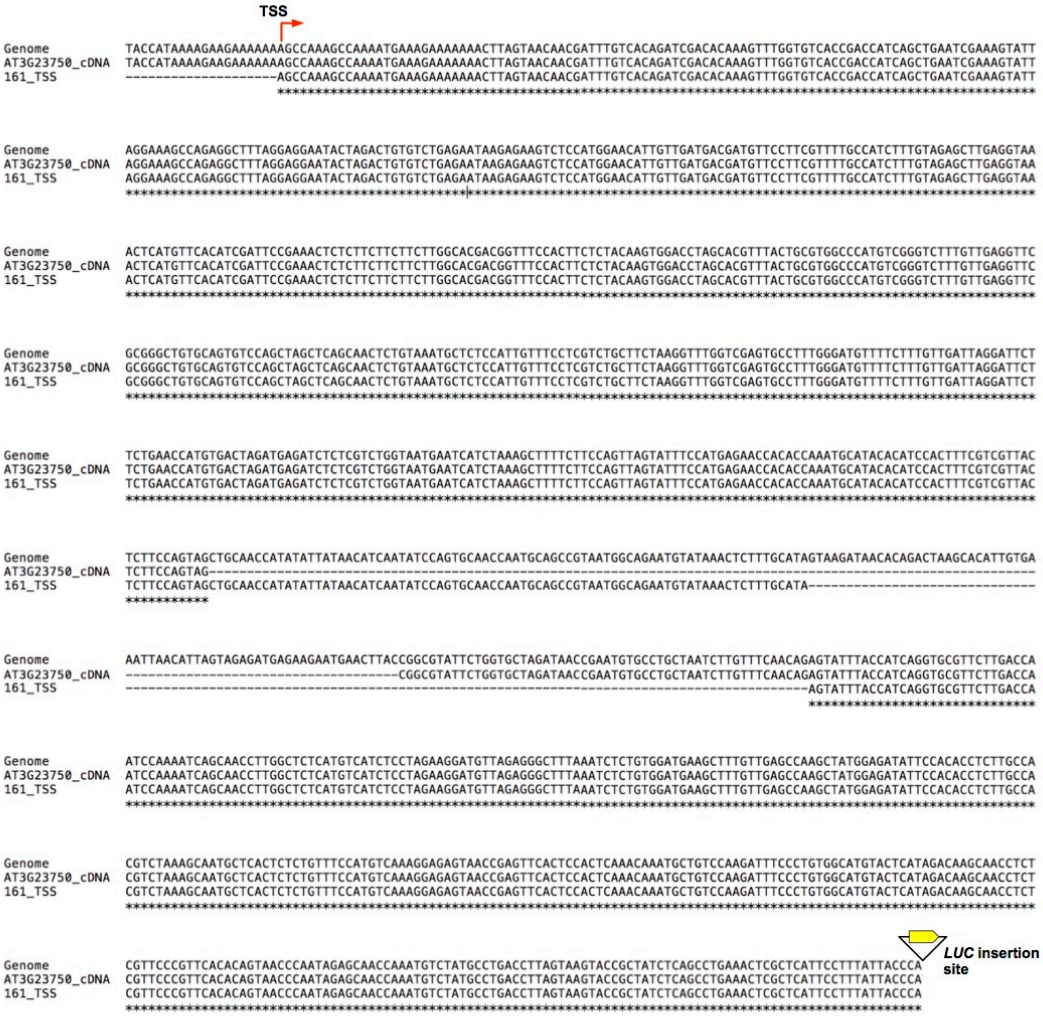


Figure 5.S3. Sequence alignment of TSS of T2:161 line. Multiple alignments among genome, cDNA sequence of AT3G23750, and TSS flanking sequence of T2:161 line were showed. The red arrow indicated the TSS of the T2:161 line. Dash lines in the alignments indicated spliced regions. The alignment was calculated by using Mafft (v7.221) (Katoh and Standley, 2013).

Table S1. List of the LUC lines.

