Genome-wide Studies on Structure of Arabidopsis Promoters and Application to a Promoter of Aluminum-activated Malate Transporter 1

Genome-wide Studies on Structure of Arabidopsis Promoters and Application to a Promoter of *Aluminum-activated Malate Transpoter1*

(ゲノム科学的手法によるシロイヌナズナプロモーター構造理解と酸性土壌に関わるリンゴ酸トランスポーター***ALMT1*** 遺伝子プロモーターへの適用**)**

2016

The United Graduate School of Agriculture, Gifu University

Science of Biological Resources

(Gifu University)

TOKIZAWA, Mutsutomo

Genome-wide Studies on Structure of Arabidopsis Promoters and Application to a Promoter of *Aluminum-activated Malate Transpoter1*

(ゲノム科学的手法によるシロイヌナズナプロモーター構造理解と酸性土壌に関わるリンゴ酸トランスポーター***ALMT1*** 遺伝子プロモーターへの適用**)**

TOKIZAWA, Mutsutomo

Contents

Chapter1

Genome-wide identification of transcription start sites in *Arabidopsis thaliana*

1-1. INTRODUCTION

Any gene requires a promoter that determines position, direction, frequency and timing of transcription. Recent studies on RNA polymerase II (pol II)-dependent promoters revealed there are a few types of core promoter elements that determine the position and direction of transcription. In mammals, Pol II-dependent promoters are divided into two types according to core promoter elements. One is the TATA-type promoter containing the TATA box, which have a sharp peak of transcription start sites (TSSs) (Suzuki et al., 2001, Carninci et al., 2006) and are rich in genes showing high (Moshonov et al., 2008) and tissue-specific expression (Schug et al., 2005). The other is the CpG type, containing CpG islands, and these have broad shaped TSSs (Suzuki et al., 2001, Carninci et al., 2006) and are rich in genes showing ubiquitous expression (Schug et al., 2005). Coverage of the two types in the 30,969 mouse promoters are 10.4% for the TATA type and 51.1% for the CpG type (Taylor et al., 2006). There are some coreless promoters in mouse, but not much is known about them.

Higher plants share the TATA box as a core promoter element with mammals and also yeast (Suzuki et al., 2001, Basehoar et al., 2004). A previous genome-wide promoter analysis comparing higher plants with mammals revealed that plants do not have CpG-type promoters (Yamamoto et al., 2007), but have plant-specific core elements, Y Patch, GA and CA

elements (Yamamoto et al., 2007, Yamamoto et al., 2007, Yamamoto et al., 2009). Like mammals, higher plants have functionally differentiated promoters according to the core type. The TATA-type promoters are rich in genes showing environmental responses, and genes with ubiquitous expression tend to have coreless-type promoters (Yamamoto et al., 2007, Yamamoto et al., 2011). These studies have revealed conserved and plant-specific core promoter elements and their functional characteristics.

It is also reported that human genes with a TATA-type promoter have a more compact gene structure regarding the lengths of the mRNA and introns, and also the number of introns than ones with a TATA-less promoter (Moshonov et al., 2008). In plants, TATA-type promoters have a shorter length of 5' UTR, but longer promoters (Yamamoto et al., 2011). These studies indicate that core promoter types affect not only transcriptional characteristics but also the structure of the corresponding genes.

In spite of the knowledge mentioned above, our understanding of genic promoters in transcription of protein-coding genes in higher plants is still limited. For example, we do not know which element, or structure, is responsible for transcription of the coreless-type promoters. Transcriptional regulatory elements have not been fully identified in a plant genome, so the structure of promoters cannot be completely defined. As for genic promoters for non-coding RNAs, our knowledge is more limited (Xie et al., 2005). In addition, little is known about non-genic-type promoters, including intragenic, antisense, and orphan promoters, parts of which have been experimentally identified but not characterized (Yamamoto et al., 2009, Mejia-Guerra et al., 2015, Alkhateeb et al., 2016).

The position and direction of promoters can be experimentally identified by analysis of transcription start site (TSS) tags. Two reliable methodologies have been developed for TSS analysis: the oligo-cap method (Otsuki et al., 2005; Tsuchihara et al., 2009) and cap analysis of gene expression (CAGE) (Shiraki et al., 2003, Kodzius et al., 2006, Takahashi et al., 2012)

based on the Cap-Trapper method for preparation of so-called full-length cDNA. Large-scale random sequencing of TSS tags and mapping to the corresponding genome provides information about the position, direction and strength of transcription. These methods have contributed to deeper understanding of promoter structure, and also provide pivotal data to detect position-dependent promoter elements including all the core elements and some of the transcriptional regulatory elements (FitzGerald et al., 2004, Carninci et al., 2006, Ni et al., 2010). Deep TSS analysis also revealed that transcription of a gene is often provided by multiple promoters (Carninci et al., 2006, Yamamoto et al., 2009). However, it is not clear how they contribute to gene expression.

TSS analysis has also been applied to plant genomes and some trends in transcriptional characteristics have been revealed (Yamamoto et al., 2009, Morton et al., 2014, Mejia-Guerra et al., 2015). However, these studies provide limited coverage of genes in the corresponding genome due to the sequencing of a limited numbers of TSS tags and/or low coverage of tissue types. In addition, a limited number of TSS tags do not allow us to analyze promoters with low activities that are rich in non-genic promoters. Therefore, their characteristics have not been reported in higher plants.

As mentioned, high coverage of genes in a genome by the TSS is indispensable for promoter annotation. It also helps our methodology for prediction of transcriptional regulatory elements based on transcriptome data (Yamamoto et al., 2011), which requires a set of promoter sequences of 1 kb length aligned at the TSS. In this study, I carried out deep sequencing of TSS tags prepared from various tissues and physiological conditions of Arabidopsis. Parallel analysis of paired-end sequencing of TSS tags and careful utilization of Cap Signature (Yamamoto et al., 2009) helped me to establish reliable data sets which could remove artifacts of TSS tags. This data covers as many as 79.7% of protein coding genes, and discovers and characterizes non-genic-type promoters, including intragenic,

antisense, and orphan promoters that are not assigned to any gene model.

1-2. MATERIALS AND METHODS

Plant growth and RNA extraction

Several tissues, including roots, shoots from seedlings, flower inflorescences, and etiolated and light stress-treated seedlings of *Arabidopsis thaliana* (Col-0), were harvested for preparation of TSS tag libraries. Preparation of roots, shoots from seedlings, etiolated seedlings and flower inflorescence has been described previously (Yamamoto et al., 2009). Light stress-treated seedlings were prepared according to Kimura *et al.* (Kimura et al., 2001). The total RNA was individually extracted as described above (Tokizawa et al., 2015).

Preparation of TSS tags

CAGE libraries for paired-end sequencing were prepared from roots, etiolate seedlings, green and light stress-treated seedlings as described (Kodzius et al., 2006, Takahashi et al., 2012) with a few modifications. Fifty ug of total RNA were used for synthesis of first strand cDNA by a MMLV reverse transcriptase (Superscript II, Invitrogen/Thermo Fisher Scientific K.K., Yokohama) with 2.5 ug of modified random octamer-containing primers (5'-GTT-CAG-ACG-TGT-GCT-CTT-CCG-ATC-TNN-NNN-NNN-C-3'). Double strand cDNAs were amplified by PCR for 15 cycles (94 for 30 s, 56 for 30 s, 68 for 1 min) from the linker-ligated single strand cDNAs using KOD FX DNA polymerase (Toyobo, Osaka) with NEB index primers for Illumina sequencing (New England Biolabs Japan, Tokyo). Amplified DNA fragments with lengths from 100 to 500 bp were purified using agarose gel electrophoresis and spin columns (Wizard SV Gel and PCR Clean-Up System, Promega KK, Tokyo). DNA concentration and size distribution were checked by a

Bioanalyzer (Agilent Technologies, Hachioji, Japan). The prepared CAGE libraries were subjected to paired-end sequencing using a Genome Analyzer IIx (GA IIx, Illumina) at the Nara Advanced Institute of Technology.

Oligo-cap libraries were prepared according to Tsuchihara *et al*. (Tsuchihara et al., 2009) from 100 ug each of total RNA from leaves, roots, flower inflorescences, etiolated seedlings, green seedlings, and light stress-treated seedlings (Yamamoto et al., 2009), and subjected to single-end sequencing by GA IIx (Illumina) at the Institute of Medical Science of University of Tokyo.

Random sequencing and data processing

CAGE and oligo-cap TSS libraries were subjected to sequencing analysis of paired- and single-end reads, respectively, of 35 bp long, using GA IIx (Illumina). Totals of 241,906,027 and 81,481,461 raw reads were obtained from CAGE and oligo-cap libraries, respectively. Raw reads were filtered to remove sequences of low quality using Trimmomatic (ver. 0.33) (Bolger et al., 2014) and those with no cap-signature were removed manually. The remaining sequence tags were mapped to the Arabidopsis genome (Col-0, TAIR10, (Lamesch et al., 2012)) using BWA (Li and Durbin, 2009) and MAQ (Li et al., 2008), allowing unique mapping and a maximum of 2bp mismatches. Sequence reads that mapped to nuclear ribosomal DNA (rDNA), chloroplast DNA and mitochondria DNA were removed from the TSS libraries. The final numbers of TSS tags from CAGE and oligo-cap libraries were 26,817,002 and 33,172,231, respectively. Both TSS tag sets are C1 type and illustrated in Supplementary Fig. S1-2.

***In silico* analysis**

The 5' UTRs of the TAIR10 gene model were extended using data sets of the paired-end

CAGE analysis as shown in Fig. 1-1A. For each gene model, the most extended paired-end tag was selected from the CAGE data using home-made R scripts, and the extension process was repeated 20 times. When the extended 5' UTR, designated as the Maximum 5' UTR (Max 5' UTR), went into the next gene model, the closer end of the invading gene model was set as the 5' end of the Max 5' UTR so as not to overlap the two gene models. This rule facilitated classification of promoters into genic, intragenic, and antisense groups. When two gene models were nested, the larger one was selected for categorization of promoters.

The steps for clustering TSS tags are illustrated in Supplementary Fig. S1-1. I applied the following two rules: 1) TSS tags with a distance of more than 20 bp were divided into different clusters, and 2) secondary peaks (see Supplementary Fig. S1-1) with a distance of more than 100 bp were divided into different clusters.

TSS tags of C1 from single-end sequencing of oligo-cap libraries were filtered to prepare C2 tags which have a mismatch at the cap signature as illustrated in Fig. S1-2, and used to calculate the ratios of C2 tags shown in Fig. 1-2B and Supplementary Fig. S1-3.

TSS clusters mapped to the *Arabidopsis* genome were then categorized into genic, intragenic, antisense, and orphan groups according to relative positions of their peak TSS positions and the expanded gene models containing the Max 5' UTR as shown in Fig. 1-2A.

Detection of core elements for each promoter is done using previously identified octamers for the core elements (Yamamoto et al., 2007, Yamamoto et al., 2009). No mismatch was allowed for the detection. GO analysis was achieved based on information of TAIR10 Gene Ontology (https://www.arabidopsis.org). A list of Arabidopsis genes thought to be of cyanobacterial origin has been prepared by Martin *et al.* (Martin et al., 2002). A total of 797 genes speculated as being cyanobacterial in origin were included in the GO analysis. Statistical significance of results was judged by Fisher's exact test.

Single nucleotide polymorphisms (SNPs) in 80 Arabidopsis accessions with the reference

genome of Col-0 were detected according to the genome sequences determined by Cao *et al.* (Cao et al., 2011). Considering that coverage of genome sequencing for each accession is not complete, promoter sequences from -1,000 to +100 bp relative to the peak TSS whose SNP info is available for more than 40 accessions throughout the region were subjected to SNP analysis. 16,896 genic top promoters and 22,981 orphan promoters were subjected to the analysis. I did not consider the number of accessions showing SNP from the reference genome but the presence or absence of SNP among the analyzed population at each promoter position was used for calculation of the SNP ratio. Observed ratios were normalized with base composition as shown in Supplementary Fig. S1-12, and subjected to 21-bin smoothing as mentioned. The net generation rate for each base (Fig. 1-7D and E) was calculated by adding appropriate SNP ratios for generation and subtracting appropriate SNP ratios for disappearance as follows: A = CA+GA+TA-AC-AG-AT, C = AC+GC+TC-CA-CG-CT, G = AG+CG+TG-GA-GC-GT, T = AT+CT+GT-TA-TC-TG.

1-3. RESULTS

Determination of maximum 5’ untranslated region with paired-end sequencing data of TSS tags

In a previous study, experimentally determined promoters, that are TSS clusters, were associated with gene models based on the distance from a translation start site of each gene model, and promoters having the same orientation as the downstream neighboring gene model within 1 kb from its ATG were considered as ones belonging to the gene model (Yamamoto et al., 2009). In this study, I have introduced experimental association of promoters with gene models.

First, 5' untranslated regions (UTRs) of Arabidopsis gene models in TAIR10 have been extended using paired-end sequencing data of TSS tags to determine the maximum 5' UTR for each gene model, designated as Max 5' UTR. As shown in Fig. 1-1A, paired-end TSS tags were used for "walking" towards the 5' direction from an established gene model to extend the 5' UTR that had been determined based on EST and full-length cDNA information (Lamesch et al., 2012). This walking started from the 5' UTR of TAIR10, or if no 5' UTR information was available, from the CDS. The extended 5' UTR was designated as the Max 5' UTR. The paired-end sequencing data of 27 M TSS tags extended the 5' UTRs of 27,526 genes from TAIR10, and I was able to assign a Max 5' UTR to 29,090 genes, which corresponds to 87.3% of the Arabidopsis genes (Fig. 1-1B). The median length of the extension of the 5' UTR was 201 bp, and the resulting Max 5' UTRs have a median length of 319 bp as shown in the panel. Data of the paired-end sequencing of TSS tags were used only for determination of Max 5' UTR.

Determination of promoter positions and association with gene models

Next, I prepared six TSS tag libraries from leaves, roots, flowers, etiolated seedlings and also light stress-treated seedlings by the oligo-cap method (Yamashita et al., 2011), and subjected them to single-end sequencing. In this analysis, 33,172,231 (33 M) TSS tags containing Cap-Signature (Yamamoto et al., 2009) were successfully mapped to the Arabidopsis nuclear genome.

Mapped TSS tags were then subjected to clustering as illustrated in Supplementary Fig. S1-1 to give 324,461 TSS clusters. These clusters, *i.e*. promoters, have been classified into four categories according to the extended gene models containing Max 5' UTRs, namely genic, intragenic, antisense, and orphan (Fig. 2A). The numbers of classified promoters are 59,628, 193,208, 42,070, and 34,549 for genic, intragenic, antisense and orphan categories, respectively (Fig. 1-2B). Of them, genic promoters had the highest activity, which is on average 465.54 tags/promoter, while the other categories showed much less activity (11.15 to 75.65 tags/promoter, Fig. 1-2B).

Association of promoters to gene models

The identified 59,628 genic promoters cover 22,211 genes, and they include 21,672 protein-coding genes which correspond to as much as 79.7% of the Arabidopsis protein-coding genes. This coverage is considerably higher than the 35.4% (9,627 genes) reported in a previous study (Yamamoto et al., 2009) and the 64.8% (17,619 protein-coding genes) in Morton *et al*. (Morton et al., 2014). Much lower coverage was observed for the ones for non-coding RNA genes (10.7% for miRNA and 41.8% for ncRNA) than for the genic promoters for protein-coding genes (Fig. 1-2B). This low coverage may be due to the involvement of transcription by non-pol II or the lower stability of the unprocessed transcripts with a cap. Coverage for pseudogenes is also low (16.6%), and this is a reflection of the large portion of

non-transcribed genes in this category.

In the Arabidopsis genome, 14,168 promoters have been reported as a sum of intragenic, antisense, and orphan groups (Yamamoto et al., 2009). The deeper sequencing in this study has allowed me to detect 193,208 intragenic promoters, 42,070 antisense promoters, and 34,549 orphan promoters (Fig. 1-2B). Detection of a large number of intragenic promoters, which make up 59.5% of the total promoters identified in this study, is also reported in a maize TSS study (around 50% for exon promoters, (Mejia-Guerra et al., 2015)). Expression levels of antisense promoters are one order lower than the genic promoters (tag/promoter, Fig. 1-2B).

Evaluation of tag data

All of the TSS tags contain Cap-Signature, which is an artificial addition of C at the 3' end of the (-) strand of cDNA (G at the 5' end of the (+) strand) in a cap-dependent manner by a MMLV reverse transcriptase (Potter et al., 2003, Yamamoto et al., 2009). I call these tags with Cap-Signature C1 tags. A problem with C1 tags occurs when the genomic sequence at -1 relative to TSS is G, because in this situation it is impossible to distinguish the first G in the TSS tag coming from the cap-dependent reverse transcription from the one derived from transcript sequence. The latter situation can include undesirable artifacts from non-capped RNA species, possibly processed or digested fragments of transcripts.

I prepared a C2 tag set from the C1 tags by filtering out all the tags whose genomic sequence corresponding to Cap-Signature is G (Supplementary Fig. S1-2). The C2 tag set is more reliable than the C1 tags, but does have a blind spot - excluding all the TSSs with G at the -1 position in the genome. Therefore, C2 tags alone are insufficient for promoter analysis, but good for evaluation of C1 tag clusters. Using the C2 tag set, I evaluated the promoter categories.

Fig. 1-2B shows the results of the C2 assessment of genic, intragenic, antisense, and orphan promoter fractions. In this assessment, a C1 cluster was judged as C2 tag-supported, if the coverage by C2 tags is higher than 50%. All of the categories of genic promoters including protein coding, miRNA, ncRNA and pseudogene are supported by the C2 tag set more than 80%. Intragenic and orphan promoters also showed high coverage (over 50%) by C2 tags. In contrast, antisense promoters are covered for 42.1%, which is considerably lower than the genic, intragenic, and orphan promoters (Fig. 1-2B and Supplementary Fig. S1-3). As these results suggest that the antisense promoters defined by C1 tags potentially include a considerable number of artifacts, I decided to do a comparative analysis of antisense and other promoter categories using only C2-supported promoters as shown in Fig. S1-3.

