


Kozak Sequence Acts as a Negative Regulator for De Novo Transcription Initiation of Newborn Coding Sequences in the Plant Genome

Takayuki Hata^{1,2}, Soichirou Satoh¹, Naoto Takada¹, Mitsuhiro Matsuo², and Junichi Obokata ^{*,2}

¹Graduate School of Life and Environmental Sciences, Kyoto Prefectural University, Sakyo-ku, Kyoto, Kyoto, Japan

²Faculty of Agriculture, Setsunan University, Hirakata, Osaka, Japan

*Corresponding author: E-mail: junichi.obokata@setsunan.ac.jp.

Associate editor: Naruya Saitou

Abstract

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.

Key words: de novo transcriptional activation, Kozak sequence, artificial evolutionary experiment, transcription start site, promoter evolution, gene evolution.

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 noncoding 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 et al. 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 neighboring eukaryotic promoters or by utilizing the prokaryotic plastid promoter

© The Author(s) 2021. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Open Access

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 (Friedrich and Soriano 1991; Springer 2000). A recent study reported that newly inserted promoterless coding sequences were transcribed without trapping any endogenous genes or transcription units, which indicated the origination of brand-new promoters in the plant genome (Kudo et al. 2020). However, the throughput of these studies was too limited to illustrate the general features of how newborn genes acquire transcriptional competency.

Here, we experimentally simulated gene origination processes in the plant genome to elucidate the manner in which newborn genes become transcriptionally active shortly after their birth. To overcome the low-throughput drawback of promoter/gene-trapping experiments, we previously applied a massively parallel reporter assay (Inoue and Ahituv 2015) to the conventional promoter-trapping screenings, and established transgenic *Arabidopsis thaliana* T87 cell lines individually harboring promoterless *LUC* open reading frames (ORFs) (Satoh et al. 2020). Based on the precise mapping of *LUC*-TSSs, we identified de novo TSSs; they occurred de novo ~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.

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 (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 *LUC*-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 analyzed transcription start sites (TSSs) and insertion loci of the promoterless *LUC* genes. For this sake, we modified the conventional TSS determination method (Cap-trapper 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 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 nonspecifically amplified endogenous cDNAs and erroneous reads generated during the library preparation and sequencing steps (supplementary figs. S1, S2; supplementary methods S1, Supplementary Material online).

Figure 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 (fig. 1B). In total, we identified 550 *LUC* inserts and 858 corresponding TSSs across the *A. thaliana* genome (fig. 1C). Among the 550 *LUC* inserts, 74% were associated with a single TSS and the remainder were associated with two or more TSSs (fig. 1D). The *LUC* inserts were unbiasedly distributed over the *A. thaliana* genome (Satoh et al. 2020), whereas the *LUC* loci identified in this TSS analysis were overrepresented in the genic regions (fig. 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 (fig. 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 data sets of wild-type (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 (fig. 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 (fig. 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 (fig. 2C).