Chromosome[*1]	Insertion site [*1]	Strand[*1]	Border sequence [*2,*3]	Barcode sequence [*2]	LUC transcription level[*4]	WT transcription (FPKM)[*4]	Insertion type[*4]	Distance to the nearest neighboring gene (bp)[*5]		Chromatin states[*6]	Preferential chromatin marks[*7]
								on the same strand	on the opposite strand		
Chr1	901792	+	TCAAGGCCTCGACGTTATCAGCTTACA	AAGCCTCTCGTC	0	9.2	AT1G03610; Genic Sense	−2137	2527	25	histone acetylation,H3K4me3,H3K4me2,H2A.Z
Chr1	1192342	-	CAAGGCCTCGACGTTATCAGCTTACAG	TGATTAGATCAA	0	None	AT1G04420; Genic AS	−930	5776	28	H3K4me3,H3K4me2,H2A.Z
Chr1	1299593	+	AAGGCCTCGACGTTATCAGCTTACAG	CCCTATATGCAA	0	None	Intergenic	2268	−591	20	accessible DNA
Chr1	1965690	-	CAAGGCCTCGACGTTATCAGCTTACAG	GTGGTTTGTCTA	0	None	Intergenic	−639	150	17	accessible DNA
Chr1	3748754	-	CAAGGCCTCGACGTTATCAGCTTACAG	CGCTCTAAGTCA	0	None	Intergenic	1312	−233	16	accessible DNA
Chr1	3888164	+	AGGCCTCGACGTTATCAGCTTACAG	AGAGATTGTCTA	4870	None	Intergenic	526	−7803	17	accessible DNA
Chr1	3888171	-	TCAAGGCCTCGACGTTATCAGCTTACA	TTTTTTTTTCAA	0	None	Intergenic	−7810	519	17	accessible DNA
Chr1	5836536	-	CCTCGACGTTATCAGCTTACAG	AAAAGTTTTCAA	0	None	Intergenic	−1080	1054	16	accessible DNA
Chr1	5999781	+	CAAGGCCTCGACGTTATCAGCTTACAG	ACACAATCTTGG	19422	2.0	AT1G17460; Genic Sense	−1306	−2996	6	H3K4me1,H3K36me3
Chr1	7078843	+	TCAAGGCCTCGACGTTATCAGCTTACA	CGGTCTCTTAAC	0	None	Intergenic	−1237	55	30	rare signal (intergenic)
Chr1	7852701	-	TCAAGGCCTCGACGTTATCAGCTTACA	GAGTGTTGTCTA	0	None	Intergenic	6929	383	30	rare signal (intergenic)
Chr1	11441693	-	CAAGGCCTCGACGTTATCAGCTTACAG	CCATATTGTCCA	0	None	Intergenic	1220	−1584	14	H3K27me3
Chr1	17629578	+	TCAAGGCCTCGACGTTATCAGCTTACA	ATTTAAAGTCCA	0	None	Intergenic	5179	−10684	13	H3K27me3,H2A.Z
Chr1	17629583	+	AGGCCTCGACGTTATCAGCTTACAG	ACAATTGTGCAA	0	None	Intergenic	5174	−10689	13	H3K27me3,H2A.Z
Chr1	19873522	-	TCAAGGCCTCGACGTTATCAGCTTACA	TATAACACAATT	1537	None	Intergenic	−3427	−32	18	accessible DNA
Chr1	21330400	-	CCTCGACGTTATCAGCTTACAG	GATGGGTAGTCA	3840	None	Intergenic	112	−4148	2	H3.3, histone acetylation,H3K4me2,H2A.Z
Chr1	22981288	+	AAGGCCTCGACGTTATCAGCTTACAG	GGTGTTTGCAAA	0	None	AT1G62190; Genic Sense	−3475	748	3	H3K4me1,H3.3,H3.1
Chr1	23521018	-	GGCCTCGACGTTATCAGCTTACAG	GTAGCAATGCAA	0	None	Intergenic	−9021	1479	18	accessible DNA
Chr1	28062066	+	TCAAGGCCTCGACGTTATCAGCTTACA	GGTTAGTTACTA	16151	None	AT1G74690; Genic AS	−947	−13939	5	H3K4me1,H3K36me3,H3.3,H3.1
Chr1	28240131	-	TCAAGGCCTCGACGTTATCAGCTTACA	CGGCCCGCTTGT	29	None	Intergenic	4186	941	29	weak signal (intergenic)
Chr1	29680745	+	AGGCCTCGACGTTATCAGCTTACAG	GGTCTTTGTCTA	51231	None	AT1G78930; Genic AS	−3208	77	23	accessible DNA,H3K36ac,H3K56ac,H4K16ac,H3K4me3
Chr2	714944	-	TCAAGGCCTCGACGTTATCAGCTTACA	GAGTTAGTGTCC	0	None	Intergenic	−14027	−2125	14	H3K27me3
Chr2	6168468	-	CAAGGCCTCGACGTTATCAGCTTACAG	GATTCAATGTCA	0	None	Intergenic	2665	−489	12	H3K27me3,H2A.Z
Chr2	7565422	-	CAAGGCCTCGACGTTATCAGCTTACAG	CACTTTTGTAAA	0	None	AT2G17420; Genic AS	8413	−1670	7	H3K4me1,H3K36me3,H3K4me2