I also evaluated the TSS data by comparison with a previous study by the Cap-Trapper method (Yamamoto et al., 2009). In this analysis, I looked at about 200-fold more TSS tags than in the previous one. As shown in Supplementary Fig. S1-4, highly expressed promoters have no or a small change in the peak TSS position, and 62.8% have a shift of less than 21 bp in the peak TSS. Low expression promoters showed longer shifts in the peak TSS, and I interpret the results as indicating more accurate detection of the peak TSS by deeper sequencing, which is evident in low expression promoters. Assuming this interpretation, the comparison demonstrates good consistency of the new TSS data with the previous analysis.

Comparison of features between genic promoters and other categories

Several core promoter elements have been found in plant promoters, including the TATA Box, Y Patch, GA and CA elements (Yamamoto et al., 2007, Yamamoto et al., 2009). Of these, Y Patch, GA and CA elements are thought to be plant specific. Yamamoto *et al.* showed that there are positive correlations between the expression level and the ratio of core elements, including the TATA box, Y Patch, GA, and Inr (Yamamoto et al., 2009). Moreover, they also

reported that the promoter shape is determined by the core promoter type (TATA and Y Patch; sharp type, GA and Coreless; broad type). I have confirmed these promoter characteristics based on the newly generated TSS data (Supplementary Fig. S1-5).

In contrast to the well-studied genic type of promoters, little is known about non-genic types, including intragenic, antisense, and orphan promoters, especially in plants. To characterize these non-genic types of promoters, I compared trends of core promoters and TSS types in these promoters with genic ones. In this analysis, I selected only C2-supported promoters for all the analyzed categories (Supplementary Fig. S1-3 B). Antisense promoters are much less supported by C2 tags (Fig. 1-2B) and thus a considerable portion of this category has the potential to contain artifacts in the C1 population.

The ratio of promoter categories among promoter fractions according to expression level, measured by tag number per promoter, is shown in Fig. 1-3A. The ratio of genic promoters is very low (0.11) in the lowest expression fraction (1 tag), and increases in accordance with an elevation in the expression level. The highest ratio of genic promoters is 0.94 in the highest fraction of expression (≥4097 tags), demonstrating that almost all of the highly expressed promoters are in the genic category. In contrast, non-genic promoters including intragenic, antisense and orphan, show the opposite tendency, and their promoter ratios decrease with an elevation in expression level. In the case of orphan promoters, I noticed highly expressed promoters in this category locate proximal to rDNA genes whose expression is extremely high, so I assume that these transcriptional activities are influenced by the very strong rRNA promoters and thus their activation is due to exceptional situations. When these promoters, located within 4 kb of rDNA, were removed from the analysis (dotted gray line in Fig. 1-3A), the observed trend of the orphan promoters became similar to intragenic and antisense ones. These results reveal contrasting features between genic and non-genic promoters.

I then looked into core elements of the promoter categories. Panel B in Fig. 1-3 shows

the ratio of promoters containing each core element of the promoter categories. Coreless means the promoters have neither TATA box, Y Patch, CA nor GA elements. The genic category was further classified into genic-protein coding and genic-miRNA.

Comparing with genic-protein coding promoters, promoters for genic-miRNA have a considerably higher ratio of TATA and a lower ratio of CA element and Coreless. An example of the structure of a promoter for a miRNA and a list of identified promoters for miRNA genes are shown in Supplementary Fig. S1-6. In contrast, intragenic, antisense and orphan promoters show lower ratios of TATA, Y Patch, GA and CA elements and higher ratios of Coreless. This analysis reveals that the composition of core elements is different among these three groups: genic-protein coding, genic-miRNA, and non-genic intragenic, antisense and orphan. Interestingly, the last group also shares the same characteristics of low expression as shown in Fig. 1-3A, suggesting that these two results are related.

Supplementary Fig. S1-7 shows the composition of core elements for each promoter category fractionated with expression level. The results indicate that the trends of the core elements regarding expression level are essentially the same regardless of the promoter categories. Therefore, the trends of core elements, where TATA Box, Y Patch, GA and CA elements are rich in highly expressed promoters and Coreless promoters are rich in low expression fractions, are conserved among genic and non-genic promoters. The expression level of each promoter category is consistent with the composition of core elements, *e.g.*, highly expressed genic promoters are rich in the TATA type and poor in the Coreless type. In addition, possible differences in *in vivo* half-life of transcripts from different promoter categories would also contribute to the expression profiles as shown in Panel A, *e.g.*, genic transcripts which are mRNA would be more stable than other types of transcripts.

Fig. 1-3C shows analysis of local sequences around TSSs. Previous studies have identified a dinucleotide motif, YR Rule (+1 is underlined, Y: C or T, R: A or G), that is

applicable to plants (Yamamoto et al., 2007) and mammals (Carninci et al., 2006). I compared coverage of YR rule and also dinucleotide sequences of the rule among the genic and non-genic promoter categories. This analysis shows that the highest coverage is observed in genic promoters, which is higher than 80%. Antisense and orphan promoters also show high coverage, but intragenic promoters are covered by the motif with a very low ratio of 42%. When usage of all the dinucleotide sequences was examined at the TSS, intragenic promoters were revealed to have the least preference in the -1/+1 sequence (Supplementary Fig. S1-8).

The shapes of the TSS clusters are compared in Fig. 1-3D. The vertical axis of the graph indicates the peak ratio, which shows the relative amount of TSS tags of a peak TSS over the total TSS tags of the cluster. If all the TSS tags of a cluster come from one specific TSS, the peak ratio is 1.0 and the shape of the cluster is considered as "sharp". The panel indicates that genic, antisense and orphan promoters are a mixture of sharp and broad peak-TSS clusters, but the majority of intragenic promoters are broad type. These data are consistent with the results of Panel C, because the high peak ratio correlates with high coverage of YR rule (Yamamoto et al., 2009).

Low TATA ratio in promoters for chloroplast and mitochondrial proteins

TATA-type promoters have several functional features that are shared with plants and mammals: high expression, sharp-peak TSS clusters, and "regulated" rather than constitutive gene expression (Schug et al., 2005, Carninci et al., 2006, Yamamoto et al., 2009, Yamamoto et al., 2011). I have addressed the question as to which genes prefer the TATA-type promoter by subjecting genes driven by TATA-type promoters to Gene Onthology (GO) analysis.

Three types of GO classifications, "molecular function", "biological process" and "cellular components", were used and I found that classification by "cellular component" showed the highest variance (data not shown). As presented in Fig. 1-4A, the TATA ratio of

all the genes used for GO analysis was 0.32. Extremely high ratios were obtained for "cell wall" and "extracellular" genes. The reason for these higher ratios is not known, but one possible reason is that they are rich in regulated gene expression (Supplementary Fig. S1-9).

On the other hand, significantly lower ratios were obtained for "chloroplast", "mitochondria", and "plastid" genes. One possibility for the lower TATA ratio is that these GO categories contain genes with low expression levels, considering the characteristics of TATA-type promoters. However, expression levels of the GO categories as shown in Fig. 1-4B indicate that this is not the case because their expression level is comparable to the average (ALL) or is higher. Another possibility is these GO categories contain a smaller number of "regulated" genes, which also fits with the characteristics of TATA-type promoters. GO analysis of stress-responsive genes as shown in Supplementary Fig. S1-9 reveals that this possibility is also not viable. These analyses indicate that there must be some other reason for the low TATA ratio observed.

One common feature of "chloroplast", "mitochondria" and "plastid" genes is that they are rich in genes transferred from the corresponding organellar genomes (Smith, 2014). Therefore, I examined the TATA ratio of putative genes from the chloroplast genome, which are computationally identified as genes with a cyanobacterial origin ("Cyano origin") (Martin et al., 2002). This group is not limited to genes for "chloroplast", "mitochondria" and "plastid", as around 2/3 of the genes correspond to other GO categories, such as "other membranes", "other intracellular components" and "other cytoplasmic components", as shown in Supplementary Fig. S1-10. Therefore, the "Cyano origin" group is composed of considerably different genes from the "chloroplast" or "mitochondria" categories, but the TATA ratio of this group turns out to be lower than these two GO categories (Fig. 1-4A). These results suggest that genes with organellar origin are richer in TATA-less promoters than the global average.

Gene expression is predominantly determined by a single promoter

In mammals, a gene is transcribed from multiple genic promoters (Carninci et al., 2006, Ni et al., 2010). My results from Arabidopsis indicate that the number of promoters for a gene is 2.70 on average (58,551/21,672 for genic-protein coding, Fig. 1-2B). To understand how multiple genic promoters contribute to the expression of a gene, I hypothesize two possibilities; 1) multiple promoters equally contribute to gene expression and the sum of each promoter activity is important to determine the expression level of a gene, and 2) one specific promoter predominantly contributes to the expression of a gene, and the contribution of the other companion promoters is negligible.

To examine these two hypotheses, I calculated the ratio of the most active promoter in a gene to the total promoter activity for the gene expression, measured by the tag counts of each promoter (Fig. 1-5A). The most highly expressed promoter was selected from genic promoters for a gene as a "top" genic promoter, and the coverage of a top promoter over the total expression level of the gene of focus was calculated as the dominance of the top promoter in a gene. Results indicate that dominance is as high as 81.8% in genes with the least expression (31-60 tags/gene), and it increases in accordance with the elevation of the total expression level of a gene. The highest coverage obtained was 98.8% for the most highly expressed genes (7681-tags/genes). Fig. 1-5B shows the number of genic promoters for a gene regarding expression level, and as shown, the number of promoters for a gene was around three regardless of the expression level of a gene. Of these three, only one promoter contributes to most of the gene expression. These results also suggest that an increase in gene expression is solely supported by the top promoter. These conclusions are illustrated in Fig. 1-5C.

High expression correlates with short 5' UTR

The length of the 5' UTR is different among species. Maria-Guerra *et al*. reported that the median length of 5' UTR in Arabidopsis (112 bp) was significantly shorter than those in maize (154 bp), mouse (159 bp), human (171 bp) and also Drosophila (191 bp) (Mejia-Guerra et al., 2015). I calculated the lengths of the 5' UTRs for all the genic promoters and genic top promoters as the distance from the peak TSS to the translation start site (ATG) based on my quantitative TSS data. The median length of the 5' UTR for genic top promoters is 83 bp long (dashed line in Fig. 1-6A), shorter than the 112 bp reported by Maria-Guerra *et al*. The difference in the length is suggested to be due to the selection of promoters; I selected genic top promoters but Maria-Guerra and colleagues analyzed total genic promoters.

The length of the 5' UTR is reported to have a negative correlation with expression levels in plants (Yamamoto et al., 2011) and mammals (Rao et al., 2013). Analysis of the 5' UTR length and expression levels based on this TSS data, shown in Fig. 1-6A, confirmed the negative correlation within a 5' UTR length of 400 to 50 bp, but the correlation is reversed when the length is shorter than 50 bp. These features indicate an optimum length of a 5' UTR for the highest expression to be around 50 ~ 60 bp. The median lengths of top genic promoter and companion genic promoters are also shown as vertical dashed lines, and as indicated, the lengths are 143 bp and 83 bp, respectively. The shorter length of the 5' UTR of top promoters compared with those of companion promoters is consistent with the higher expression of top promoters.

The same analysis was applied to TATA and Coreless types of genic promoters (Fig. 1-6B and Supplementary Fig. S1-11). For both populations, the relationship between the length of the 5' UTR and expression level was conserved in terms of overall trends and peak positions. The comparative analysis suggests a common characteristic regardless the core promoter type. A difference between the two populations is the median length of the 5' UTR,

where the TATA top clusters have a median length of 70 bp, shorter than the 85 bp of the Coreless top clusters, and this is consistent with the higher expression of the TATA-type promoters than the Coreless type (Fig. 1-6B and Supplementary Fig. S1-11).

Trends of single nucleotide polymorphism in the promoter region

Finally, I surveyed trends of promoter mutations using the released genomes of 80 accessions which represent natural variations of Arabidopsis in Europe, North Africa, and West and Central Asia (Cao et al., 2011). Using the Columbia accession as a reference genome, genome sequences of the accessions were aligned, and information of single nucleotide polymorphisms (SNPs) was summarized according to promoter positions. In this analysis, the observed SNP ratio was normalized by base composition at the corresponding promoter position (Supplementary Fig. S1-12).

Taylor *et al*. reported that the mutation rate of TATA-type promoters in mammals is lower than the ones of the other types from ~-50 to -1,000 bp relative to the TSS, suggesting that the TATA-type promoters are more conserved and thus more mature than the other types of promoters (Taylor et al., 2006). Using the SNP data of Arabidopsis, I examined if the same tendency is also observed in a plant genome.

Fig. 1-7A shows the normalized SNP ratios for the TATA, GA and Coreless types of genic top promoters with an average of 16,896 promoters. Results show that the three types of promoters show similar overall profiles as the summed data (All), where SNP ratios decrease from -400 bp relative to the TSS to +100 bp. In the case of orphan promoters, such conservation was not observed at all, and the SNP ratio was rather flat from -1,000 to +100 bp (Supplementary Fig. S1-13). These results reveal contrasting behavior between genic and orphan promoters.

In the region from -1,000 to -400 bp, the TATA type has the highest SNP ratio of the three

promoter types, and the Coreless type has the lowest. The order is reversed around -300 bp, and in the downstream region from this point, the TATA type shows the lowest SNP ratio and the Coreless type the highest. Therefore, the highest conservation of the TATA type in Arabidopsis is observed in a narrower region from -200 to ~-50 bp than in mammals.

Considering the results in Fig. 1-7A, the two extreme promoter types, the TATA and Coreless, were further compared. I analyzed the base composition, the generation ratio of each base and individual SNP ratios according to the promoter position. Panels B and C in Fig. 1-7 show the base composition of the TATA and Coreless types, respectively. Both types have higher AT content than GC throughout the promoter region, consistent with the low GC content of the Arabidopsis genome (32.4 ~ 33.0% in the non-coding region, (The_Arabidopsis_Genome_Initiative, 2000)). Appearance rates of AT and GC pairs are even from -1,000 to -300 bp, which means there is no strand bias of base composition in the region.

In addition to the even region, the TATA type contains several uneven regions showing strand bias in addition to the TATA box and the TSS sites. They are, (a) a higher appearance rate of A over T from -200 to -50, (b) a higher rate of T over A from -50 to -30 bp, that is the upstream neighboring side of the TATA box, (c) a higher rate of C over G from -100 to +100, and (d) a higher appearance rate of A over T from +1 to +50 (Fig. 1-7B). These strand biases are all more clearly observed in the TATA type than the Coreless type (Fig. 1-7B and C). It should be noted that my selection of TATA-type promoters was rather strict (Yamamoto et al., 2009), so some promoters containing a weak TATA box or a TATA-like sequence at the position of the TATA box are judged as "TATA negative" and thus part of them can be classified as a Coreless type. (b), (c), and (d) appear in a report of an Arabidopsis study by Alexandrov *et al*. (Fig. 3, (Alexandrov et al., 2006)), but my data (Fig. 1-7B) demonstrates more drastic differences. The functional implication of (a) is not known. I suggest that (c)

and also (b) reflect the generation of Y Patch in the region, because sequence, location of bias and the preferential appearance in the TATA-type promoters fit with the characteristics of Y Patch (Yamamoto et al., 2009). The region for (d) and its function will be discussed later.

I then looked into types of SNP. Using all the SNP ratios, I summarized "net generation rates" for A, C, G, and T (Panels D and E in Fig. 1-7). They are calculated as a sum of corresponding SNP ratios for changing to a specific base after subtraction of corresponding SNP ratios for changing from the base. For example, the net generation rate for A is calculated as the following formula of SNP ratios: [CA + GA + TA - AC - AG - AT]. The right graphs in Panels D and E are close up for a region from -50 to +50. As shown in the panels, generation ratios for a pair of A and T, and also of C and G, are very close to each other indicating parity of generation rates between (+) and (-) strands. Generation rates for A and T are both positive and ones for C and G are negative throughout the promoter region, indicating that A/T are being generated and C/G are disappearing by SNPs for both the TATA and Coreless types of promoters. This trend is reflected in the current base composition of Arabidopsis promoters: ratios of A and T are around 2 times more than those of C and G for both types of promoters (Fig. 1-7B and C).

As for the strand bias in the base composition observed in Panel B, (b), (c) and (d) were found to be reflected in the strand bias in the net generation rates for the TATA type, which is less obvious in the Coreless type as in the case of the base composition (Fig. 1-7D and E). Therefore, these strand biases in the base composition are growing. I could not detect a corresponding generation pressure for the strand bias of the base composition for (a).

Subsequently, I looked at which SNP types contribute to these differences in the generation ratios. SNP ratios of each SNP type in the TATA and Coreless types are shown in Panels F to I. One obvious feature of both promoter types is that the ratios of a SNP pair for complementary strands are generally very close (CT-GA, AT-TA, AG-TC, AC-TG, CA-GT,

and CG-GC). This suggests that generation and fixation of SNPs essentially occur in a strand-independent manner.

Of all the SNP types shown in Panels F to I, the highest SNP ratios are the CT and GA pair for both promoter types. This pair of SNPs occurs by deamination of cytosine changing it to uracil, which is reported to happen at a high frequency (Collins and Jukes, 1994, Ossowski et al., 2010). My results indicate that this CT SNP happening on both DNA strands provides the highest contribution to the generation ratios for A and T as shown in Panels B and C, which is driving the AT-rich promoters of Arabidopsis (Fig. 1-7B and C). The CT-GA pair provides the biggest contribution to the reduction of total SNP rates at the proximal region of promoters for all types of promoters (-400 to +100, Panel A). This reduction is due to higher pressure for conservation in the upstream region (-1,000 to -400) in a strand-independent manner.