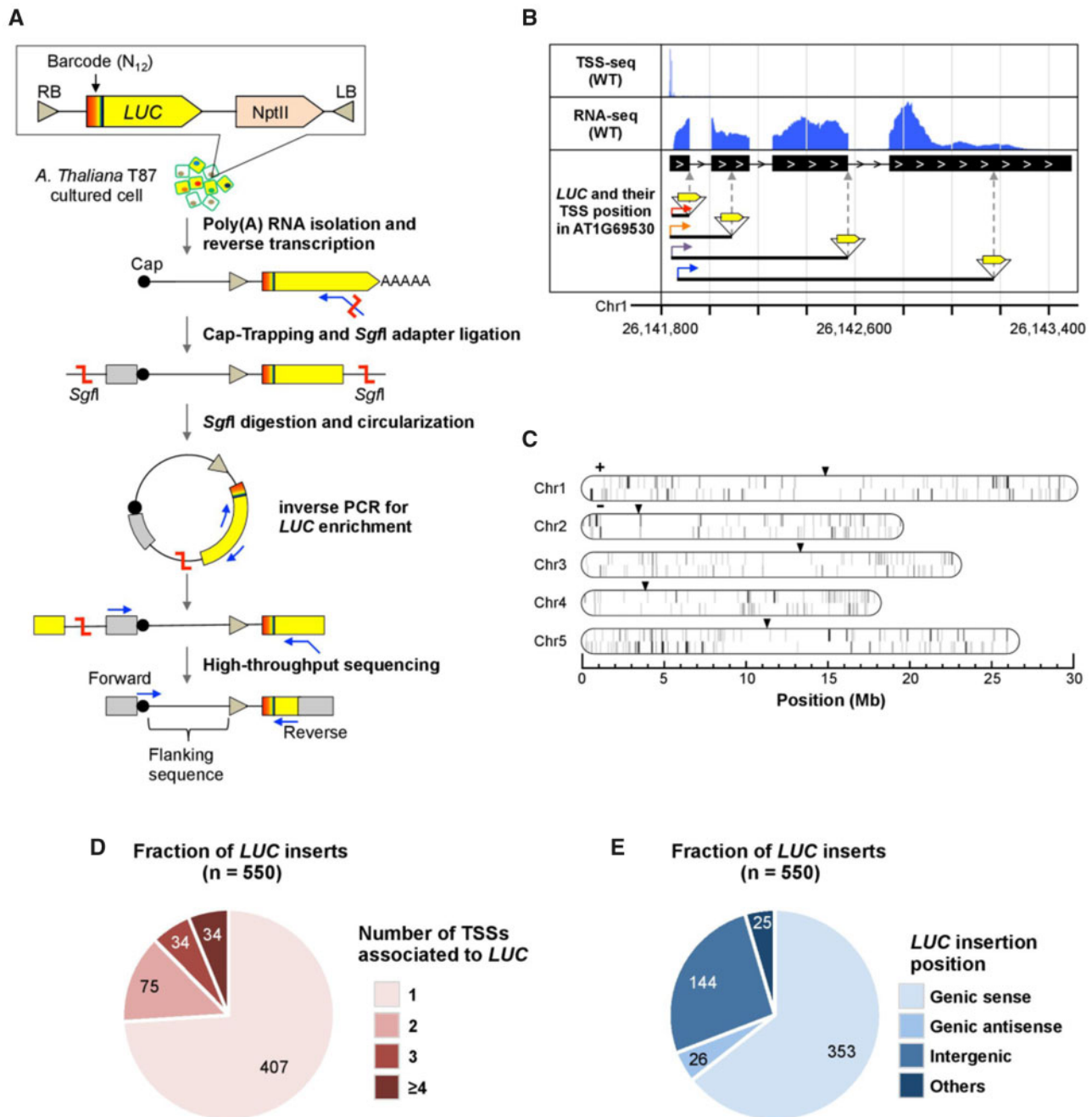


Fig. 1. Determination of the TSSs of promoterless *LUC* genes at single-nucleotide resolution. (A) Experimental design of the parallel determination of promoterless *LUC* insertion sites and their corresponding TSSs. cDNAs reaching the 5' end of *LUC* RNAs were prepared by the Cap-trapper method followed by inverse PCR. Amplified cDNAs were subjected to paired-end sequencing. For details, see the Materials and Methods. (B) Example of determined *LUC*-TSSs in the genome viewer. The colored 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.

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 1) *LUC* loci relative to the WT genes, 2) TSS loci relative to the WT genes,

and 3) types of *LUC*-TSS initiation (fig. 2C), to give 72 TSS types (fig. 3A). Among these 72 types, we identified 17 types in this study (fig. 3B; supplementary fig. S3, Supplementary Material online). 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

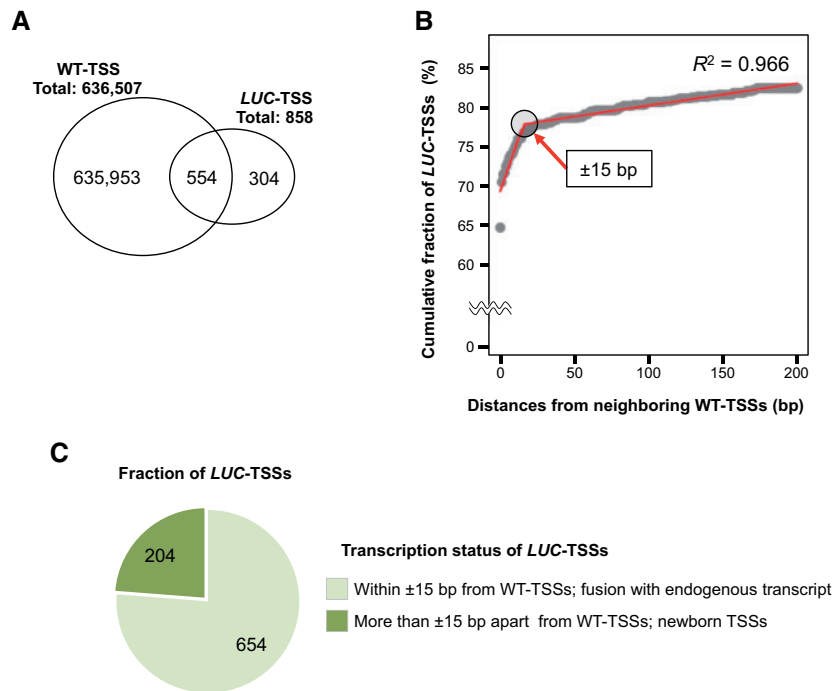


Fig. 2. Categorization of *LUC*-TSSs with respect to WT-TSSs. (A) Venn diagram summarizing the overlap between the positions of WT-TSSs and *LUC*-TSSs at single-nucleotide resolution. (B) The gray dots show a cumulative fraction of *LUC*-TSSs according to the distances from their nearest WT-TSSs. The red line indicates the linear approximation of the gray 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).

transcription units (fig. 3B; supplementary fig. S3, Supplementary Material online). We found that transposable elements were also sources of transcriptional activation (supplementary fig. S3, Supplementary Material online).