Chr2	9746308	+	TCAAGGCCTCGACGTTATCAGCTTACA	CTTTATATTTGT	0	None	Intergenic	2463	−16	21	accessible DNA
Chr2	13304758	+	TCGACGTTATCAGCTTACAG	CTTCGTGTGTCA	0	None	AT2G31220; Genic Sense	−4456	1835	10	H2A.Z
Chr2	14098779	+	TCAAGGCCTCGACGTTATCAGCTTACA	GCTTATTTGCAA	1217	17.8	AT2G33255; Genic Sense	−68	−1865	20	accessible DNA
Chr2	14946009	-	TCAAGGCCTCGACGTTATCAGCTTACA	TTAATATAGCAA	0	None	Intergenic	−652	54	23	accessible DNA,H3K36ac,H3K56ac,H4K16ac,H3K4me3
Chr2	15486445	-	TCAAGGCCTCGACGTTATCAGCTTACA	GATAACAAGTCA	0	None	AT2G36890; Genic AS	2583	15569	15	H3K27me3,accessible DNA
Chr2	15952230	-	GCCTCGACGTTATCAGCTTACAG	CCTTTAAATCAA	0	None	Intergenic	377	−5145	13	H3K27me3,H2A.Z
Chr2	17190745	+	AGGCCTCGACGTTATCAGCTTACAG	ACACTTTGGTCA	0	None	Intergenic	−1994	283	13	H3K27me3,H2A.Z
Chr2	18679855	+	TCAAGGCCTCGACGTTATCAGCTTACA	ACTACAGGAATC	107857	17.7	AT2G45300; Genic Sense	2226	4261	1	H3.3
Chr2	19541479	-	GCCTCGACGTTATCAGCTTACAG	TCACAGCACTGC	0	24.8	Intergenic	−6	1473	23	accessible DNA,H3K36ac,H3K56ac,H4K16ac,H3K4me3
Chr3	2933730	+	TCAAGGCCTCGACGTTATCAGCTTACA	CTATGTTGTCCA	58482	1.1	AT3G09550; Genic Sense	−5515	887	3	H3K4me1,H3.3,H3.1
Chr3	3856342	-	CGACGTTATCAGCTTACAG	TCCAGTAAGCAA	16727	5.2	AT3G12100; Genic Sense	−2622	1501	6	H3K4me1,H3K36me3
Chr3	4184931	-	AAGGCCTCGACGTTATCAGCTTACAG	CTACTAAATCAA	105130	None	AT3G13082; Genic AS	6420	−956	25	histone acetylation,H3K4me3,H3K4me2,H2A.Z
Chr3	5808395	-	AAGGCCTCGACGTTATCAGCTTACAG	CATAAAATACAA	0	1.2	AT3G17030; Genic Sense	871	−5927	26	histone acetylation,H3K4me3,H3K4me2,H2A.Z
Chr3	5993945	-	CAAGGCCTCGACGTTATCAGCTTACAG	GTAGGTGTTATC	0	None	Intergenic	−1033	−5447	17	accessible DNA
Chr3	6659067	+	CCTCGACGTTATCAGCTTACAG	TCACATCCGTCA	0	None	Intergenic	97	−79	23	accessible DNA,H3K36ac,H3K56ac,H4K16ac,H3K4me3
Chr3	8560168	-	TCAAGGCCTCGACGTTATCAGCTTACA	GGAACCTCGCAT	5837	None	AT3G23750; Genic AS	2661	5326	3	H3K4me1,H3.3,H3.1
Chr3	9339821	+	TATCAGCTTACAG	ATCGTTATGCAA	0	None	AT3G25660; Genic AS	−5017	−1756	1	H3.3
Chr3	19056749	-	GCCTCGACGTTATCAGCTTACAG	CGTAAACTGCAA	7	None	Intergenic	264	−4055	17	accessible DNA
Chr3	20810801	+	CAAGGCCTCGACGTTATCAGCTTACAG	ACTGTTTGGCAA	0	None	AT3G56080; Genic AS	12345	−2649	1	H3.3
Chr3	22693250	+	TCAAGGCCTCGACGTTATCAGCTTACA	GGGAAATTTCAA	2637	None	Intergenic	659	−416	19	accessible DNA
Chr3	23006872	-	TCAAGGCCTCGACGTTATCAGCTTACA	AGGATTGTCAAC	0	None	AT3G62130; Genic AS	1883	211	21	accessible DNA
Chr4	850103	+	AAGGCCTCGACGTTATCAGCTTACAG	TGTGTACTAAGG	0	None	Intergenic	−7739	1107	30	rare signal (intergenic)
Chr4	9054156	-	TCAAGGCCTCGACGTTATCAGCTTACA	TTATGAAAGCAA	0	None	Intergenic	3310	−1004	30	rare signal (intergenic)
Chr4	9309047	+	TCAAGGCCTCGACGTTATCAGCTTACA	GGGGATTGTAAA	0	None	Intergenic	2308	−585	29	weak signal (intergenic)
Chr4	10851730	+	GTTATCAGCTTACAG	TACTTTAATCAA	0	None	AT4G20050; Genic AS	−4338	−2615	12	H3K27me3,H2A.Z
Chr4	12666239	+	AGGCCTCGACGTTATCAGCTTACAG	AAGGTTATGCAA	0	None	AT4G24520; Genic AS	−3583	1119	7	H3K4me1,H3K36me3,H3K4me2