A careful look at the graphs in Panels F to I reveals parity of a SNP pair is broken at some local regions. A large break was found in the CG-GC pair around the TSS of the TATA type (-80 to +100 bp of the inserted graph, Panel H). This strand bias in the SNP ratio is reflected in the higher generation ratio for C over G (Panel D), and also in case (c) of strand bias in the base composition, which has a higher rate of C over G (Panel B) for the TATA type.

Another large break in the parity is the CA-GT pair for the TATA type. In the inserted graph of Panel H in Fig. 1-7, the GT ratio is higher than the CA ratio in the upstream region of the TSS, and in the downstream region, the relationship reverses. The strand break in the upstream region is consistent with a higher generation rate of T over A (Panel D), and also of higher T over A in the base composition in an upstream neighboring region to the TATA box ((b), Panel B). The strand bias of the CA ratio over GT, in a downstream region of the TSS is also consistent with a higher generation rate of A over T (Panel D), and higher A over T in the base composition ((d), Panel B).

As the downstream region of the TSS contains the coding region, I wanted to know the relative position of the observed bias from the ATG. TATA promoters were divided into three groups by the length of their 5' UTR, and SNP ratios for CA and GT were calculated again for each group. The results shown in Supplementary Fig. S1-14 indicate that the peak position of CA and strand bias for CA over GT shifted downstream with longer 5' UTRs. This suggests that the position of this strand bias is not determined by the TSS but by the ATG. Therefore, I re-analyzed the base composition, the net generation rate for each base, and the SNP ratios in accordance with the relative positions of the TATA-type promoters to the ATG (Fig. 1-8).

Results show a clear gap of strand bias for A over T in the base composition between the upstream and downstream regions of the ATG, and the highest bias was observed at the upstream neighbor of the ATG ((d'), Fig. 1-8A and B). This clear gap demonstrates a better focus than the results shown in Fig. 1-7B. They indicate that the strand bias observed downstream of the TSS in Fig. 1-7B locates on the upstream side of the ATG within the 5' UTR. This feature correlates with the higher generation rate for A over T in the region (Fig. 1-8C and D).

When corresponding SNPs for this strand bias were surveyed, as shown in Fig. 1-8E and F, the most contribution was observed in the CA-GT pair among the SNP pairs, but all the possible pairs for the generation of A/T (CT-GA, AT-TA, and CA-GT) and their disappearance (AG-TC and AC-TG) occur with a bias towards generation of the strand bias for A over T on the upstream side of the ATG within the 5' UTR. These biases for A are less obvious for the Coreless type of promoters (Supplementary Fig. S1-15), suggesting that these trends are specific to the TATA type. The functional significance of the generation of As in this region is not clear, but one possible hypothesis is that an increase of As in the 5' UTR leads to an elevation of translational efficiency, as demonstrated in the functional studies of Arabidopsis and yeast 5' UTRs (Kawaguchi and Bailey-Serres, 2005, Dvir et al., 2013).

1-4.DISCUSSION

Comprehensive identification of Arabidopsis promoters by deep TSS sequencing

This analysis provides substantial sequencing data of Arabidopsis TSS tags that corresponds to a 175-fold increase of that in a previous report (Yamamoto et al., 2009) and an 8-fold increase of that in the report by Morton *et al*. (Morton et al., 2014). This deeper analysis using tag libraries from various tissues enabled the discovery of a much larger number of promoters in the Arabidopsis genome, leading to high coverage of genic promoters for protein-coding genes (21,672/27,206 = 79.7% in this study, compared with 35.4% by Yamamoto *et al.*, and 64.8% by Morton *et al*.), and a more exact estimation of promoter numbers per gene (~13 promoters and 2.7 genic promoters per gene) (Fig. 1-2). The determination of genic promoters is based on my extended gene models using independent paired-end analysis of TSS tags, so the association is experimentally validated and reliable. Another feature of the high reliability of the promoter analysis is the reference to C2 tag information when necessary. This approach helped me avoid misleading characterization of the antisense type of promoters.

Promoter types and their characteristics

Genic promoters for miRNA Many miRNAs are transcribed by RNA pol-II in mammals (Bracht et al., 2004, Cai et al., 2004, Lee et al., 2004) and also in plants (Xie et al., 2005). However, the coverage ratio of genes for miRNA obtained in this study is significantly less (10.7%) than that of protein-coding genes. One possible reason for the low coverage ratio is that the stability of unprocessed transcripts with a cap for miRNA is low. Another possibility is that miRNA is generated not only by RNA pol-II but that another RNA polymerase, such

as pol-IV (Onodera et al., 2005), also contributes. The higher TATA ratio of the identified promoters for miRNA (Fig. 1-3B) is consistent with a previous report (Xie et al., 2005), and this might reflect many stress-responsive genes for miRNA (Sunkar et al., 2012), because the TATA-type promoters are enriched with "regulated" genes rather than "constitutive" ones (Yamamoto et al., 2011).

Intragenic promoters Intragenic promoters have a lower ratio of core promoter elements. In particular the TATA ratio of intragenic promoters (3.8%) is considerably lower than that of genic promoters (19.5%) (Fig. 1-3B). In addition, the peak TSS in the intragenic promoters shows a poorer fit with the YR rule (42%) than the peak TSS for the genic promoters (84%), and the preference for dinucleotide sequence at the -1/+1 is less than other promoters (Supplementary Fig. S1-8). The unique features of intragenic promoters that were observed may suggest a different mechanism of the promoters for transcriptional initiation from genic promoters, which may be enabled by an "open" state of intragenic regions regarding chromatin structure and of the double strand DNA.

Antisense promoters Judging from the C2 tag analysis (Fig. 1-2B), antisense promoters possibly contain the highest contamination of uncapped RNA from the preparation of TSS tags. This might be due to strong hybridization with mRNA during the process of TSS tags preparation. Antisense promoters also have lower TATA ratios than genic promoters (Fig. 1-3B). Low TATA ratios of antisense promoters are also reported in mammalian studies (Orekhova and Rubtsov, 2013, Lin et al., 2016). In addition, I found that the shape of promoters as determined by the distribution of TSSs is the sharpest among all the promoter categories (Fig. 1-3D). In genic promoters, the sharp shape correlates with a high TATA ratio (Yamamoto et al., 2009), but antisense promoters have low TATA ratios, so this correlation breaks down. It is not clear what factor determines the sharp shape of promoters without the TATA box in the antisense promoters.

Orphan promoters These are promoters which could not be associated with any gene models, and are thought to be a mixture of promoters for unidentified genes and also for so-called transcriptional noise (Hüttenhofer et al., 2005) when complete shut-off of transcriptional activity at all the unnecessary genomic regions is difficult. To date, very little characterization of this type of promoter has been done. My initial attempts to understand orphan promoters have revealed that they have lower expression levels (Fig. 1-3A), lower TATA, Y, GA, and REG ratios and a higher ratio of Coreless types (Fig. 1-3B), and also less conservation of promoter sequences from -400 to +100 bp relative to the TSS (Supplementary Fig. S1-13). Further to the last finding, I did not detect any selection pressure on orphan promoter sequences whereas there was some pressure on genic promoters (Fig. 1-7A). This suggests there is less biological importance of orphan promoters than genic ones.

In summary, non-genic promoters and genic promoters for miRNA generally share core elements with genic promoters for protein-coding genes, and thus shared mechanisms for transcriptional initiation are expected. Exceptional characteristics were observed for intragenic promoters, so these may employ a different mechanism for transcriptional initiation.

Enrichment of TATA-less promoters for chloroplast and mitochondrial proteins

Nakamura *et al.* reported TATA-less promoters in the photosynthesis-related nuclear genes of higher plants are considerably more abundant than non-photosynthesis-related nuclear genes (62.5% for photosynthesis-related genes vs. 9.1% for non-photosynthesis-related genes) (Nakamura et al., 2002). My GO analysis extended this feature to genes for chloroplast and mitochondrial proteins (Fig. 1-4A). It does not correlate with a functional aspect of the TATA-type promoters, which is high expression (Fig. 1-4B). One of the common features for nuclear-encoded chloroplast and mitochondrial proteins is their origin: both

groups are rich in genes originating from the corresponding organellar genomes (Smith, 2014). As another set of genes sharing this feature, genes with cyanobacterial origin, also showed enrichment of TATA-less promoters, this feature is apparently important for the enrichment of TATA-less promoters.

5' UTR for the TATA-type promoters

High expression is one of the characteristics of genes with TATA-type promoters (Supplementary Fig. S1-7), and this is thought to be due to the high promoter activity of the TATA type. TATA-type genes have shorter 5' UTRs, which correlates with their higher expression levels (Fig. 1-6). Negative correlations between 5' UTR length and the expression level have also been reported in yeast (Lin and Li, 2012), plants (Yang, 2009, Yamamoto et al., 2011) and chickens (Rao et al., 2013). Interestingly, Kawaguchi and Bailey-Serres report that 5' UTRs with a length of 60 nt showed the highest ribosome loading in Arabidopsis (Kawaguchi and Bailey-Serres, 2005), and this length is very close to the peak length of 5' UTRs giving highest expression level, that is, the highest accumulation of transcripts (Fig. 1-6). It is thus reasonable to assume that short 5' UTRs from the TATA-type promoters contribute to a high translation efficiency. Taking these results and interpretations into account, I suggest that these two features of the TATA-type genes, high promoter activity and short 5' UTR, are not functionally connected but considered to be a consequence of two parallel selection pressures toward high gene expression.

A third feature found in TATA-type genes is the enrichment of As on the upstream neighboring side of the translational initiation codons (Fig. 1-8A). This local area shows strand bias towards A over T, and the enrichment of As at this site is ongoing by the biased appearance of SNPs (Fig. 1-8C, D and F). This position is a part of the Kozak sequence important for efficient translation (Kozak, 1981) and the Arabidopsis consensus fits with this

sequence (-3:A/C, +4:G). Interestingly, As at positions -1, -3, -4, -9, and -10 nt from the AUG are enriched in mRNA populations in polysome fractions of Arabidopsis (Kawaguchi and Bailey-Serres, 2005). Similar results are reported in yeast studies (Gingold and Pilpel, 2011, Dvir et al., 2013). Therefore, the highest base compositions for A at the upstream neighboring region of the ATG in Arabidopsis TATA-type genes provide high translational efficiency. The trend is stronger in genes with the TATA-type promoter than ones with the Coreless type. Again, this is considered to be a consequence of another parallel selection pressure towards high gene expression.

Refinement of promoter sequences towards characteristics of the majority

Analysis of SNPs among Arabidopsis natural variations revealed several trends. Comparison of appearance rates of a SNP pair sensitively detected strand bias in specific regions in the promoter structure. Assuming generation of SNPs occurs evenly along DNA strands, detected strand biases mean the presence of corresponding selection pressure during the fixation of SNPs among Arabidopsis populations.

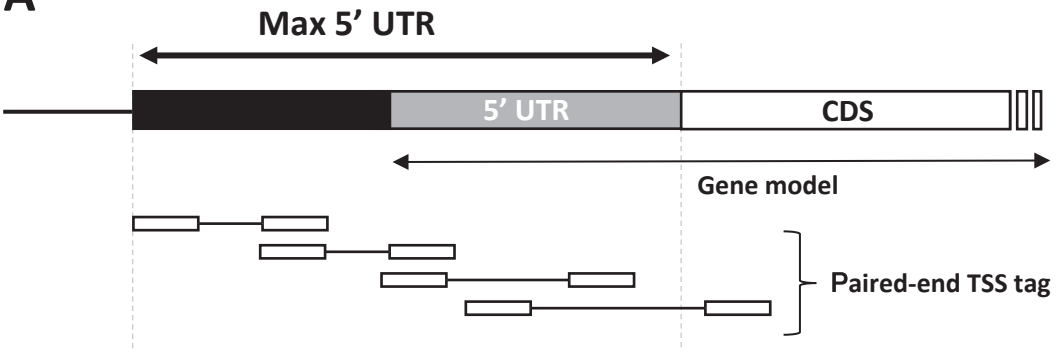
One surprise from a series of SNP analyses is that the majority of detected biases and trends of SNPs are already reflected in the majority of the current promoter sequences. This is true for low GC content throughout the promoter region, overrepresentation of C over G around the TSS for the TATA-type promoters, dominance of A in the upstream neighboring region of the ATG for the TATA type, and strand bias towards T at the upstream neighboring region of the TATA box for the TATA type. A possible interpretation of this reflection is that these biased SNPs occur in promoters whose sequence has not followed the characteristics of the majority of promoters, representing succession from premature to mature promoter sequence. It should be noted that all the local biases of base compositions and SNP ratios are more evident for the TATA-type promoters than the Coreless type. These differences suggest that the TATA-type promoters have more restriction on their full promoter function than the

Coreless type.

Three genic promoters for one gene

These paired-end and single-end deep sequencing of various TSS tag libraries has revealed that there are 2.7 genic promoters per protein-coding gene on average (58,551/21,672, Fig. 1-2B). My genome-wide analysis first identified that one promoter among them is the determinant of expression of the corresponding gene, contributing more than 80% of gene expression, and its companion promoters give minimum contribution to gene expression (Fig. 1-5A). I also found that the dominance of the top promoter gets higher in accordance with the expression level of the gene. This suggests that elevation of a gene's expression is achieved by an increase in the top promoter, and the companion promoters give no contribution. Logically to say, this path requires fewer steps than parallel enhancement of all the genic promoters for the focused gene and thus seems reasonable, because addition of a transcriptional positive regulatory element is known to enhance the corresponding promoter in a synergistic, not additive, way.

A



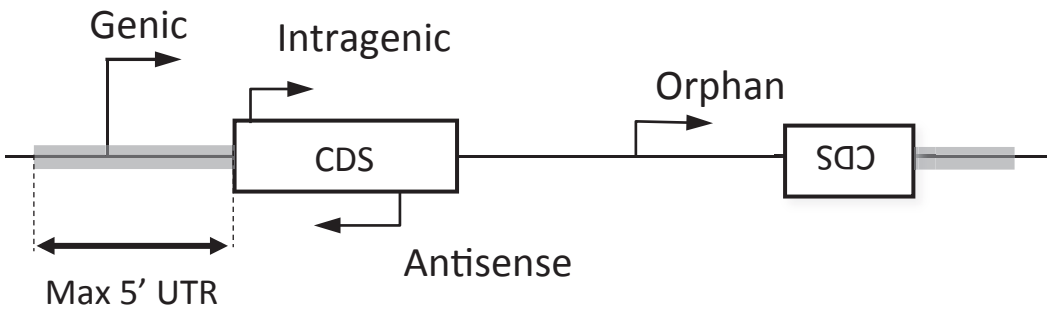
B

Description	number of genes (coverage)
Max 5' UTR	29,090 (87.3%)
Extended 5' UTR from TAIR10	27,526 (82.6%)
No extension of 5’ UTR info from TAIR10	1,564 (4.7%)
	length
Median length of Max 5' UTR	319 bp
Median length of extended Max 5' UTR	201 bp

Figure 1-1. Determination of Maximum 5’ untranslated regions (Max 5’ UTRs) using pair-end TSS analysis

(A) Schematic diagram of determination of Max 5’ UTR. The max 5’ UTR was determined by extending a gene model towards the 5’ direction using pair-end sequencing data of TSS tags (shown by a pair of white boxes connected by a line), (see Materials and Methods). The start point of "walking", the extension of the gene model in the 5' direction, is the 5' end of the gene model of TAIR10, which is the 5' end of the 5' UTR or that of the CDS when EST data for this region is not available. The downstream end of a TSS tag in the first walk should be within the gene model. (B) Summary of Max 5’ UTRs.

A



B

Category	Tag	Promoter number	Tag/ promoter	Gene	Coverage	Ratio of C2 tag-supported cluster
Total	33,172,231	324,461	102.24			77. 6%
Genic		59,628	465.54	22,211/33,323	66.7%	89.2%
-protein coding		58,551		21,672 / 27,206	79.7%	89.3%
-miRNA		33		19 / 177	10.7%	97.0%
-ncRNA (other,sn,sno)		387		200 / 478	41.8%	84.2%
-pseudogene		359		153 / 924	16.6%	83.3%
Intragenic		193,208	11.15	22,751/33,323	68.3%	82.5%
Antisense		42,070	31.77	14,825/33,323	44.5%	42.1%
Orphan		34,549	75.65			73.8%

Figure 1-2. Summary of genome-wide TSS identification using oligo-cap method in Arabidopsis
(**A**) Schematic diagram of categorization of TSS clusters. The 33 M TSS tags were grouped into 324,461 TSS clusters (= promoters) as explained in Supplementary Figure S2. Promoters were categorized as follows: Genic, clusters in Max 5' UTRs with the same direction as the gene; Intragenic, clusters in translated regions with the same direction as the gene; Antisense, clusters within transcribed regions with the opposite direction to the gene; Orphan, clusters which were not associated with any gene models. (**B**) Summary of TSS categorization. Genic promoters were further classified into protein coding, miRNA, ncRNA, and pseudogene. Pre-tRNA, pre-rRNA and transposable element (TE) genes were excluded from the analysis. C2 tag-supported clusters were determined as C1 clusters whose tags are supported by C2 tags for more than 50 %. For C2 tags, see Supplementary Figure S2.