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,” $A-\alpha-2$ type in fig. 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” ($C_1-\beta_1-1$ type in fig. 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” ($A-\alpha-1$ type in fig. 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 ($C_1-\gamma_1-2$ type in fig. 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 (fig. 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 (Yamamoto et al. 2009; Haber and Stark 2018; Andersson and Sandelin 2020). We confirmed that the *A. thaliana* protein-coding genes utilized the same initiation dinucleotide motif based on the TSS-seq

of WT cells (fig. 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 (fig. 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 (fig. 4A–E, left panels). In addition to the AT-rich region described above, we were unable to find any characteristic motifs commonly found in 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, that is, variant of histone H2A (H2A.Z) and lysine (K) trimethylation of histone H3 (H3K36me3) as active transcription marks and lysine dimethylation 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 (fig. 4A, right panel). The epigenetic landscapes of the

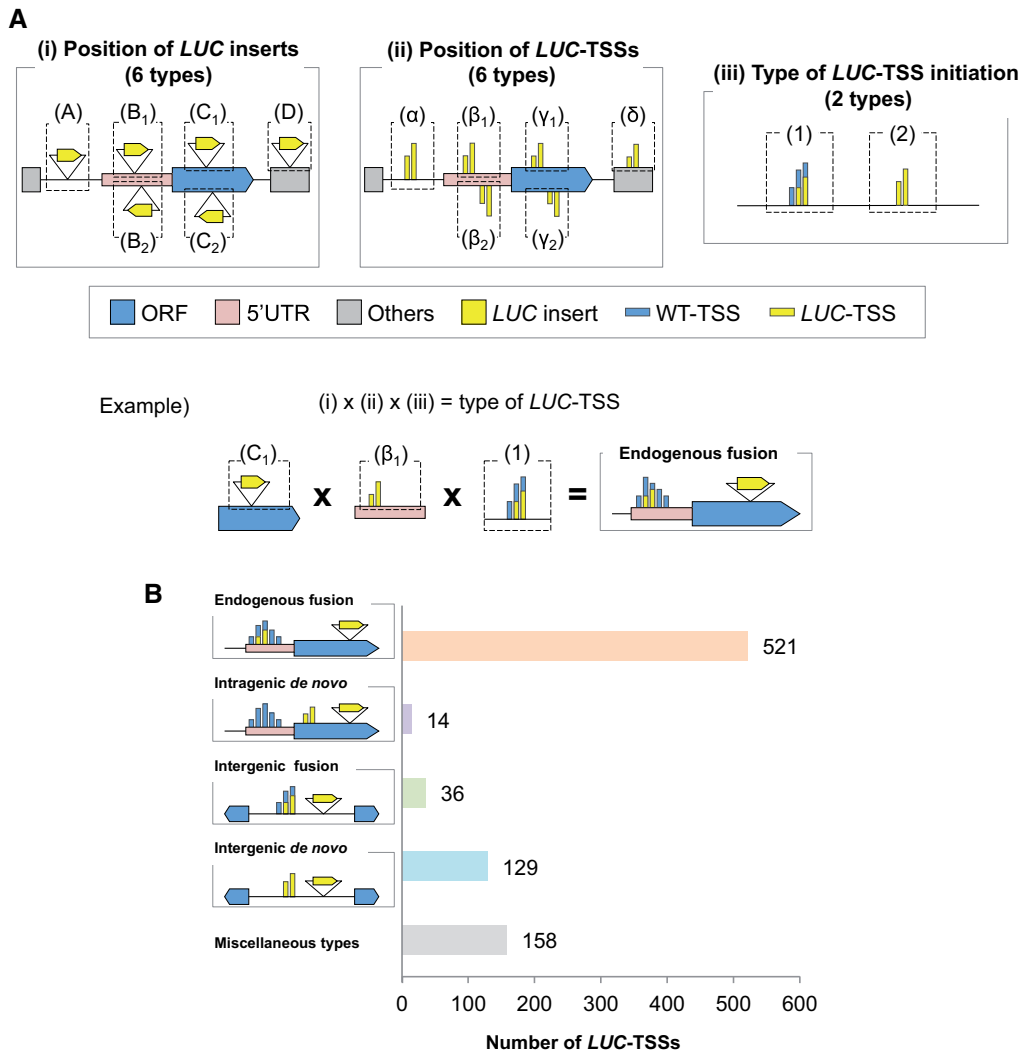


FIG. 3. Classification of *LUC*-TSSs according to the combination of their genomic loci and types of TSS initiation. (A) *LUC*-TSSs were classified according to the combination of the position of (i) the *LUC* insert, (ii) the *LUC*-TSS relative to *Arabidopsis thaliana* annotated genes, and (iii) the types of *LUC*-TSS initiation, as categorized in figure 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 supplementary figure S3, Supplementary Material online.