Chr4	14211097	+	GACGTTATCAGCTTACAG	ATTCAAGACTGT	100934	12.3	AT4G28760; Genic Sense	3681	758	4	H3K4me1,H3.3
Chr4	14729510	+	TATCAGCTTACAG	ATGCTTGGTCAA	0	None	Intergenic	5309	891	29	weak signal (intergenic)
Chr4	15700350	-	GGCCTCGACGTTATCAGCTTACAG	AATCTTCTTGAG	2210	None	Intergenic	−5250	554	30	rare signal (intergenic)
Chr4	16800071	+	AGGCCTCGACGTTATCAGCTTACAG	GTTTTGTGTTAA	0	None	Intergenic	1916	−294	29	weak signal (intergenic)
Chr5	41601	+	AGGCCTCGACGTTATCAGCTTACAG	GGGCTTTATAAT	1563	None	Intergenic	3212	167	18	accessible DNA
Chr5	1005874	-	AGGCCTCGACGTTATCAGCTTACAG	GTGTAGCCTCAG	0	None	AT5G03790; Genic AS	1516	−8026	11	H3K27me3,H2A.Z,H3K4me2
Chr5	2319523	+	AGGCCTCGACGTTATCAGCTTACAG	GATTACCGTCAA	79916	None	AT5G07340; Genic Sense	−2649	498	1	H3.3
Chr5	3070791	-	GTTATCAGCTTACAG	GTACGGGAGTCA	0	None	AT5G09860; Genic AS	−4902	2565	2	H3.3, histone acetylation,H3K4me2,H2A.Z
Chr5	3914425	-	AGGCCTCGACGTTATCAGCTTACAG	ACTATATCAATA	0	None	Intergenic	1623	43	27	H3K4me2,histone acetylation,H3K4me3,H2A.Z
Chr5	5198938	-	TCAAGGCCTCGACGTTATCAGCTTACA	TTGGTATGTATT	20	None	AT5G15920; Genic AS	−6868	773	4	H3K4me1,H3.3
Chr5	7292543	-	AAGGCCTCGACGTTATCAGCTTACAG	GGGGAGCTGTGA	469	4.8	AT5G22030; Genic Sense	−3061	8688	6	H3K4me1,H3K36me3
Chr5	7743110	-	TCAAGGCCTCGACGTTATCAGCTTACA	TCTGGTTTGCAA	44595	4.3	AT5G23080; Genic Sense	−4575	−132	1	H3.3
Chr5	8547813	+	AAGGCCTCGACGTTATCAGCTTACAG	GCAAATTGTAAA	4557	None	Intergenic	4930	−24	17	accessible DNA
Chr5	12056489	+	TCAAGGCCTCGACGTTATCAGCTTACA	TGTAGAGTGTCA	0	None	Intergenic	20404	25314	35	H3K9me2,DNA methylation,H2A.X
Chr5	14550668	-	TCAAGGCCTCGACGTTATCAGCTTACA	CTTAGTTCGTCA	0	0.2	Intergenic	−4438	−2465	29	weak signal (intergenic)
Chr5	15010915	-	TCAAGGCCTCGACGTTATCAGCTTACA	ACGTAGGGTCAA	0	4.9	AT5G37790; Genic Sense	−4710	−12696	25	histone acetylation,H3K4me3,H3K4me2,H2A.Z
Chr5	15033452	-	CAAGGCCTCGACGTTATCAGCTTACAG	CGAGGTGGTCAA	0	None	Intergenic	−22300	2745	21	accessible DNA
Chr5	15157637	-	AAGGCCTCGACGTTATCAGCTTACAG	GGGAAAACTCAA	0	None	Intergenic	8015	638	14	H3K27me3
Chr5	16423914	+	TTATCAGCTTACAG	GATAATCATTTT	6351	None	Intergenic	−84	1969	23	accessible DNA,H3K36ac,H3K56ac,H4K16ac,H3K4me3
Chr5	18431247	+	CAAGGCCTCGACGTTATCAGCTTACAG	AGACGCCTGTCA	100516	11.4	AT5G45490; Genic Sense	−9285	1211	26	histone acetylation,H3K4me3,H3K4me2,H2A.Z
Chr5	19849566	-	AAGGCCTCGACGTTATCAGCTTACAG	GCCACGTTGCAA	0	None	AT5G48960; Genic AS	−1316	−2131	17	accessible DNA
Chr5	20000062	-	TCAAGGCCTCGACGTTATCAGCTTACA	GCTAGTATGCAA	0	10.7	AT5G49330; Genic Sense	−3748	−16193	29	weak signal (intergenic)
Chr5	23960200	-	AAGGCCTCGACGTTATCAGCTTACAG	TGACTTGAGTCA	1736	9.0	AT5G59410; Genic Sense	−13770	−493	7	H3K4me1,H3K36me3,H3K4me2
Chr5	25633773	-	CAAGGCCTCGACGTTATCAGCTTACAG	AGAGCTGTACAC	39012	0.4	AT5G64060; Genic Sense	−564	−5214	8	H3K4me1,H3K4me2,H2A.Z
Chr5	26549175	-	TATCAGCTTACAG	GCATATAATAGC	0	None	AT5G66500; Genic Sense	529	2704	7	H3K4me1,H3K36me3,H3K4me2
Chr5	26741918	+	AGGCCTCGACGTTATCAGCTTACAG	GATTCCGTGCAA	0	None	AT5G66980; Genic Sense	1361	−793	11	H3K27me3,H2A.Z,H3K4me2