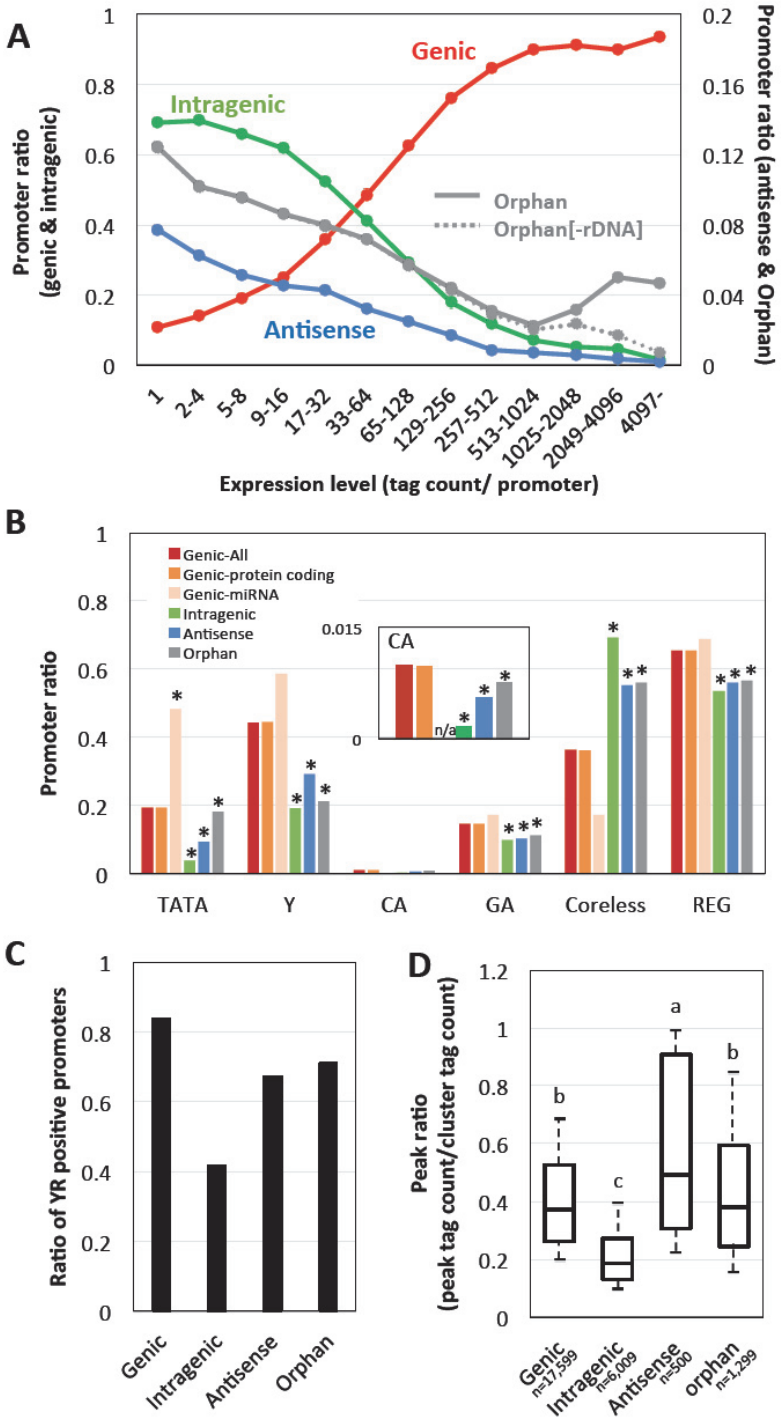


Figure 1-3. Expression characteristics, core promoter ratio and ratio of TSS type in promoter category
Promoters supported by C2 tags (Supplementary Figure S4) were used for the analysis. (**A**) Ratio of promoter category in relation to expression level. The gray dotted line indicates the result of orphan promoters after the exclusion of 170 promoters that locate within 4 kb of rDNA. Promoters which have more than one category due to nested gene models were eliminated from the analysis. (**B**) Core promoter ratio for each promoter category. The insert is an enlarged graph for the CA ratio. Asterisks indicate a significant difference from Genic-All shown in red (Fisher's exact test, $p<0.05$). (**C**) Ratio of TSS type matching YR rule for each promoter category. (**D**) Comparison of peak ratio for each promoter category. The peak ratio is an indicator of promoter shape, and a high ratio means a peaky TSS cluster. Promoters which have over 33 tags were used for this comparison. Horizontal lines of the box plot figures indicate the median peak ratio and boxes include second and third quartiles. The top and bottom whiskers indicate 75-90% and 10-25% of the population of the peak ratio, respectively. The different letters indicate a significant difference (pairwise Wilcoxon rank-sum test with Bonferroni correction, $p<0.05$).

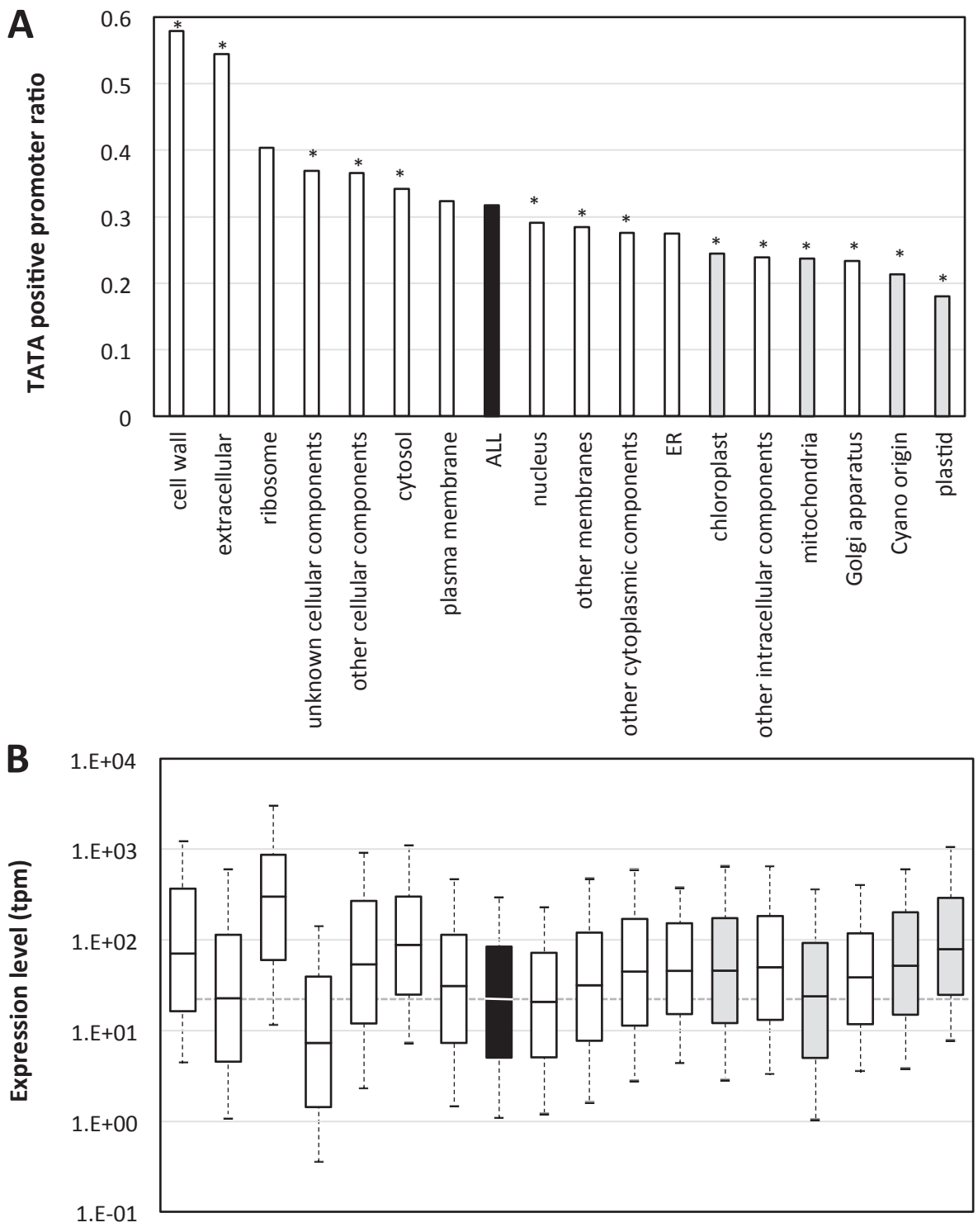


Figure 1-4. Preference of TATA-positive and -negative genic promoters among GO categories
(**A**) TATA ratio of genic promoters for different cellular components is shown. Cellular components information was obtained from TAIR 10. "Cyano origin" is a group of Arabidopsis nuclear genes that are suggested to have cyanobacterial origin (Martin *et al*, 2002 (29)). Asterisks indicate a significant difference from ALL promoters (Fisher's exact test, $p<0.05$). (**B**) Expression levels of genes for each cellular component are presented as a boxplot with median, second and third quartiles. The upper and lower whiskers indicate 75-90% and 10-25%, respectively. The horizontal dashed line shows the median expression level of ALL promoters.

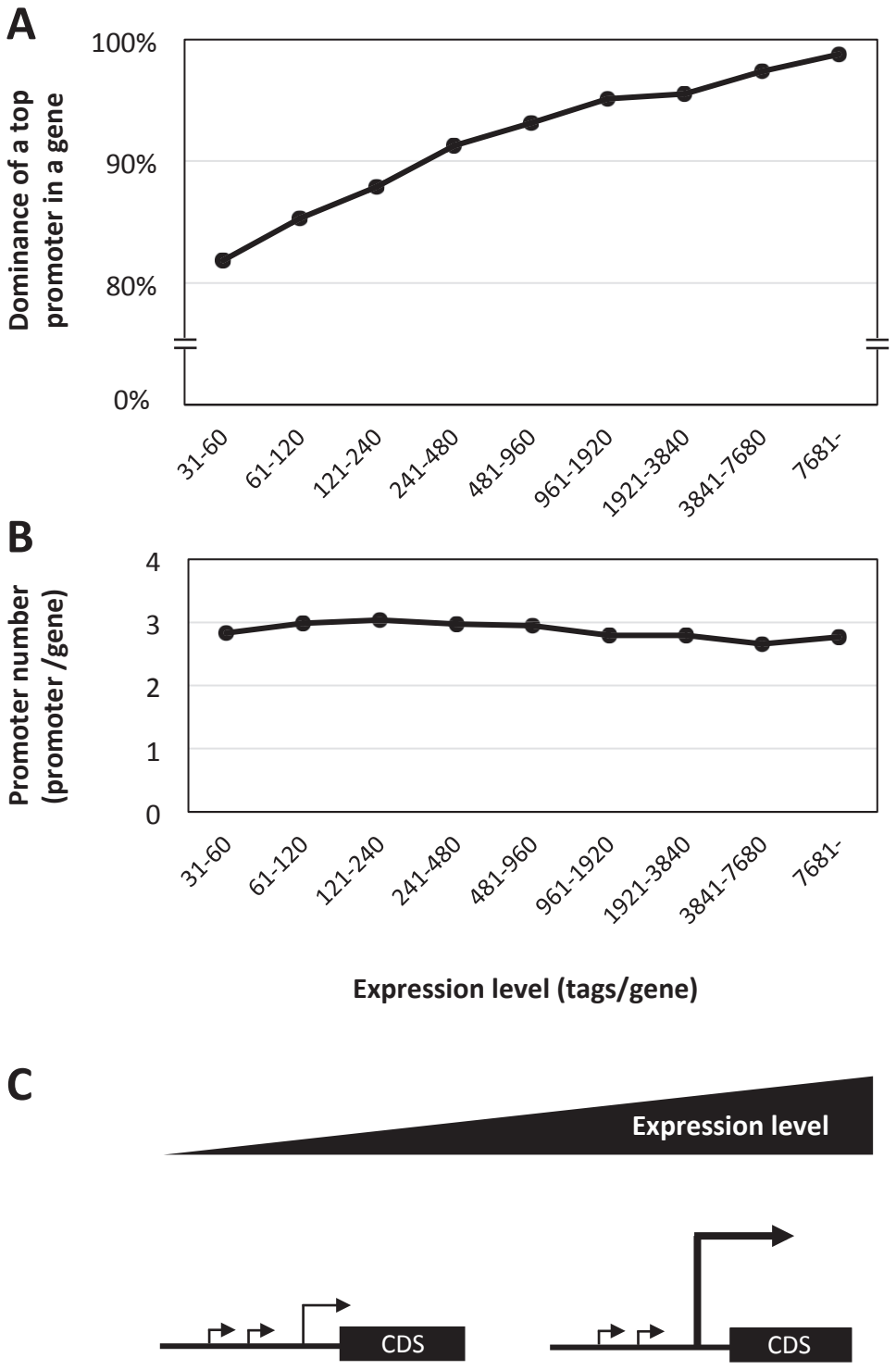


Figure 1-5. Gene expression is predominantly determined by the top genic promoter
(**A**) Expression ratio of most dominant (top) promoter among genic promoters for gene is shown. The horizontal axis shows the expression level of a gene, and the vertical axis indicates the ratio of the tag count of the top promoter in a gene to total tag count of the genic promoters for the corresponding gene. 16,892 genes containing over 30 tags were selected for the analysis. (**B**) Number of genic promoters per gene is shown in regard to expression level. (**C**) Schematic diagram of contribution of multiple promoters to gene expression.

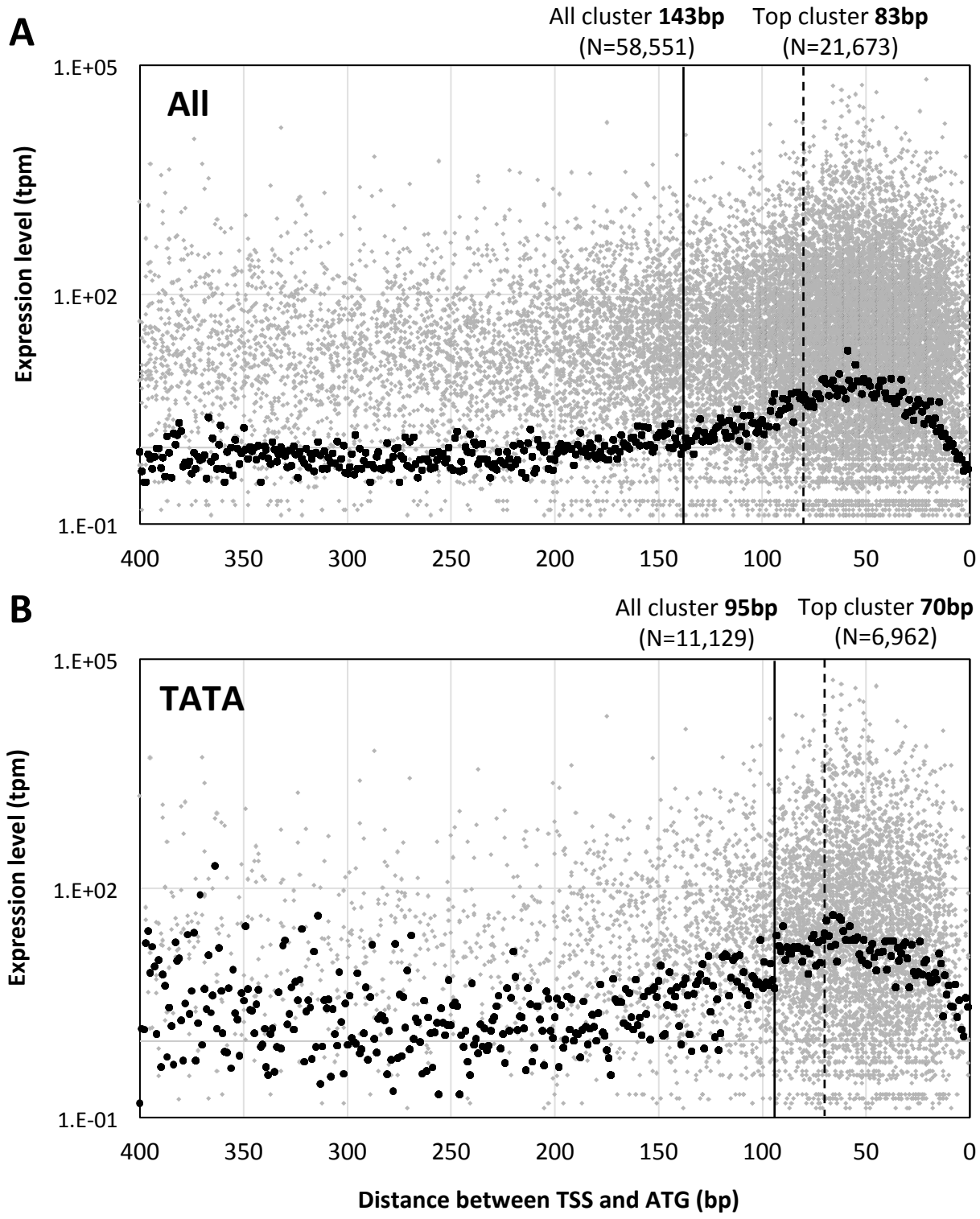


Figure 1-6. Relationship between length of 5' UTR and expression level
The graphs show relationships between TSS position and expression level for genic promoters. (**A**) Gray and black dots represent individual promoters and median expression level for each distance of All promoters (58,551). Median 5' UTR lengths of All (58,551) and top (21,673) promoters are also shown by solid and dashed vertical lines, respectively. (**B**) Gray and black dots represent individual promoters and median expression level for each distance of TATA-type promoters (11,129). Median 5' UTR lengths of All (11,129) and top (6,962) promoters are also shown by solid and dashed vertical lines, respectively.