“Endogenous fusion” type around its TSSs were similar to those of WT-TSSs (fig. 4A, 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 (fig. 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 (supplementary fig. S4, Supplementary Material online). We also found promoter-specific epigenetic patterns in the “Intergenic fusion” type, indicating that unannotated WT transcription was trapped in this case (fig. 4D, right panel). In contrast with these observations, no significant epigenetic patterns were detected around “Intergenic *de novo*” TSS loci (fig. 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 (fig. 5A). Next, we examined the *de novo* TSSs. Surprisingly, “Intergenic *de novo*” TSSs initiated

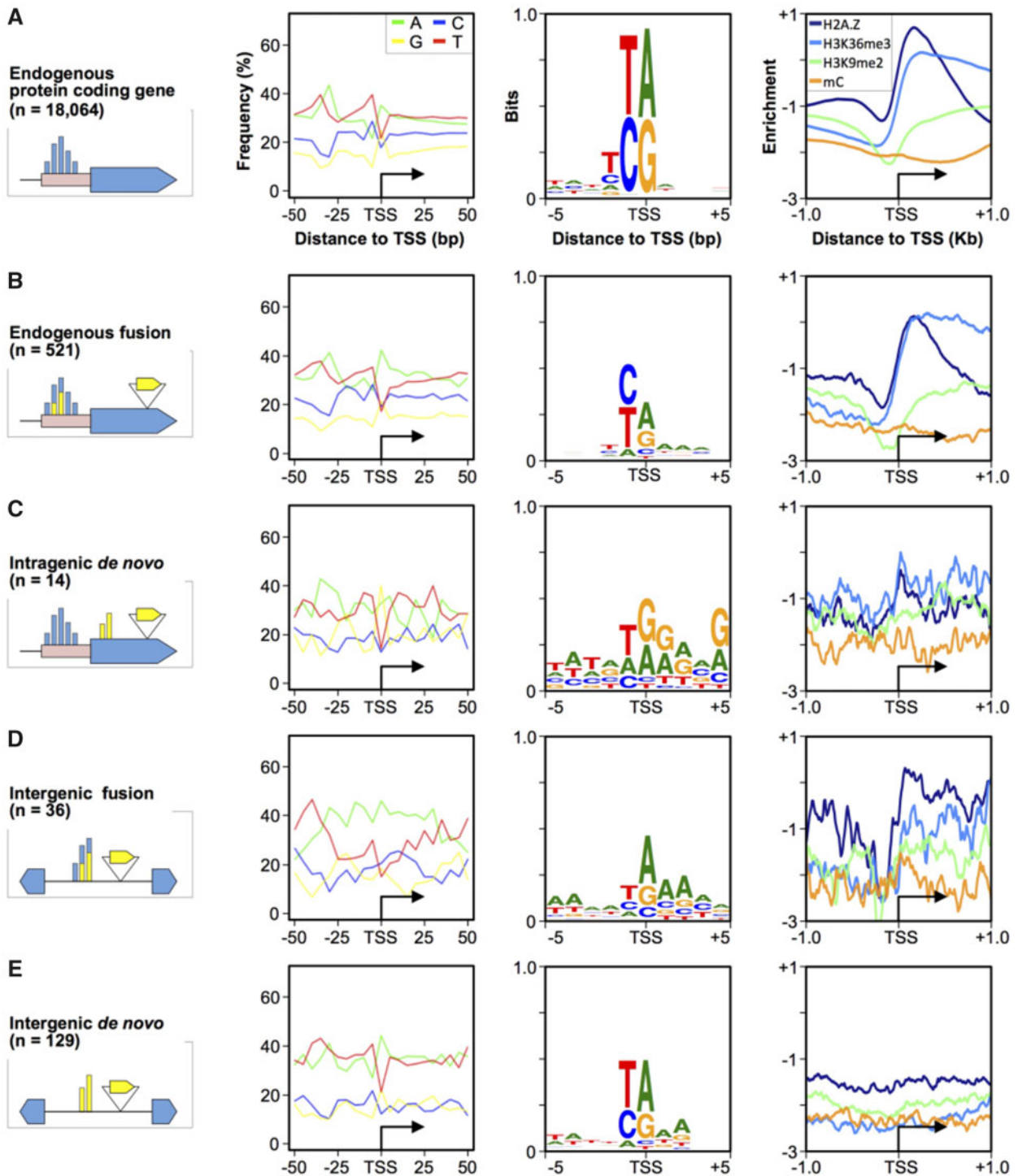


FIG. 4. Sequence and epigenetic characteristics of the *LUC*-TSSs. (Left panels) Nucleotide frequency at 5 nt resolution centered on the TSSs of (A) endogenous protein-coding genes ($n = 18,064$) and *LUC*-TSSs classified as (B) “Endogenous fusion” type ($n = 521$), (C) “Intragenic de novo” type ($n = 14$), (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.