*1. Genomic coordinate is based on the TAIR10 version of *A. thaliana* genome (1-based coordinate).

*2. Border sequence and Barcode sequence are showen with the "5' (RB) -> 3' (LUC)" as sense orientation.

*3. Border sequences identidied in individual inserts are shown with the predicted sequence (5'-TCAAGGCCTCGACGTTATCAGCTTACAG-3') as a reference.

*4. For the details, see Materials and Methods.

*5. Relative distances from the LUC insertion site to the nearest neighboring gene on the same/opposite strand. Negative number means that the LUC gene was located downstream of the neighboring gene.

*6. Chromatin states of LUC insertion sites were analyzed by utilizing a Plant Chromatin State Database (Nucleic Acids Res. 2018 4;46(D1):D1157-D1167.).

*7. Preferential chromatin marks of each chromatin state were according to the PCSD (http://systemsbiology.cau.edu.cn/chromstates).

Table 5.S2. 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	

Chapter 6:

General discussion

To generate genetic novelty is one of the fundamental properties of the genome. As I summarized in Chapter 1, 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, the artificial evolutionary approach described in this thesis sheds light even on evolutionary losers within a much shorter timescale (Chapters 3–5). Our approach based on the use of the cultured cells will be a useful model to investigate the molecular mechanisms underlying promoter birth, thus providing a homogeneous and simple experimental system (Chapter 3 and 4). In contrast, 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 (Chapter 5).

The main aim of this thesis is to reveal the initial appearances of newborn genes that comparative genomics could never approach. For this purpose, we carried out an artificial evolutionary experiment focusing on how newborn coding sequences acquire their initial transcriptional competency shortly after their birth. Firstly, we established high-efficient transformation method of *A. thaliana* T87 cultured cells to obtain massive transformants that can be applied for the MPRA approach (Chapter 2). Then, we established MPRA-based promoter-trap screening and found a novel transcriptional response of the plant genome to the incoming coding sequences (Chapter 3). We characterized such newly activated transcription in greater detail and found the *de novo* transcription (Chapter 4). In the fifth part, we showed genetic behavior of this *de novo* transcription based on the artificial evolutionary experiment in the *A. thaliana* plants under the similar experimental scheme in the cultured cells (Chapter 5).

In this last chapter, I discuss how this project advances our understandings of gene evolution and propose future perspectives.

De novo transcription: a key player on the genome evolution

In the conventional view of HGT/EGT, transcriptional activation of transferred DNA has been postulated mainly by the gene-trapping, in which the transferred DNA is inserted into the pre-existing gene or transcription unit and forms transcriptional fusion (Chapter 1). Contrary to this model, our result indicated that the gene-trapping type of transcriptional activation (i.e., conventional model) might be only a chip of an iceberg of the total functional gene transfer event. We found that only a small portion of the expression of the transferred DNA could be explained by the trapping of pre-existing genetic materials; rather the majority of them were activated via different manner, i.e., integration-dependent stochastic transcriptional activation (Chapter 3). This activation occurs independently of the genomic loci, and inherent transcriptional or heterochromatic status; rather, it does at fixed frequency of each DNA insertion event (Chapter 3). In contrast to the conventional model, this activation seems less harmful to the preexisting host gene networks. Moreover, as the activation could occur at entire genome, the chance of transcriptional activation would become larger than previously thought. Therefore, this transcriptional activation mechanism might contribute to endow transcriptional competency to the incoming foreign DNA, thus, to expand the possibility of the evolutionary innovation of host genome.

The molecular characteristics of such newly activated transcription was provided in Chapter 4. Precise TSS mapping and characterization of promoterless LUC genes led us the transcriptional initiation manners of newly originated coding sequences in the plant genome (Chapter 4). Significantly, identification of de novo TSS provided an evolutionary model of eukaryotic promoter (Figure 4.7 in Chapter 4). Moreover, the collective results suggested a profound role of the coding sequences in their own transcription activations; the coding sequences might act as a *cis*-determinant of the pol II recruitment (Chapter 4).