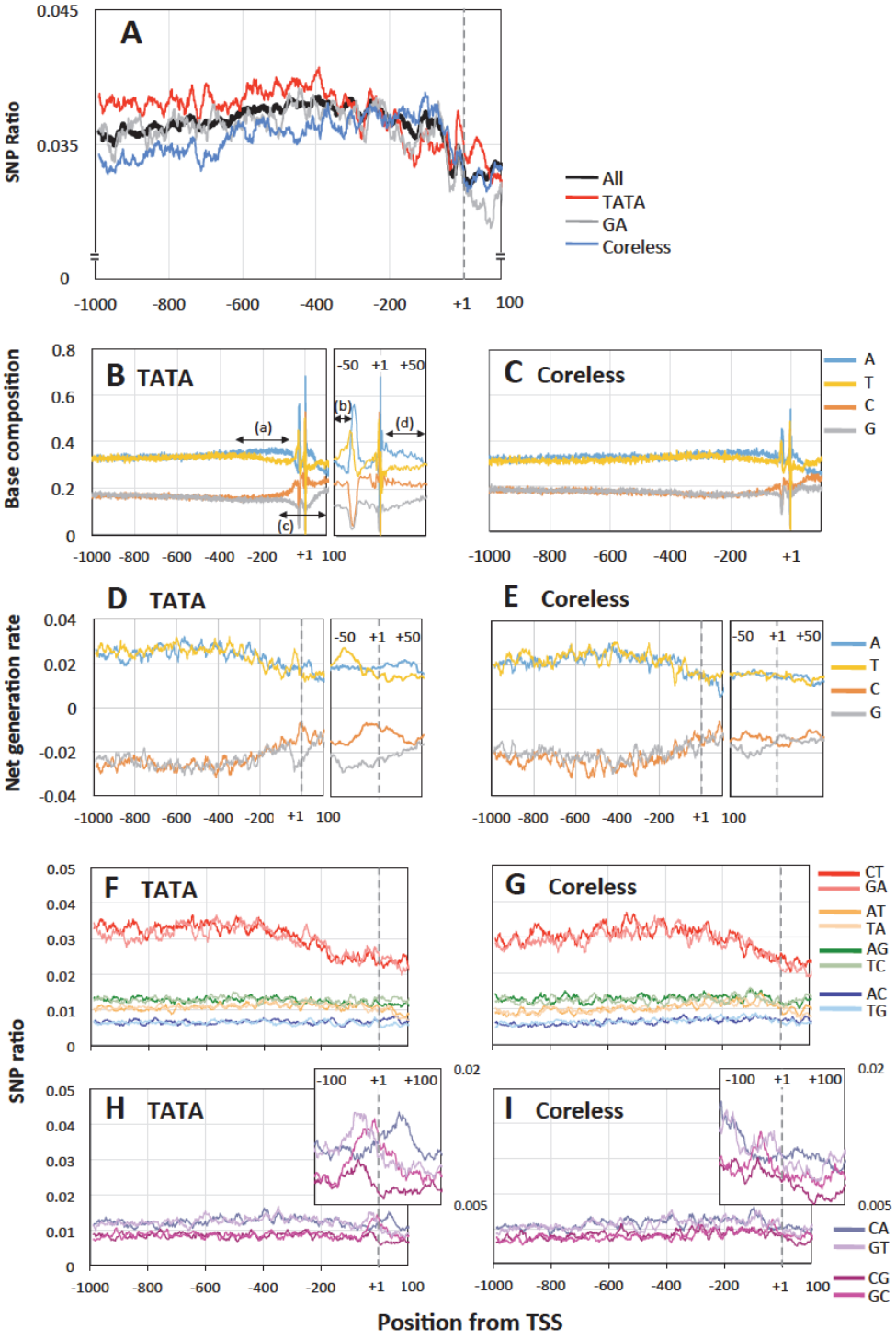


Figure 1-7. Analysis of SNP appearance among Arabidopsis intra-species variants
Genome sequences of 81 Arabidopsis accessions (Cao *et al*., 2011 (30)) were subjected to SNP analysis of genic top promoters. SNP ratios were normalized according to the base composition for each promoter position (Supplementary Figure S3). A total of 16,896 genic top promoters whose SNP data is available for ≥40 accessions was used for the analysis. The numbers of analyzed promoters of the TATA, GA, and Coreless types were 4,383, 3,003, and 4,188, respectively. The SNP ratio data was subjected to smoothing with a bin of 21bp unless otherwise stated. **(A)** The sum of the SNP ratios in each core promoter type is shown. **(B** and **C)** Base compositions of promoter regions in TATA (**B**) and Coreless (**C**) types are shown. Graphs were not smoothed. The small letters in panel B indicate the regions significantly showing strand bias [(a) higher occurrence rate of A over T from -200 to -50, (b) higher rate of T over A from -50 to -30, (c) higher rate of C over G from -100 to +100, and (d) higher appearance rate of A over T from +1 to +50]. **(D** and **E)** Net generation rates (see Materials and Methods) for each nucleotide in TATA (**D**) and Coreless (**E**) types are shown. **(F - I)** Individual SNP rates are shown for TATA (**F** and **H**) and Coreless (**G** and **I**) types.

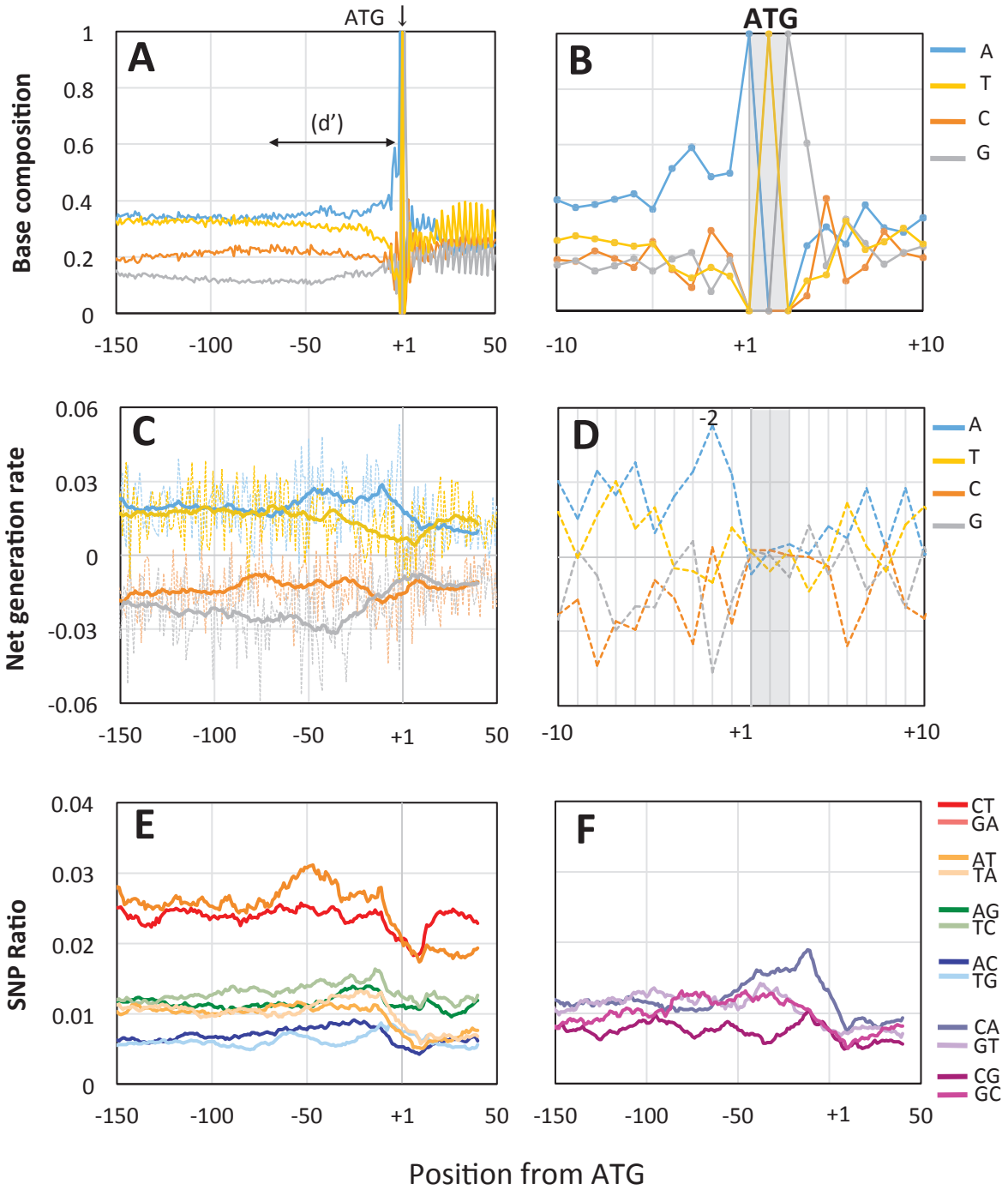


Figure 1-8. SNP appearance of the TATA type around ATG
The TATA type of genic top promoters, with a 5' UTR shorter than 150 bp, was subjected to the analysis. (**A** and **B**) Base composition relative to ATG is shown. Small letter [(d')] in the panel A indicates the regions significantly showing strand bias. (**C** and **D**) Net generation rates of each nucleotide type in the 5' UTR of the TATA-type promoter are shown. The rates before and after smoothing with 21bin are shown with dashed and solid lines, respectively. (**E** and **F**) Each normalized SNP ratio with 21bin smoothing is shown in different colors. Regions from -150 to +50 bp (**A**, **C**, **E** and **F**) or -10 to +10 bp (**B** and **D**) relative to the ATG are shown.

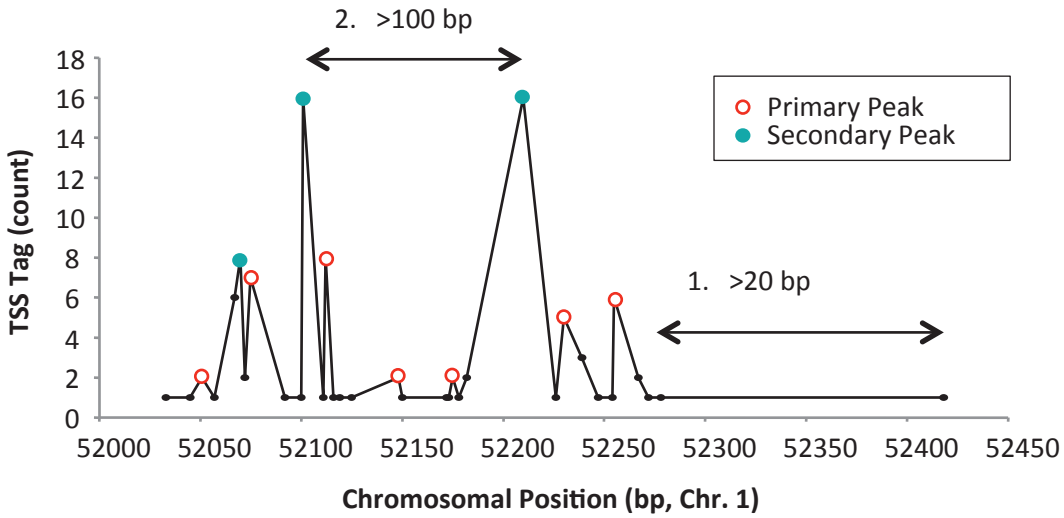


Figure S1-1. Schematic diagram for clustering of TSSs
A primary peak means a local maximum point of TSS accumulation, and a secondary peak means a local maximum point of primary peaks. Positions without any TSS tags were ignored for the clustering. Two criteria for dividing TSS clusters were applied. First, TSSs with a distance of more than 20 bp were divided into two clusters. Second, two secondary peaks with a distance of more than 100 bp were divided into two clusters. This clustering method allowed determination of 324,461 TSS clusters from 33 M TSS tags.

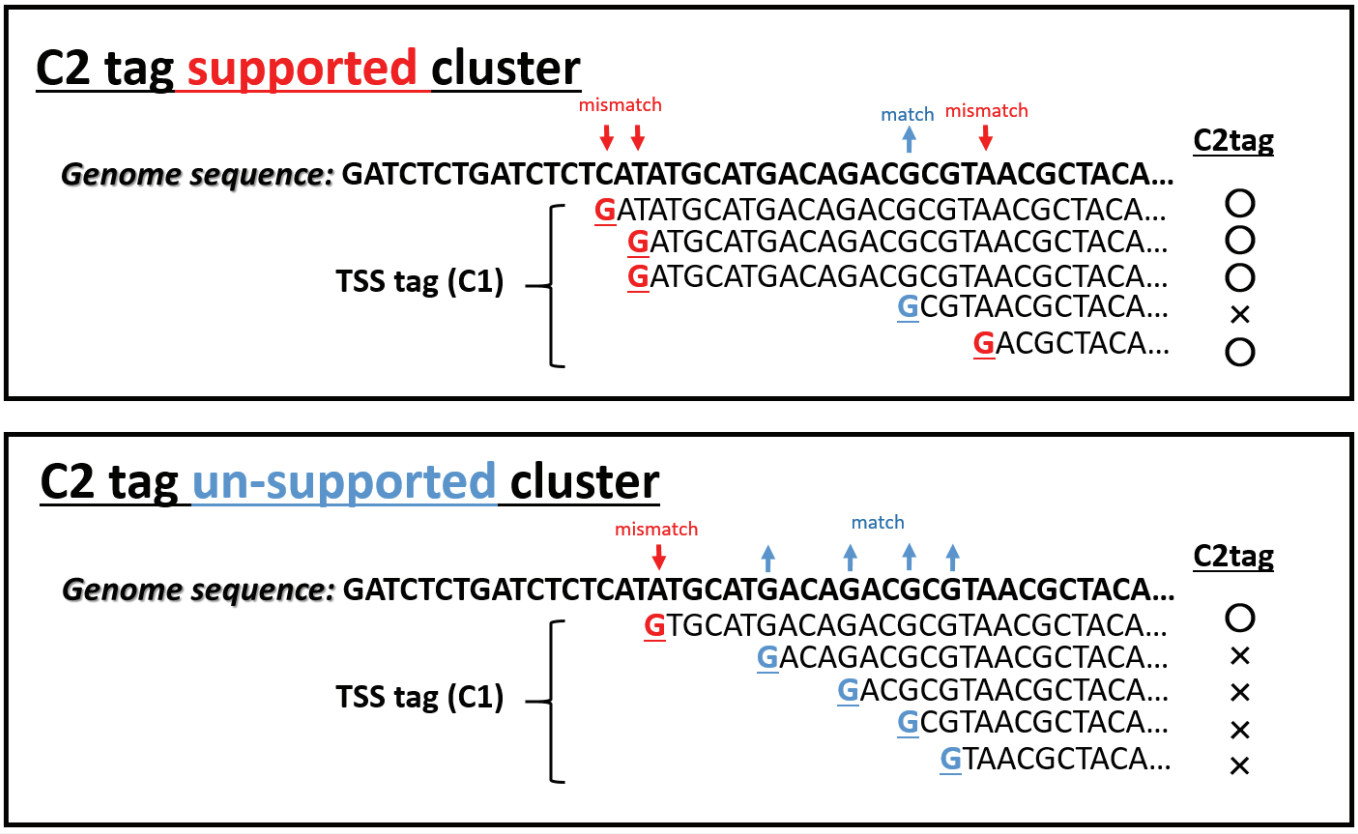
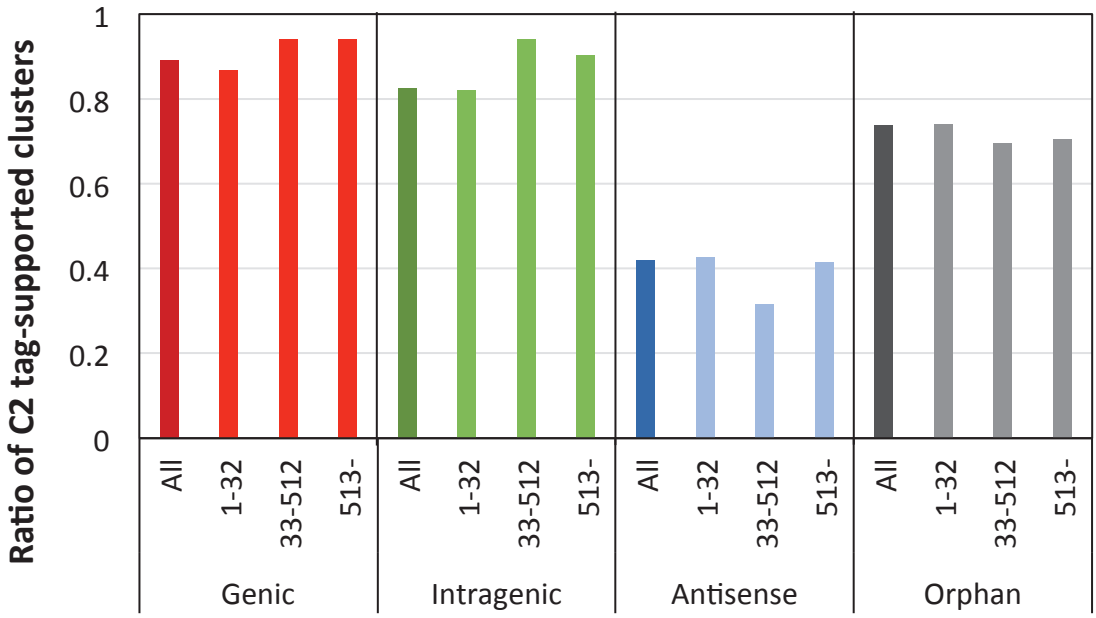


Figure S1-2. Schematic diagram for TSS clusters supported by C2 tags
Both C1 and C2 tags contain Cap Signature, a guanine (G) residue at the 5' end of a TSS tag due to the cap-dependent terminal deoxynucleotide transferase (TdT) activity of a MMLV reverse transcriptase (SuperScript II) (Yamamoto *et al*., 2009(9)). However, this signature sequence allows another interpretation when the corresponding sequence of the genome is also G. In order to distinguish real TSS tag clusters from clusters derived from end of transcript fragments which do not contain the cap structure, we prepared C2 tags whose Cap Signature does not correspond to the genomic sequence, that is, A, C, or T. Our stricter definition of TSS clusters contains C2 tags. The upper panel shows an example of a tag cluster containing C2 tags, and the lower panel shows a tag cluster containing C1 but not any C2 tags. The tag cluster of the lower panel has the potential of being a false TSS cluster.