predominantly in the close vicinity of *LUC* insertion sites (median distance, 108 bp) (fig. 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 (fig. 5B; supplementary fig. S5, Supplementary Material online). We confirmed these distribution profiles in three different biological samples (supplementary

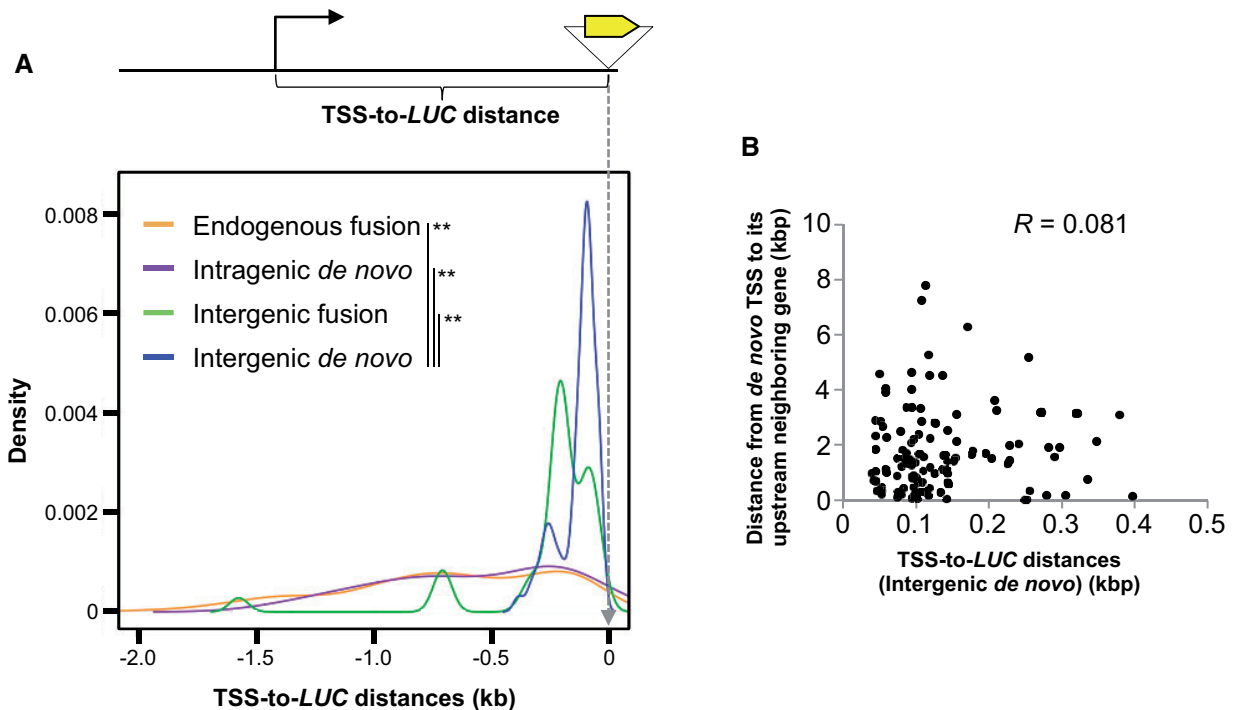


Fig. 5. De novo TSSs occur in the 5' proximity of the *LUC* inserts. (A) Density plot showing the distribution of the distance between the TSS and *LUC* insertion site (TSS-to-*LUC* distance) in each *LUC*-TSS type. The Kernel density estimation plot was generated using “density” R package. The median 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 neighboring gene distance. *R*, Pearson’s product-moment correlation test.

fig. S6, Supplementary Material online). 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 (fig. 6A; supplementary fig. S7, Supplementary Material online). This characteristic was similar to the 5'-untranslated region (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 (fig. 6B; supplementary fig. S8A, Supplementary Material online). 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) (supplementary fig. S8B, Supplementary Material online). 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 *LUC* activity. However, we found that Kozak-containing ORFs exhibited an unusual distribution around de novo TSSs: these two entities were mutually exclusive (fig. 6C and D). As shown in figure 6C, de novo TSSs did not occur within Kozak-containing ORFs (fig. 6C, middle panel; supplementary fig. S8C, Supplementary Material online), while ORFs without Kozak sequences were uniformly distributed around de novo TSS loci as well as in randomly sampled intergenic regions (fig. 6D, left and middle panels). These distribution patterns were commonly observed among three distinct biological replicates (supplementary fig. S8D, Supplementary Material online). 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 (fig. 6C, 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.