If this is the case, this mechanism could also be a candidate of the causative mechanism of the transgene transcriptional activation in the HGT/EGT process. One example is an enigmatic expression of the foreign DNA in the plant genome. Plant nuclear genome contains many escaped DNA fragments from the plastid. Based on the genome-wide transcriptome analysis, Wang *et al.* showed that such plastid-derived sequences were expressed (Wang *et al.*, 2014). Because the gene-trapping type of transcriptional activation is rare event (Chapter 3), the expression of the plastid-derived sequences might be owing to the integration-dependent

stochastic manner, or more specifically, to the *de novo* TSS. Another example is found in the functional HGT event between prokaryote and eukaryote. Some yeast species harbor a set of a biosynthetic pathway gene from bacteria via horizontal operon transfer (HOT) (Lindsey and Newton, 2019; Kominek *et al.*, 2019; Gonçalves and Gonçalves, 2019). Interestingly, the transferred polycistronic structural genes via HOT are expressed in the nucleus as monocistronic genes (Gonçalves and Gonçalves, 2019; Kominek *et al.*, 2019). How have the individual structural genes acquired their promoters? If our model is the case, the *de novo* transcription initiation could be a candidate of the molecular mechanism of this polycistronic to monocistonic conversion.

Future directions

The goal of this project is to integrate the molecular (or biochemical) process and evolutionary process in the gene evolution (Figure 6.1). As for the gene origination, a reading frame and its expression are both necessary. The birth of such genetic novelty occurs by biochemical reaction in the very short-time period. However, the evolutionary biology so far focuses on the 'young' genes. Therefore, the studies have still overlooked the intermediate processes by which such newborn genetic novelty become fixed and matured in the genome. *De novo* transcription could be an invaluable empirical model system to approach these processes, which would fill in the blank of the evolutionary biology, and moreover decipher the intrinsic property of the genome. From this viewpoint, I show two future directions here.

Scrutiny of molecular mechanism of *de novo* transcription in greater detail

An intriguing question about the integration-dependent stochastic transcriptional activation is its molecular mechanism. In Chapter 4, we aimed to analyze *cis*-regulatory elements of this type of transcription, and analyzed TSS of newly activated transcriptions of inserted reporter sequences (Chapter 4). However, so far, we could not identify whether the determined *LUC*-TSS was originated via the integration-dependent stochastic transcriptional activation manner. This was manly because of the shortcomings of the experimental design in the TSS determination in Chapter 4. For example, depending on the location where the TSSs occurred, the length of the sequencing library varies, which results in the amplification biases in PCR and sequencing steps.

This makes it difficult to analyze the transcriptional strength of each *LUC*-TSS. In addition, short-read sequence tags cannot provide the information of full-length transcripts, and hence our analysis lacked the information of spliced isoforms of *de novo* activated transcripts. These experimental limitation would be improved by utilizing molecular barcode identifier (Kivioja *et al.*, 2011) and single-molecule long-read sequencer (Werner *et al.*, 2018; Viehweger *et al.*, 2019), and the obtained results will be described in elsewhere.

The stochastic nature of this phenomenon implies this activation occurs not solely by the sequence elements; rather depending on the chromatin epigenetic configurations (Chapter 3). The idea is supported by the genetic screening of enigmatically expressed promoter-trap line in the *A. thaliana.* Kudo *et al.* showed that the insertion of coding sequences activated local chromatin remodeling and resulted in *de novo* formation of promoter-like epigenetic configurations (Kudo *et al.*, 2020). The data in the Chapter 5 also supported this idea; the transcription-related epigenetic marks were appeared all over the *de novo* transcribed regions (Chapter 5). However, because of experimental limitations, we did not show the direct evidence of such epigenetic rewiring in the cultured cell-based experiments (Chapter 3 and 4). Under our experimental condition, individual transgenic lines were identified *in silico* analysis based on the indexed barcode sequences of reporter construct as a molecular identifier. This trick of MPRA approach allows us to handle thousands of transgenic lines without establishing isogenic cell lines. However, the trick also limits the experiment; we could not obtain any information from the sequence data without identifying individual barcode sequences. There is no practical methodology to determine two distinct information (localization patterns of epigenetic marks and the molecular barcode sequence; these two are mutually irrelevant) distantly located on the chromatin with single-molecule resolution. How could we overcome this situation? Recently, single-molecular resolution techniques were reported for the chromatin accessibility assay (Wang *et al.*, 2019; Stergachis *et al.*, 2020; Shipony *et al.*, 2020), but further technical breakthrough is needed to determine the epigenetic marks on the single chromatin molecule. Now we have a plan to develop a novel method to analyze the localization patterns of DNA binding factors on the chromatin with single-molecule resolution, which could be one of the solutions to elucidate epigenetic rewiring around the newly activated transcription.

Empirical simulation of gene evolution

Is *de novo* transcription just a transcriptional noise or a 'meaningful' phenotypic trait? As the transcription level of *de novo* transcription was extremely low, it is not likely to be sufficient for functional message. However, such transcriptional noise probably turns into 'meaningful' transcript by sequence mutations or epigenetic rewiring (Figure 6.2). This assumption agreed with the proto-gene model; non-functional putative ORFs are often transcribed and translated in very low level, and such putative ORFs are in the equilibrium state between being fixed and eliminated from the genome via mutations (Carvunis *et al.*, 2012). If the *de novo* transcription is also in the similar situation, how many generations and population are needed to reach such endings? In other words, how do a *de novo* transcription become a 'gene'? The Chapter 5 was a pilot study of this line. Although the study covered the genetic behavior of *de novo* transcription only after one-generation inheritance, the obtained results indicated that a kind of genetic adaptation has already appeared in the *de novo* transcribed populations (Figure 5.1 in Chapter 5). Thus, the *de novo* transcription could be an empirical model to investigate the molecular processes of gene origination/adaptation/fixation.