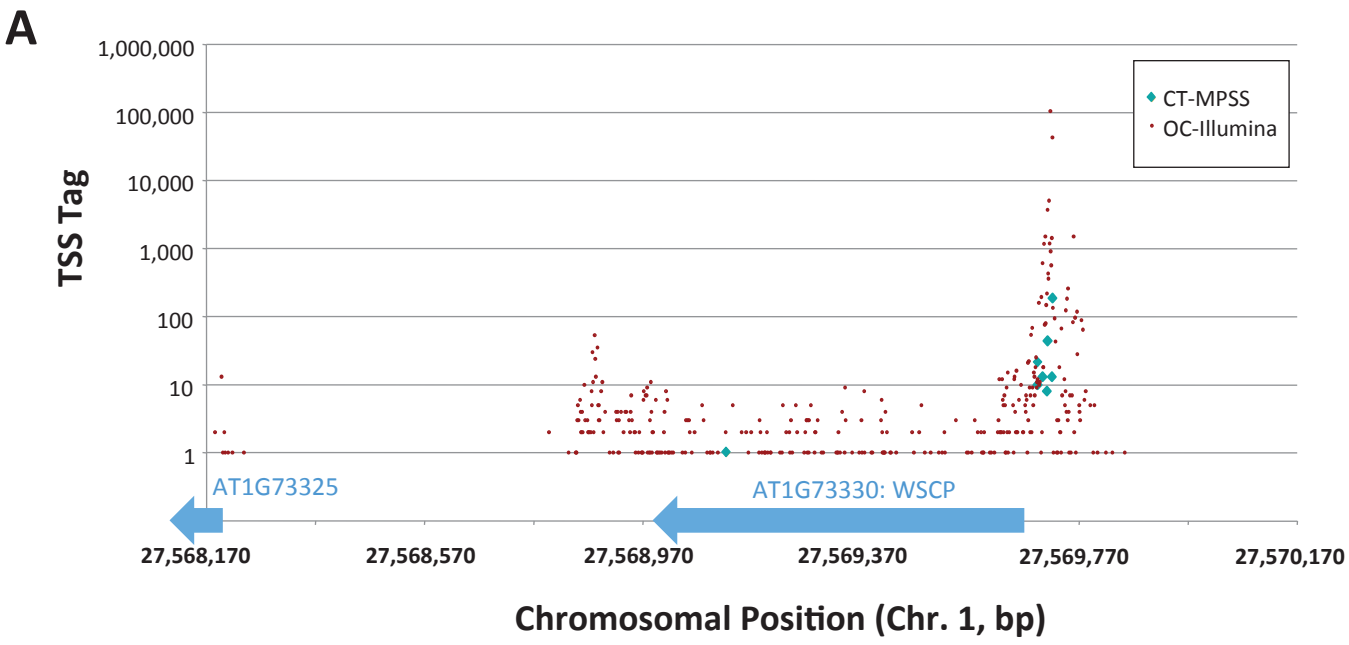
A

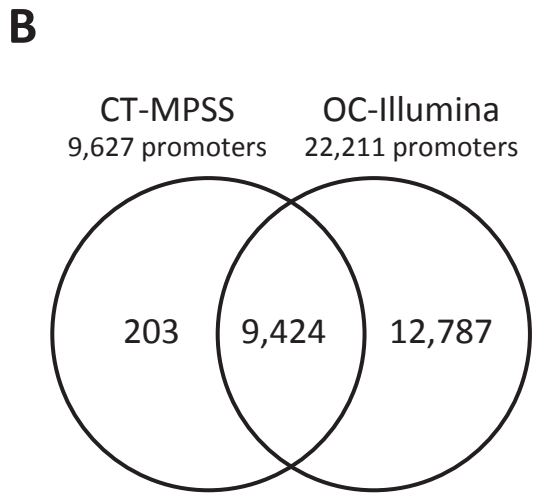


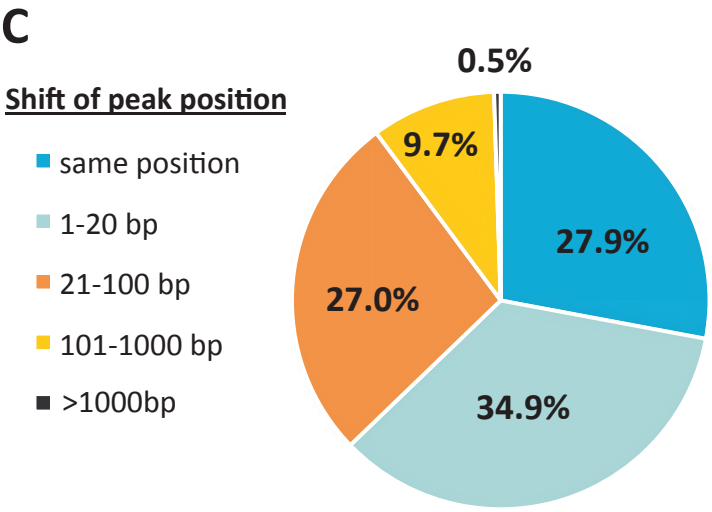
B

Category	C1 clusters	C2-supported C1 clusters
Total	324,461	251,749
Genic	59,628	53,202
_protein coding	58,551	52,292
_miRNA	33	32
_ncRNA	387	326
_pseudogene	359	299
Intragenic	193,208	159,342
Antisense	42,070	17,692
Orphan	34,549	25,503

Figure S1-3. Evaluation of TSS clusters using information of C2 tags
C2 tag-containing clusters were identified as shown in Figure S2, and the ratio of C1 tags containing C2 tags for each cluster was calculated. If the cover ratio by C2 was higher than 50%, the C1 cluster was judged as "supported by C2 tags". (**A**) Ratio of C2-supported C1 clusters in each promoter category is shown. Promoter categories were further divided according to expression level. (**B**) Numbers of C1 clusters and C2-supported C1 clusters are shown.







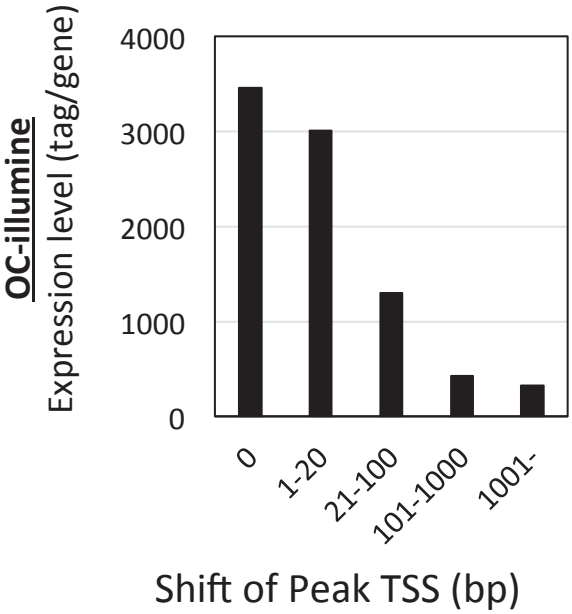


Figure S1-4. Comparison of TSS data between new analysis in this report (OC-Illumina) and previous analysis (CT-MPSS) (**A**) Comparison of the TSS tag distribution in AT1G73330 (water soluble chlorophyll-binding protein, WSCP) between OC-Illumina and CT-MPSS. (**B**) Venn diagram comparing the numbers of promoters identified by CT-MPSS and OC-Illumina. (**C**) Shift of peak position of the top cluster in OC-Illumina from the corresponding peak in CT-MPSS. 9,424 common promoters were used in this comparison. (**D**) Distance of peak shift and expression level.

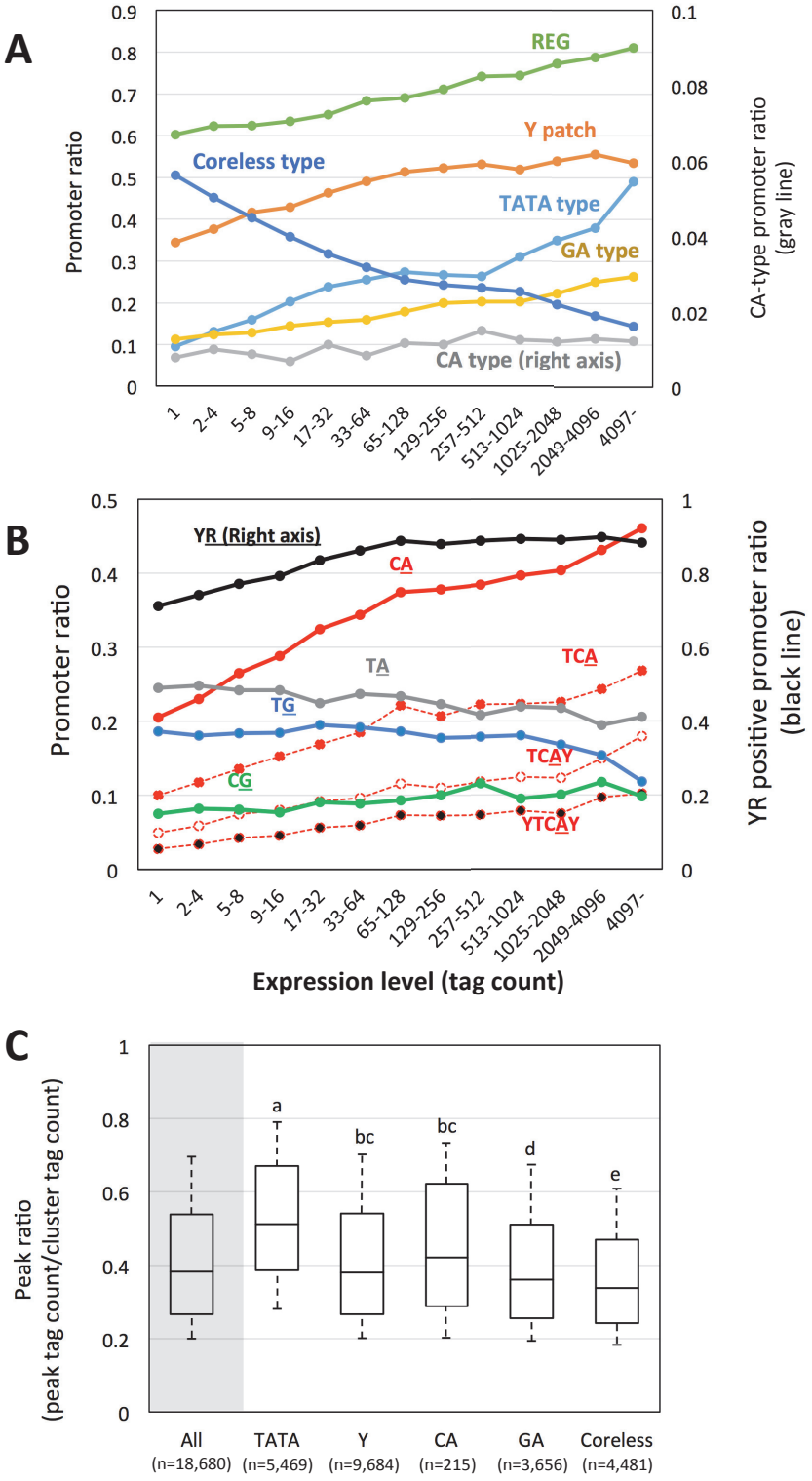


Figure S1-5. Relationship between expression level and promoter characteristics of genic promoters
(**A**) Relationship between expression level and the core promoter elements is shown. Promoter ratio of the CA type (gray line) corresponds to the right axis. (**B**) Relationship between expression level and TSS type is shown. TSS types shown in the graph include -1/+1 sequences matching YR Rule (CA, CG, TA, and TG) and Initiator (Inr)-like motifs (TCA, TCAY, and YTCAY,;+1 is underlined; Y: pyrimidine = C or T; R = purine: G or A). Ratio of YR-TSS type (black line) is shown on a scale on the right axis. (**C**) Peak ratio of promoters containing core promoter elements. Promoters which have over 33 tags were used for this comparison. Horizontal lines of the box plot figures indicate the median peak ratio and the boxes include second and third quartiles. The top and bottom whiskers indicate 75-90% and 10-25% of the population of the peak ratio, respectively. The different letters above the figure indicate significant differences (pairwise Wilcoxon rank-sum test with Bonferroni correction, $p<0.05$).

AT3G10745 miRNA

Chr3_3366564

(-100)

CCTTTAGCGGCGGCTTTACTTTAGATTCT**TCTAGGGT** TTCTAGATTGTAT**ACCCTAGA** TAA

At658-REG : TCTAGGGT At658-REG : ACCCTAGA

Chr3_3366464

TSS(+1)

GCA**TCCTATAAAG**TAAACACAAGTACTTGCAGAGACTTTAGATTAGAGGGCTAGCGACTG

At304-TATA : TCCTATAA

At281-TATA : CCTATAAA

At255-TATA : CTATAAAG

C**AGAAGAAG**AGTAACACGTCATCTCTGTGCTTCTTTGTCTACAATTTTGGAAAAAGTGAT

At736-GA : AGAAGAAG

GACGCCATTGCTCTTTCCCAAATGTAGACAAAGCAATACCGTGATGATGTCG

Figure S1-6. Example of miRNA promoter which has core promoter elements and REG.

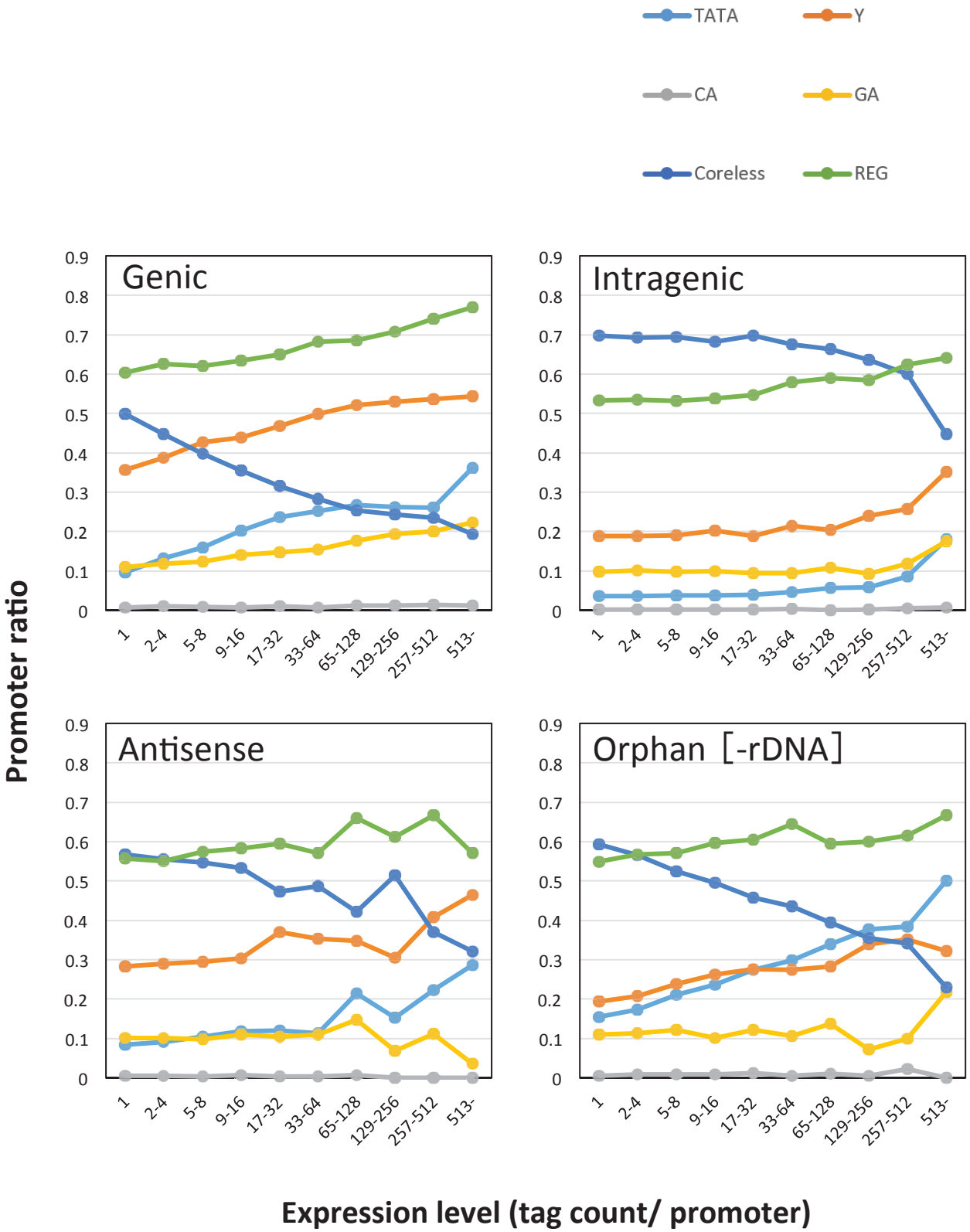


Figure S1-7. Ratio of core promoters and expression level in each promoter category