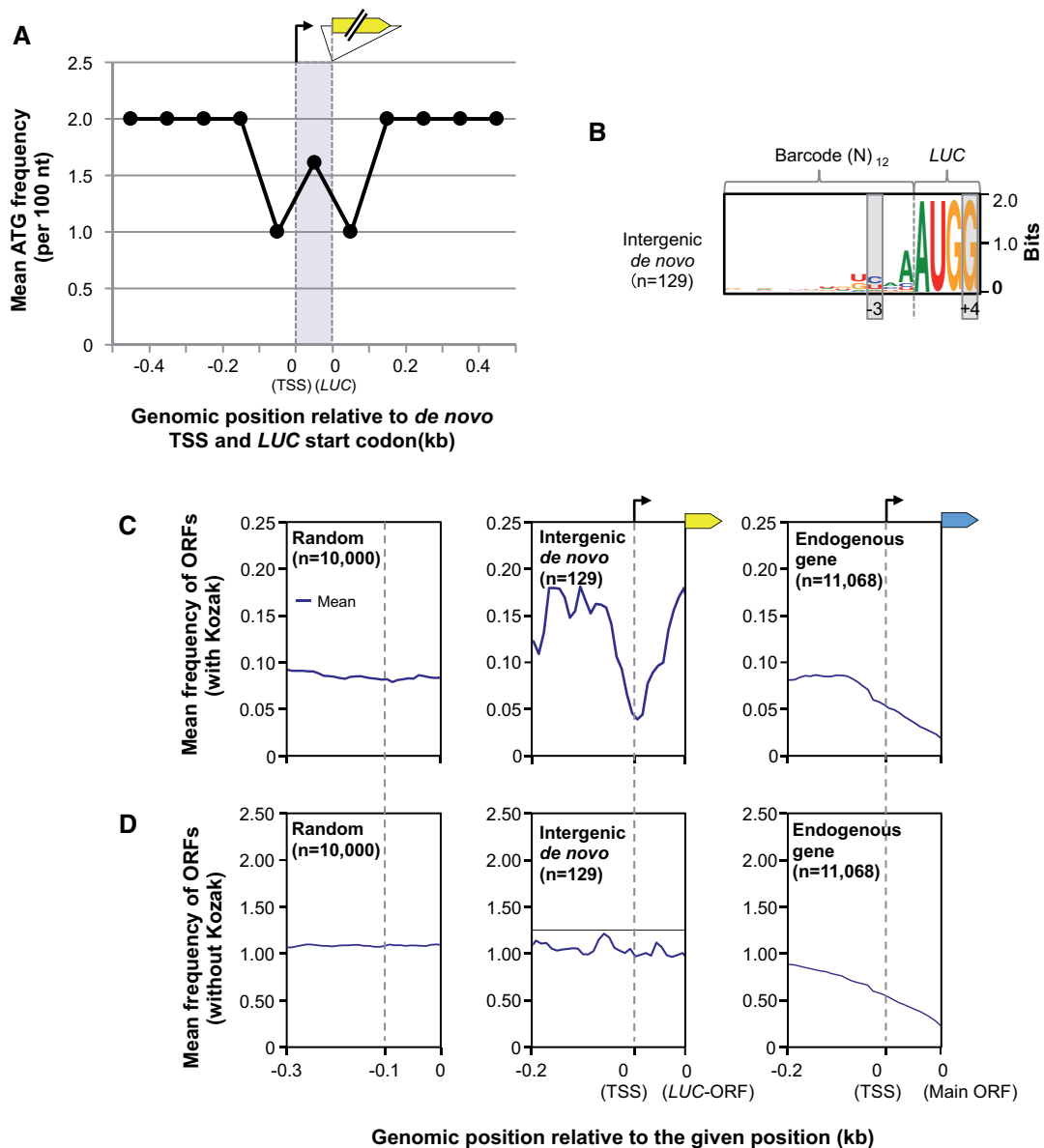


FIG. 6. De novo TSSs avoid pre-existing Kozak-containing ORFs. (A) Mean frequency of the initiation codon (ATG) per 100 bp around de novo TSS regions. The ATG frequency in the de novo TSS regions was normalized per 100 bp. (B) Sequence logo of the barcode region on the Intergenic de novo-type LUC inserts ($n = 129$). The conserved positions of a minimum Kozak motif (A/GNNAUGG) are indicated by the gray 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 gray dotted lines indicate the TSS positions.

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 (fig. 3; supplementary fig. S3, Supplementary Material online), 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 *LUC* activity (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 (fig. 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 (fig. 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 et al. 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 (supplementary methods S1, Supplementary Material online). Overcoming this experimental drawback needs further technical improvements.