From this angle, 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. The collective results would elucidate the nature of *de novo* transcription, which would open up the future of the evolutionary biology.

References of Chapter 6

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(7407), 370–374.

Gonçalves, C. and Gonçalves, P. (2019) Multilayered horizontal operon transfers from bacteria reconstruct a thiamine salvage pathway in yeasts. *Proc Natl Acad Sci U S A,* 116(44), 22219–22228.

Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S. and Taipale, J. (2011) Counting absolute numbers of molecules using unique molecular identifiers. *Nat Methods,* 9(1), 72–74.

Kominek, J., Doering, D. T., Opulente, D. A., Shen, X. X., Zhou, X., DeVirgilio, J., Hulfachor, A. B., Groenewald, M., Mcgee, M. A., Karlen, S. D., Kurtzman, C. P., Rokas, A. and Hittinger, C. T. (2019) Eukaryotic Acquisition of a Bacterial Operon. *Cell,* 176(6), 1356–1366.e10.

Kudo, H., Matsuo, M., Satoh, S., Hachisu, R., Nakamura, M., Yamamoto, Y., Yoshiharu, Hata, T., Kimura, H., Matsui, M. and Junichi, O. 2020. Cryptic promoter activation occurs by at least two different mechanisms in the *Arabidopsis* genome. unpublished data, *bioRxiv* [posted 2020 Nov 28]. Available from: https://www.biorxiv.org/content/10.1101/2020.11.28.399337v1 doi: 10.1101/2020.11.28.399337

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(5), 676–688.

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(7), 2190–2202.

Lindsey, A. R. I. and Newton, I. L. G. (2019) Some Like it HOT: Horizontal Operon Transfer. *Cell,* 176(6), 1243–1245.

Shipony, Z., Marinov, G. K., Swaffer, M. P., Sinnott-Armstrong, N. A., Skotheim, J. M., Kundaje, A. and Greenleaf, W. J. (2020) Long-range single-molecule mapping of chromatin accessibility in eukaryotes. *Nat Methods,* 17(3), 319–327.

Stergachis, A. B., Debo, B. M., Haugen, E., Churchman, L. S. and Stamatoyannopoulos, J. A. (2020) Single-molecule regulatory architectures captured by chromatin fiber sequencing. *Science,* 368(6498), 1449–1454.

Viehweger, A., Krautwurst, S., Lamkiewicz, K., Madhugiri, R., Ziebuhr, J., Hölzer, M. and Marz, M. (2019) Direct RNA nanopore sequencing of full-length coronavirus genomes provides novel insights into structural variants and enables modification analysis. *Genome Res,* 29(9), 1545–1554.

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(6), 1327–1334.

Wang, Y., Wang, A., Liu, Z., Thurman, A. L., Powers, L. S., Zou, M., Zhao, Y., Hefel, A., Li, Y., Zabner, J. and Au, K. F. (2019) Single-molecule long-read sequencing reveals the chromatin basis of gene expression. *Genome Res,* 29(8), 1329–1342.

Werner, M. S., Sieriebriennikov, B., Prabh, N., Loschko, T., Lanz, C. and Sommer, R. J. (2018) Young genes have distinct gene structure, epigenetic profiles, and transcriptional regulation. *Genome Res,* 28(11), 1675–1687.

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(10), 2822–2829.

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(4), 679–690.

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(6172), 769–772.

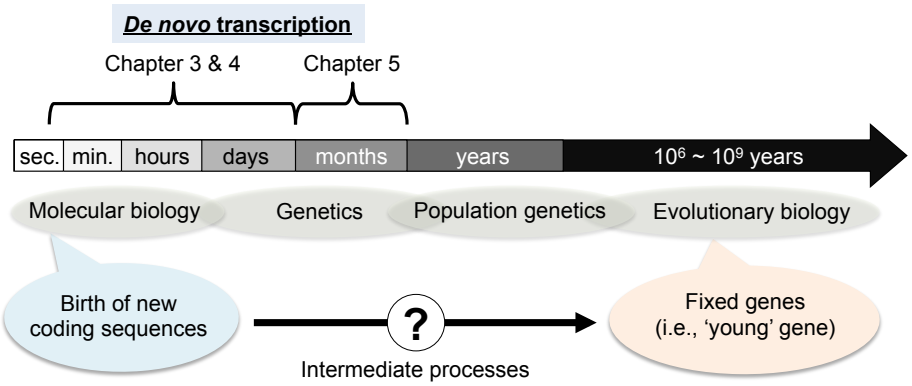


Figure 6.1. Each biological area covers different time-scales
Schematic illustration of time-scales covered by each biology area (molecular biology, genetics, population genetics evolutionary biology). While the evolutionary biology only concerns fixed genes, underlying molecular reactions and mechanisms should be a subject in much shorter time-scale. *De novo* transcription characterized in this thesis covers the timescale from the molecular biology to the genetics, which could be an empirical model to investigate the intermediate processes of gene evolution from the molecular biology to the evolutionary biology.