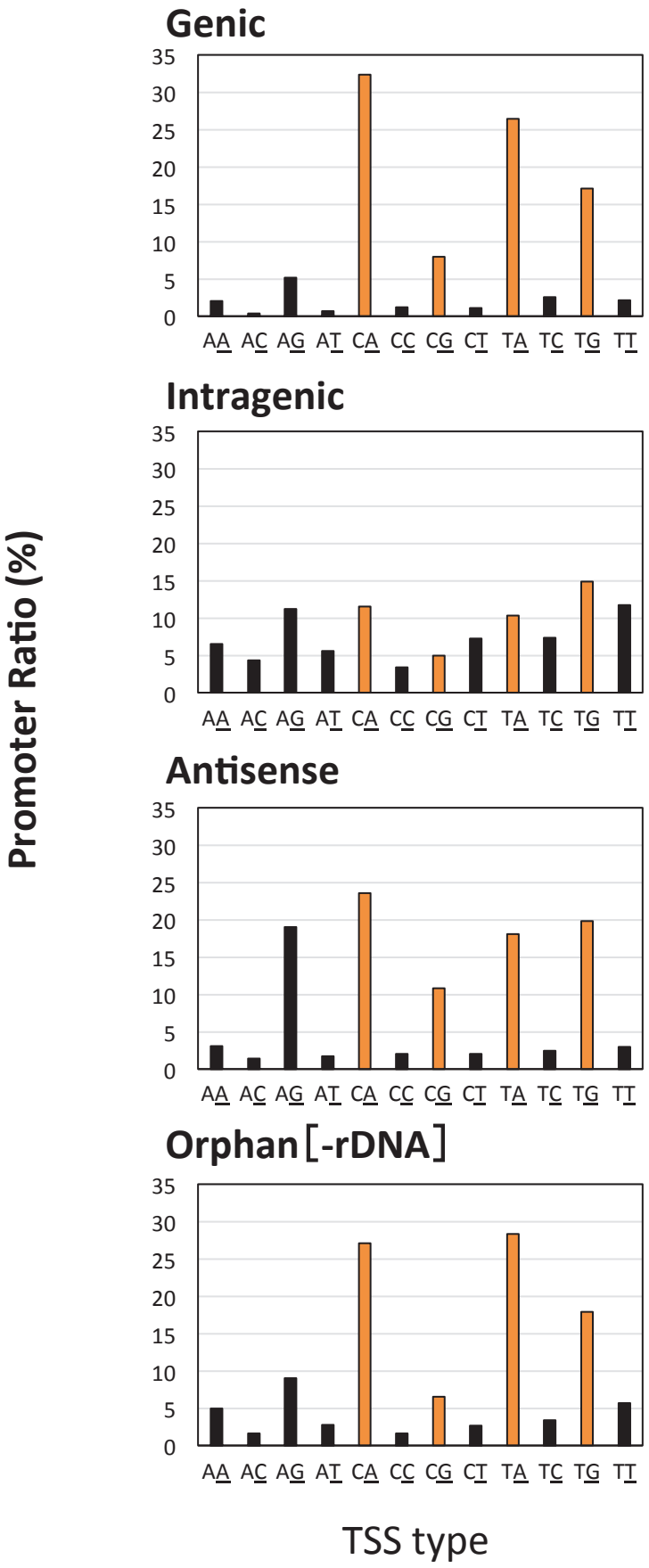


Figure S1-8. Difference in utilized dinucleotide sequences at TSS of promoter categories
The ratio of promoters having dinucleotide sequences at the TSS is shown. The underlined characters indicate the position of the TSS (+1). Sequences involved in the YR motif are shown in orange. GN dinucleotides (G<u>A</u>, G<u>C</u>, G<u>G</u> and G<u>T</u>) are excluded from the results, because C2-supported clusters were used for the analysis.

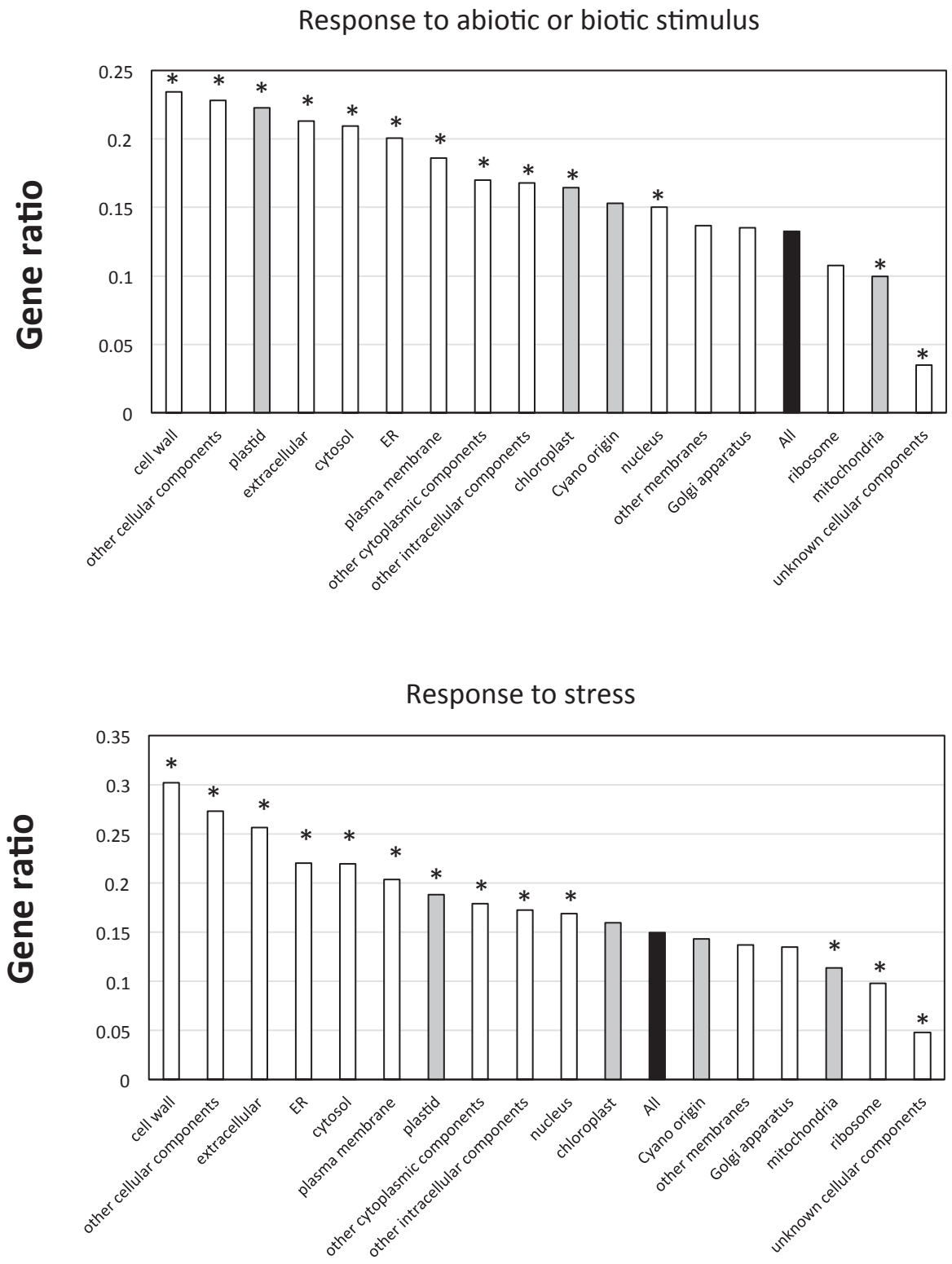


Figure S1-9. Ratios of stress-responsive genes among different cellular components
Asterisks indicate a significant difference relative to All promoters (Fisher's exact test, $p<0.05$)

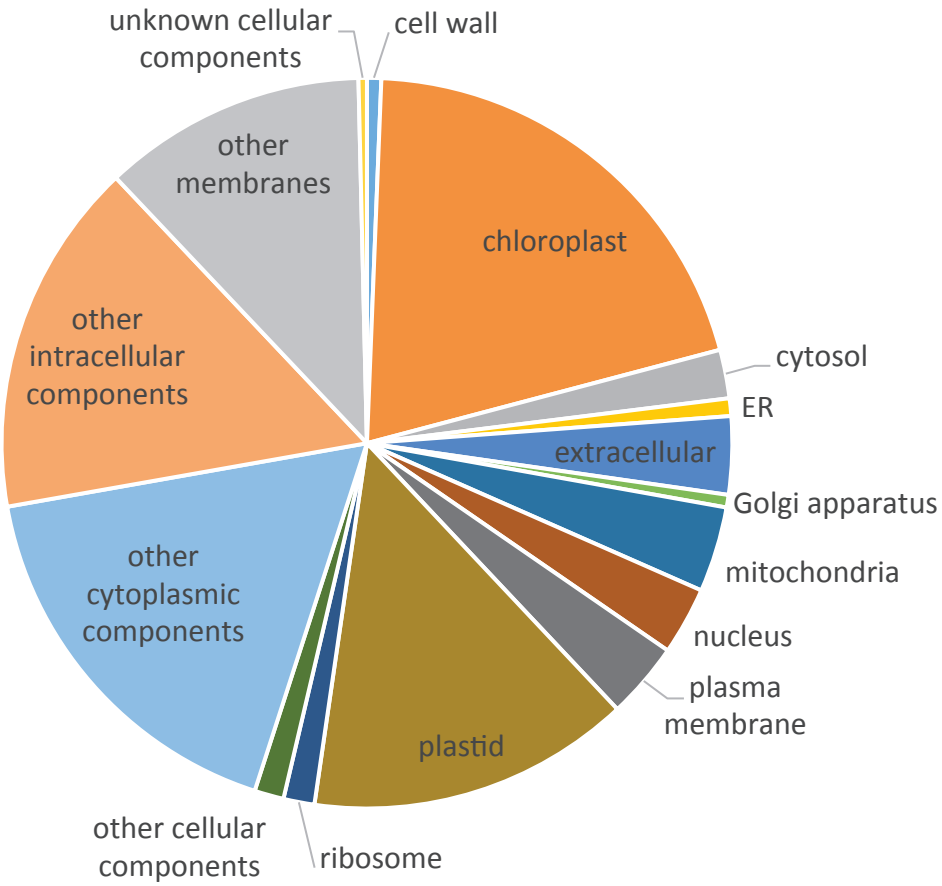


Figure S1-10. GO distribution of the gene group of Cyano origin
Classification of cellular components is shown for GO analysis.

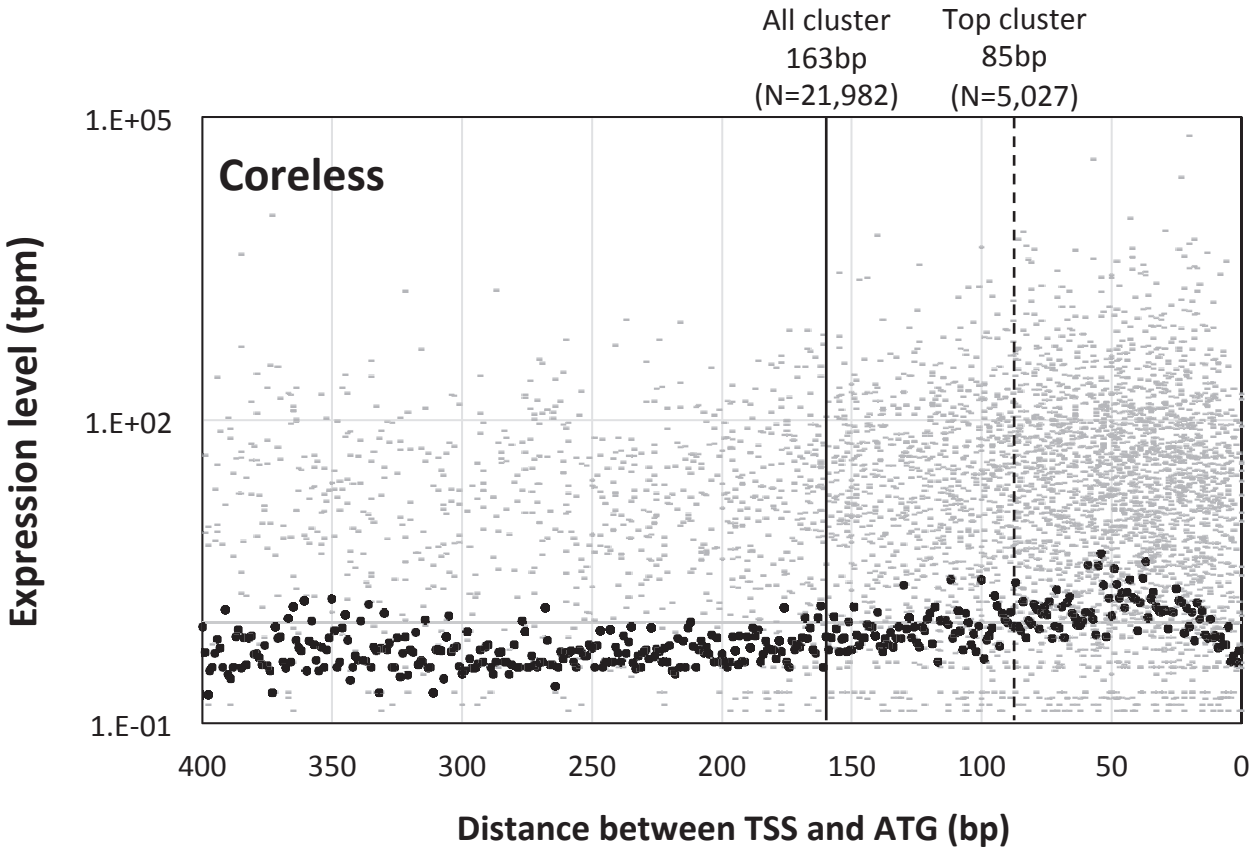


Figure S1-11. Relationship between length of 5' UTR and expression level in the Coreless type
Results of individual genes for the Coreless type of top genic promoters and the median expression levels for each distance are shown with gray and black dots, respectively. Median 5' UTR lengths of top genic promoters and all promoters for the Coreless type are shown as solid and dashed vertical lines, respectively.

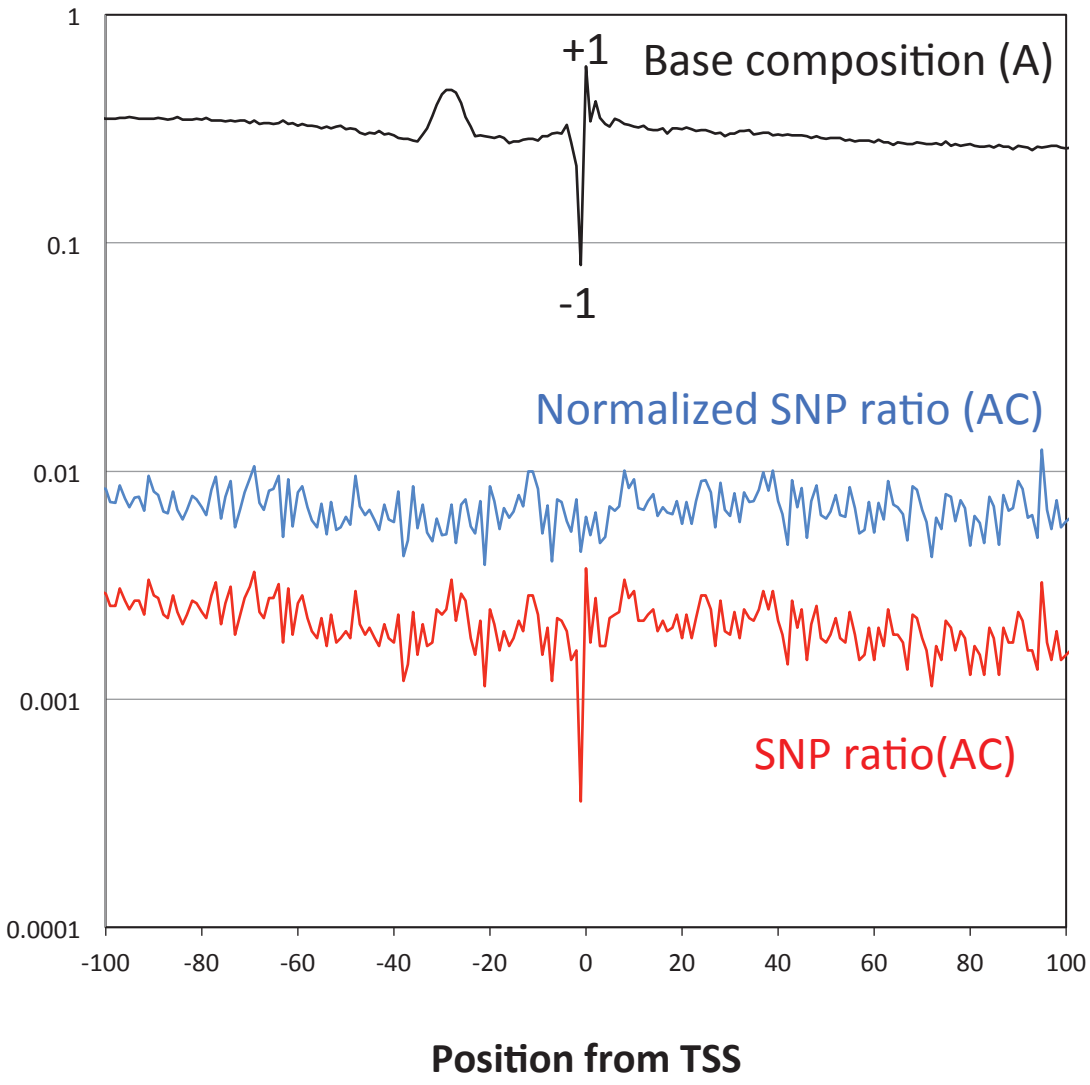


Figure S1-12. Normalization of SNP ratio according to base composition
The observed raw SNP ratio (red) was divided by the base frequency (black) to give a normalized SNP ratio (blue).