As de novo TSSs occur without inherent promoter-like epigenetic profiles (fig. 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 remodeling around the *LUC* insertion loci (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 remodeling (Satoh et al. 2020). Taken together, these findings suggest that gene insertion events stochastically activate local chromatin remodeling 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 (fig. 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 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 (fig. 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 (fig. 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 et al. 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 (fig. 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 (fig. 1C) (Satoh et al. 2020). The new TSSs do not occur within the pre-existing Kozak-containing ORFs to avoid interference with the pre-existing genetic information (fig. 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 fig. 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 et al. 2018; Werner et al. 2018; Durand et al. 2019; Zhang 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 gene origination and evolutionary studies in the plant genome. Because the current study was performed using cultured cells, the genetic behavior of de novo transcription requires further examination regarding heredity and functional adaptation with/without selective pressures.

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 ($50\text{--}70 \mu\text{E m}^{-2} \text{s}^{-1}$). 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.

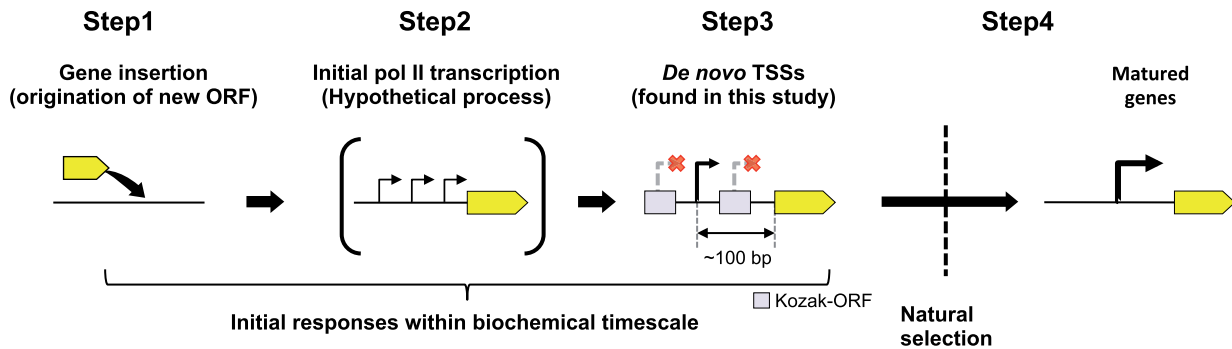


Fig. 7. Model of the evolutionary processes of new genes. Brand-new coding sequences are originated/introduced by genome shuffling or the EGT/HGT process. De novo TSSs occur in response to the origination of a new coding sequence, with satisfying an anti-Kozak rule. De novo TSSs are originated within biochemical timescale, independently of the functionality of the messages. After de novo TSS occurrence, the neighboring putative ORFs are eliminated via function-based natural selection in the evolutionary timescale.

T87 WT TSS-Seq Library Preparation

All primers used in this study are listed in [supplementary table S1, Supplementary Material](#) online. 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 μ g) was reverse transcribed using 1,000 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/, last accessed February 16, 2016). The first nucleotide of the forward reads was added by the library preparation step, and the second nucleotide was attributed to nontemplated 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. (2009). Processed paired reads were mapped to the TAIR10 release of the *A. thaliana* genome assembly (<https://www.arabidopsis.org/>, last accessed April 15, 2013) 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 (Sato 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 μ g of an anti-H2A.Z rabbit polyclonal antibody (Kudo et al. 2020) and 1.0 μ g 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. (2007) for H2A.Z, and to Yang et al. (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 μ g) 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. (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/, last accessed February 16, 2016). 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/>, last accessed January 4, 2018).

LUC-TSS-Seq Library Preparation

Transgenic T87 cells harboring promoterless *LUC* genes were established previously (Sato 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 (fig. 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 [supplementary methods S1, Supplementary Material](#) online.

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 ([supplementary methods S1 and supplementary fig. S1, Supplementary Material](#) online). 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 [supplementary methods S1 and supplementary figure S2, Supplementary Material](#) online.

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>, last accessed July 25, 2019) 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'-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>, last accessed March 14, 2018). A meta-gene 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 analyzed 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).

Supplementary Material

[Supplementary data](#) are available at *Molecular Biology and Evolution* online.

Acknowledgments

This study was supported by Japan Society for the Promotion of Science (JSPS) KAKENHI (grant numbers 23125512, 23657040, 23117006 and 25650131 to J.O., 26660008 to S.S., and 17J04887 to T.H.), and by the Strategic Research Funds at Kyoto Prefectural University.

Data Availability

Next-generation sequencing data of *Arabidopsis* T87 cells are available in the DDBJ Sequence Read Archive (accession numbers DRX190483–DRX190494).