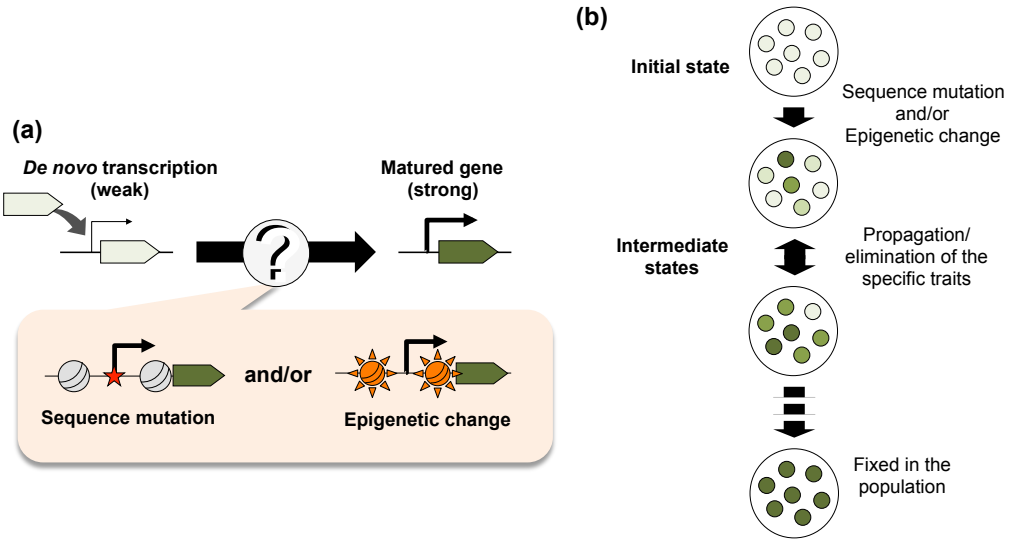


Figure 6.2. A possible process of *de novo* transcription maturation
(**a**) Emergence/insertion of new coding sequences trigger *de novo* transcription. Initially, the transcriptional strength of *de novo* transcription should be weaker than that of the matured genes. If the *de novo* transcription could be a source of new gene, its transcriptional property would alter via sequence mutation and/or epigenetic change. (**b**) Fixation process of *de novo* transcription would depend on whether the message could contribute to raise a fitness of the host genome.

Acknowledgements

It is hard to overstate my gratitude to my supervisor, Prof. Dr. Junichi Obokata, who gave me this thesis subject and supervised me since I was an undergraduate student. He has provided sound advices, good ideas, encouragement, and reliable supports in abundance over these seven years.

I am indebted to Dr. Soichirou, Satoh, Dr. Mitsuhiro Matsuo, Mr. Makoto Tachikawa and Ms. Hisaki Ishii, who have initiated me to the laboratory works and provided helpful advices in scientific questions.

I would like to thank Dr. Kushnir Sergei for the quality of his thoughtful advice and encouragement for the thesis study.

A big thank you to my friends: Naoto Takada and Moyuru Shirasu for encouragements, camaraderie, distractions, and warmth they provided, and even more importantly for all the wonderful smoking camp along the Kamogawa riverside and in Seryo, which does have a certain '*ju ne sais quoi*'.

Thanks to my lab members: Mizuho Susa, Atsushi Katahata, Ai Tsuruoka, Ayaha Tanaka, Hitomi Nakajima, Atsushi Sugioka, Hiroki Fukuizumi, Tomohiro Uchikoba, Chihiro Hayakawa, Mei Kazama, Kouhei Kawaguchi, Kouhei Nishimon, Fumika, Yamaguchi, Ayasa Yoshio, Kouki Mukae, Shirou Aso and Kento Kono for the great atmosphere that they made and that makes me enjoy coming to work there.

I would like to thank the Graduate School of Life and Environmental Sciences, Kyoto Prefectural University for accepting me. Particularly, I would like to express my gratitude to Prof. Dr. Takehiro Masumura for accepting me in charge of a doctor candidate.

I am obliged to the Faculty of Agriculture, Setsunan University for accepting me as a research student and offering helpful supports. In particular, Prof. Dr. Takashi Shiina, Dr. Yusuke Kato, Dr. Minori Numamoto, Dr. Hiromi Ikeda and Ms. Yoko Ishizaki for allowing me to join the lab seminar, for their helpful comments and advices for my thesis, and for all the enjoyable Costco parties.

All the work in this thesis was supported by grants from the Japan Society for the Promotion of Sciences (JSPS), and grant-in-aid from Kyoto Prefectural Public University Corporation. I would like to thank JSPS for accepting me as a research fellow and for funding for three years.