References

- Andersson R, 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, 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 TL, Machanick P. 2012. Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Res.* 40(17):e128.
- Barakat A, Matassi G, 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 USA.* 95(17):10044–10049.
- Cardoso-Moreira M, Long M. 2012. The origin and evolution of new genes. *Methods Mol Biol.* 856:161–186.
- Chang YF, Imam JS, Wilkinson MF. 2007. The nonsense-mediated decay RNA surveillance pathway. *Annu Rev Biochem.* 76:51–74.
- Deal RB, Topp CN, McKinney AC, Meagher RB. 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 CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29(1):15–21.
- Durand É, Gagnon-Arsenault I, Hallin J, Hatin I, Dubé AK, Nielly-Thibault L, Namy O, Landry CR. 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 RM, Souza AL, Clish CB, Gehring M. 2014. 5-hydroxymethylcytosine is not present in appreciable quantities in *Arabidopsis* DNA. *G3 (Bethesda).* 5(1):1–8.
- Friedrich G, 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. Berkeley (CA): University of California Press.
- Gibney ER, Nolan CM. 2010. Epigenetics and gene expression. *Heredity (Edinb).* 105(1):4–13.
- Haberle V, 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, Obokata J. 2020, unpublished data. *De novo* activated transcription of newborn coding sequences is inheritable in the plant genome. *BioRxiv* doi.org/10.1101/2020.11.28.402032.
- Inoue F, 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.
- Kim BH, Cai X, Vaughn JN, von Arnim AG. 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 YY, Hata T, Kimura H, Matsui M, Obokata J. 2020, unpublished data. Cryptic promoter activation occurs by at least two different mechanisms in the *Arabidopsis* genome. *BioRxiv* doi.org/10.1101/2020.11.28.399337v1.
- Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods.* 9(4):357–359.
- Lauria M, Rossi V. 2011. Epigenetic control of gene regulation in plants. *Biochim Biophys Acta.* 1809(8):369–378.
- Li C, Lenhard B, Luscombe NM. 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, Subgroup GPPD. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078–2079.
- Marquez Y, Brown JW, Simpson C, Barta A, Kalyna M. 2012. Transcriptome survey reveals increased complexity of the alternative splicing landscape in *Arabidopsis*. *Genome Res.* 22(6):1184–1195.
- McLysaght A, 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, Itoh M. 2014. Detecting expressed genes using CAGE. *Methods Mol Biol.* 1164:67–85.
- Nakagawa S, Niimura Y, Gojobori T, Tanaka H, 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, 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, Marquardt S. 2019. Transcription-driven chromatin repression of intragenic transcription start sites. *PLoS Genet.* 15(2):e1007969.
- O'Malley RC, Huang SC, Song L, Lewsey MG, Bartlett A, Nery JR, Galli M, Gallavotti A, Ecker JR. 2016. Cistrome and episcistrome 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, 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 AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26(6):841–842.
- Ramírez F, Dündar F, Diehl S, Grüning BA, 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, Obokata J. 2020, unpublished data. Plant genome response to incoming coding sequences: stochastic transcriptional activation independent of integration loci. *BioRxiv* doi.org/10.1101/2020.11.28.401992.
- Smith JE, Baker KE. 2015. Nonsense-mediated RNA decay—a switch and dial for regulating gene expression. *Bioessays* 37(6):612–623.
- Springer PS. 2000. Gene traps: tools for plant development and genomics. *Plant Cell.* 12(7):1007–1020.
- Stegemann S, Bock R. 2006. Experimental reconstruction of functional gene transfer from the tobacco plastid genome to the nucleus. *Plant Cell.* 18(11):2869–2878.
- Takahashi H, Lassmann T, Murata M, 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 SB, Carvunis AR. 2019. De novo gene birth. *PLoS Genet.* 15(5):e1008160.
- Wang D, Qu Z, Adelson DL, Zhu JK, Timmis JN. 2014. Transcription of nuclear organellar DNA in a model plant system. *Genome Biol Evol.* 6(6):1327–1334.
- Werner MS, Sieriebriennikov B, Prabh N, Loschko T, Lanz C, Sommer RJ. 2018. Young genes have distinct gene structure,

- epigenetic profiles, and transcriptional regulation. *Genome Res.* 28(11):1675–1687.
- Yamamoto YY, Yoshitsugu T, Sakurai T, Seki M, Shinozaki K, Obokata J. 2009. Heterogeneity of Arabidopsis core promoters revealed by high-density TSS analysis. *Plant J.* 60(2):350–362.
- Yang H, Howard M, 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 AR, Yu Y, Hou G, Zi J, Zhou R, et al. 2019. Rapid evolution of protein diversity by de novo origination in *Oryza*. *Nat Ecol Evol.* 3(4):679–690.
- Zuo YC, Li QZ. 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.