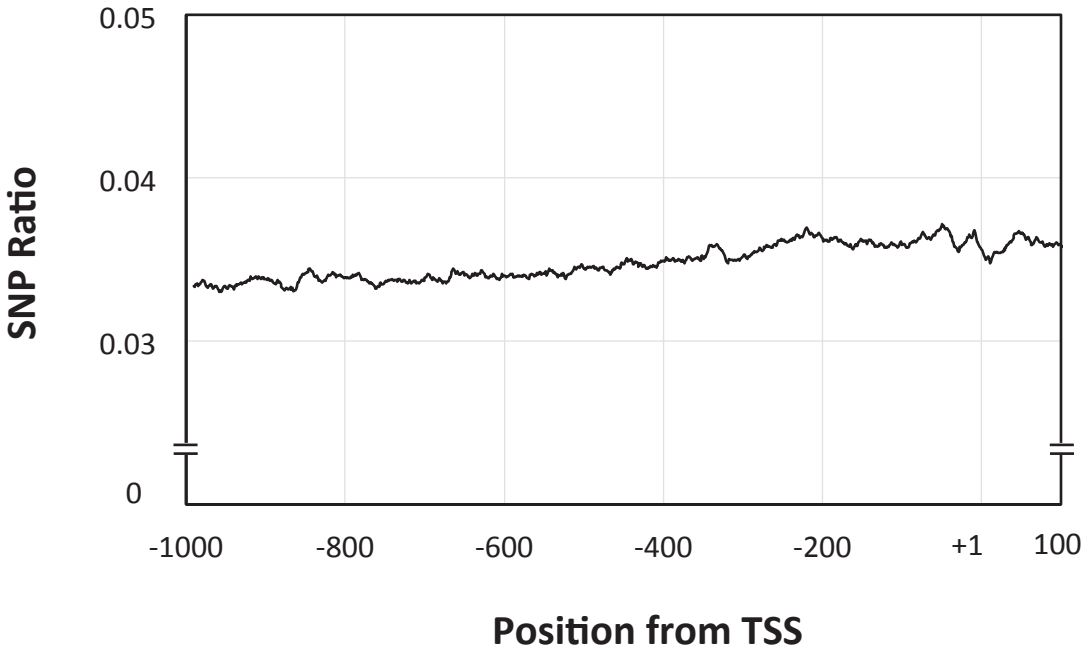


Figure S1-13. SNP ratio of orphan promoters
The sum of the SNP ratios for 22,981 Orphan promoters was subjected to smoothing with a bin of 21 bp.

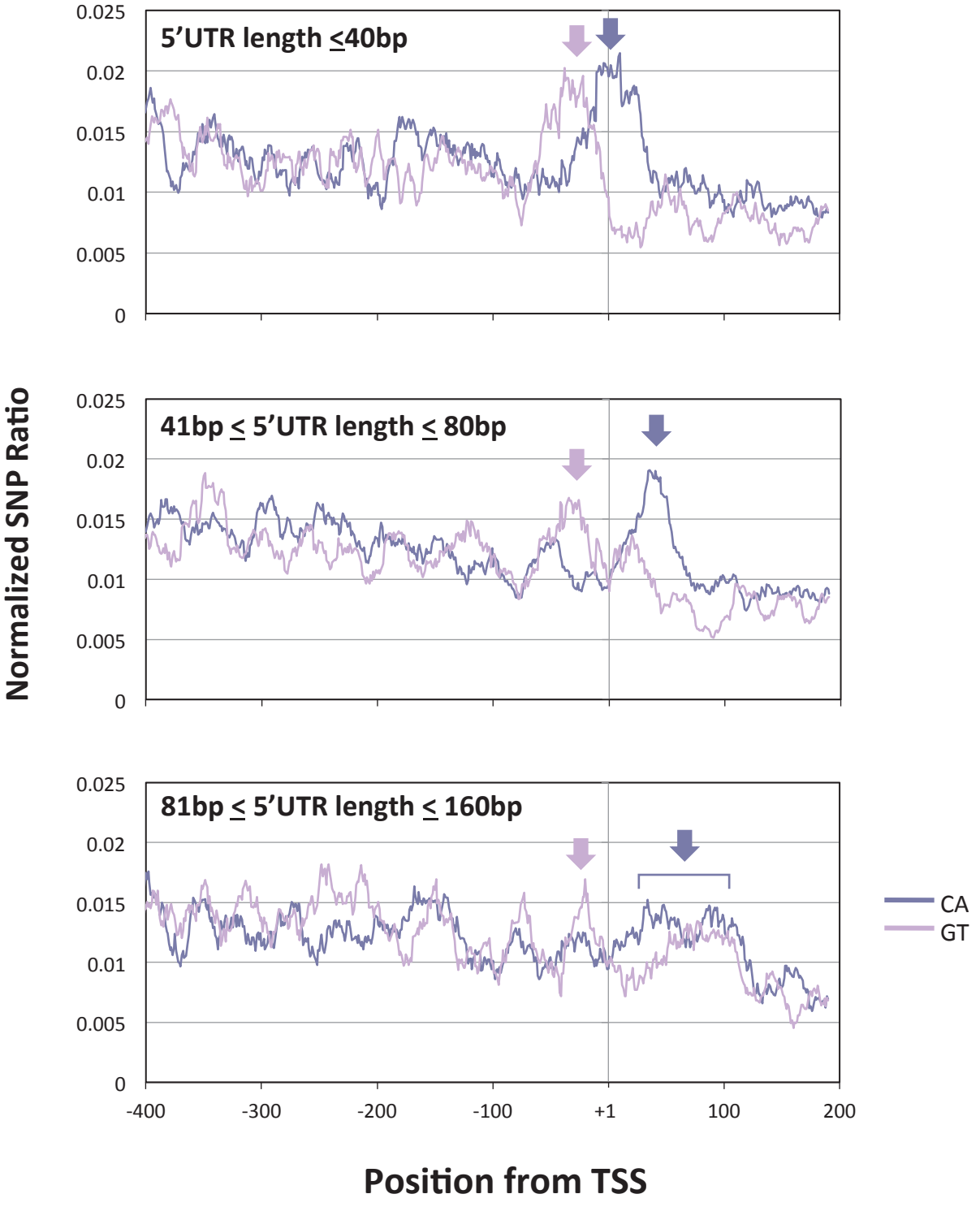


Figure S1-14. Comparison of the SNP ratios of CA and GT among three groups of genes with different 5' UTR lengths

Arrows indicate peak positions.

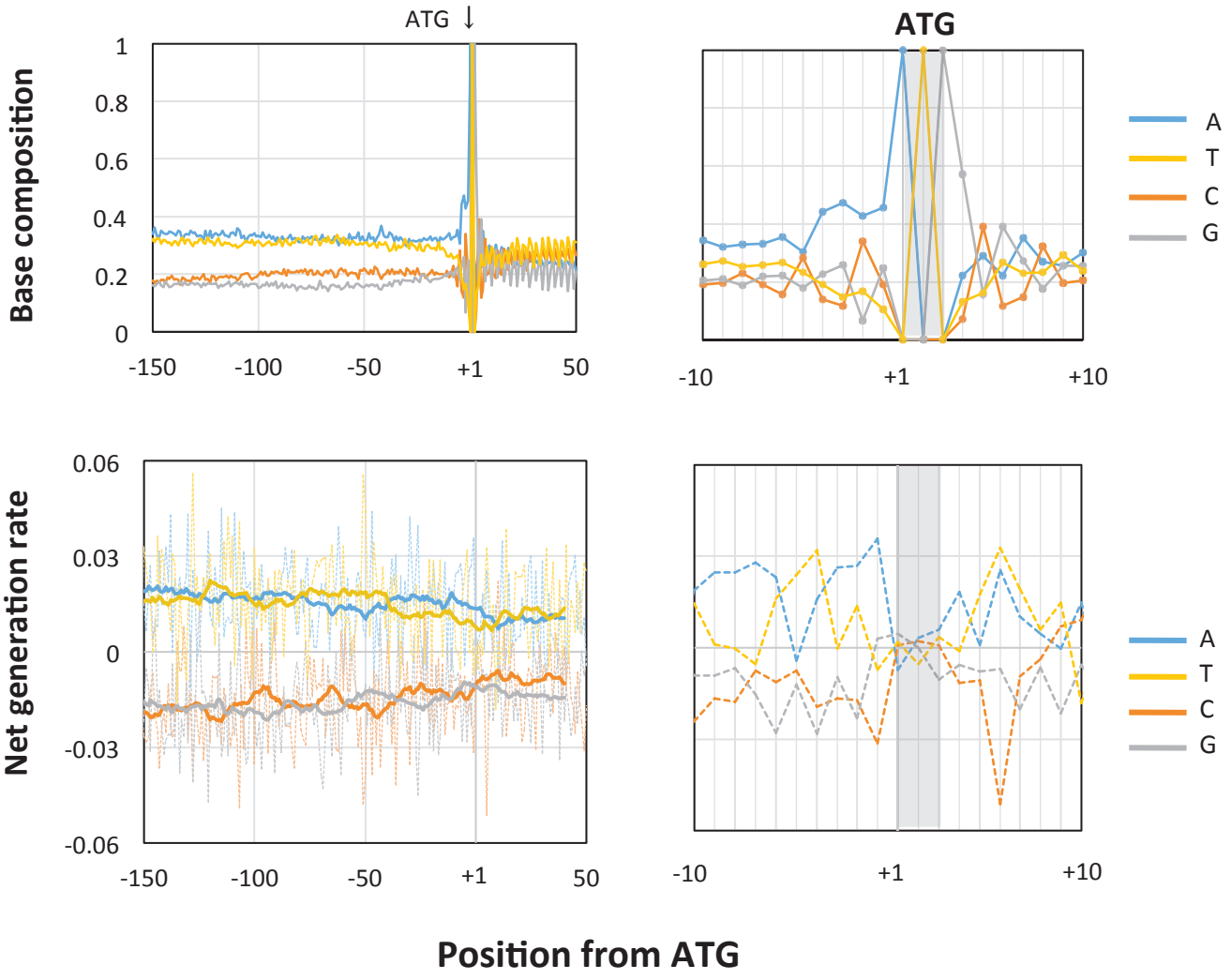


Figure S1-15. SNP ratios around ATG of Coreless promoters

(**A** and **B**) The base composition around the ATGs of 2,207 Coreless promoters, with the length of 5' UTRs shorter than 150 bp, are shown. (**C** and **D**) Net generation rates of each nucleotide type around the ATGs of the Coreless promoters are shown. Net generation rates before and after smoothing with a 21bin are shown in dashed and solid lines, respectively.

Chapter2

Analysis of *Aluminum-activated malate transporter1* promoter

2-1. INTRODUCTION

Organic acid (OA) excretion from the roots plays beneficial roles in stress adaptation processes of plants (Baetz and Martinoia, 2014). The root-exuded OAs de- toxify rhizotoxic ions, such as aluminum (Al) and copper (Kochian et al., 2004) and improve availability of phosphorus (Neumann et al., 1999) and iron (Kobayashi and Nishizawa, 2012). These roles are associated with the chemical properties of OAs, which can form chelate compounds with a variety of metals. For example, Arabidopsis (Arabidopsis thaliana) protects the root tip from Al toxicity by excreting malate and citrate through different OA transporters, namely ALUMINUM-ACTIVATED MALATE TRANSPORTER1 (ALMT1; Hoekenga et al., 2006) and a citrate-transporting multidrug and toxic compound extrusion (Liu et al., 2009). In addition, OAs can recruit beneficial rhizobacteria to the root surface by chemotaxis (Rudrappa et al., 2008). Certain bacteria form a biofilm on the root surface, which triggers systemically induced resistance (Lakshmanan et al., 2012). Excretion of OAs from the roots functions as a master switch through their pleiotropic roles in both biotic and abiotic stress tolerance. A recent molecular physiological study shows

transcriptional regulation of genes for OA transporters play critical roles in optimization of OA excretion in stress response (Liu et al., 2014).

The ALMT1 protein was first identified in wheat (Triticum aestivum; TaALMT1), which regulated a major Al tolerance mechanism in wheat through Al exclusion by Al-activated malate excretion (Sasaki et al., 2004). Functional orthologs regulating Al tolerance have been identified in Arabidopsis (AtALMT1; Hoekenga et al., 2006), Glycine max (GmALMT1; Liang et al., 2013), and other plant species. The complex transcriptional regulation of these orthologs is consistent with the pleiotropic roles of the root-excreted malate. Transcription of the Arabidopsis ortholog *AtALMT1* is activated by Al (Kobayashi et al., 2007) and by other signal inducers, including a type of microbe-associated molecular pattern peptide, flagellin22 (Kobayashi et al., 2013a). *GmALMT1* expression is induced by multiple stressors, namely Al, phosphorus deficiency, and low pH (Liang et al., 2013). Transcriptional regulation also plays roles to optimize malate excretion in terms of carbon economy during malate excretion. For example, Al induces *AtALMT1* expression in the root tips (Kobayashi et al., 2007), which are the most sensitive target of Al rhizotoxicity. Conversely, the expression level in epidermal cells of mature root tissue is greatly reduced, which may avoid unnecessary carbon loss in Al detoxification. Understanding such complex regulatory mechanisms at the molecular level will clarify the true nature of OA excretion in plant stress tolerance.

AtALMT1 is among the most highly up-regulated genes in the roots of Arabidopsis under Al-stressed conditions (Sawaki et al., 2009). Up-regulation of *AtALMT1* is initiated at an early stage (after 1 h of Al exposure) and increases continuously over a longer period (up to 12 h; Kobayashi et al., 2007). A study that combined electrostatic modeling and molecular physiology showed that Al activation of *AtALMT1* expression is sufficiently sensitive to alleviate Al toxicity (Kobayashi et al., 2013b). In addition, histochemical assays using

transgenic plants carrying the GUS reporter gene showed that *AtALMT1* expression was highly induced by Al in the whole root apex but was limited to central cells in the Al-tolerant mature root tissue (Kobayashi et al., 2007). This is likely to optimize protection of sensitive tissue from Al toxicity and minimize carbon loss by malate excretion. These complex but harmonized regulatory mechanisms are achieved by the combined action of multiple transcription factors that regulate expression levels and tissue specificity (Birnbaum et al., 2003). Al though the mechanism of transcriptional regulation has not been completely elucidated, previous studies show that Al activation of *AtALMT1* expression is completely suppressed in the dysfunctional mutant of SENSITIVE TO PROTON RHIZOTOXICITY1 (AtSTOP1; Iuchi et al., 2007). The *stop1* mutant carries a missense mutation in which His is substituted with Tyr at the essential Cys-2- His-2 motif in one of the four zinc finger domains, which indicates that STOP1 may directly bind to the *AtALMT1* promoter and activate transcription. In addition, a recent study has shown that a type of Al-suppressed repressor protein is involved in *AtALMT1* activation by Al (Ding et al., 2013). Coordinated regulation by additional transcription factor(s) is reported in the Al-inducible expression of Al tolerance genes in rice (Oryza sativa) that are regulated by the AtSTOP1 ortholog ALUMINUM RESISTANCE TRANSCRIPTION FACTOR1 (ART1; Yamaji et al., 2009). Expression of *SENSITIVE TO ALUMINUM RHIZOTOXICITY1* (*STAR1*), which encodes a half-type ABC transporter (Huang et al., 2009), requires coordination of the ABSCISIC ACID, STRESS, AND RIPENING5 (ASR5) transcription factor (Arenhart et al., 2014). A similar complex mechanism is likely to be involved in Al-inducible expression of *AtALMT1*.

Identification of cis-elements is a useful approach to analyze complex regulation of gene expression. In planta assays using transgenic plants that carry a deleted promoter:reporter gene construct are often used to map the cis-regulatory elements in the promoter region. In

planta complementation assays, involving transformation of the functional gene driven by the deleted promoters into the mutant background, are also useful to evaluate essential promoter function (Kobayashi et al., 2013a). In addition, several bioinformatic procedures have been developed to predict cis-elements (Tompa et al., 2005; Zou et al., 2011). For example, we previously developed a procedure for cis-element prediction using a microarray dataset that computed the relative appearance ratio (RAR) of the octamers (i.e. the frequency of a particular octamer in the grouped genes relative to that in the genome-wide genes) as a predictive index (Yamamoto et al., 2011b). Using this approach to identify overrepresented octamers in the promoter of salt-inducible genes, which were identified from microarray analysis, we successfully predicted the promoter regions containing experimentally validated cis-elements in the promoter of *RESPONSIVE TO DESSICATION 29A* (*RD29A*). *RD29A* is among the best characterized promoters of salt-inducible genes in Arabidopsis (Narusaka et al., 2003). Combination of in planta reporter assays and this bioinformatic approach is useful to identify the important regions of the *AtALMT1* promoter that regulate efficient response to Al exposure.

In this study, I analyzed the Al-responsive region of the *AtALMT1* promoter by integrating bioinformatics and molecular biological approaches. Overrepresented octamers in gene groups induced or suppressed by Al in the stop1 mutant enabled identification of several candidate regions in the *AtALMT1* promoter. Further analyses of these regions using GUS reporter assays clarified the complex regulation of *AtALMT1*, which involves the STOP1-binding site and interaction with repressors and activators.

2-2. MATERIALS AND METHODS

Plant Materials

Arabidopsis (Arabidopsis thaliana) accession Col-0 (JA58) was obtained from the RIKEN BioResource Center (http://en.brc.riken.jp/index.shtml). The T-DNA insertion mutant of *AtALMT1*, designated *AtALMT1-KO* (SALK_009629), was obtained from the Arabidopsis Biological Resource Center (https://abrc.osu. edu). T-DNA insertion lines of *CAMTA1* (SALK_008187), *CAMTA2* (SALK_007027), and *CAMTA3* (SALK_001152) were also obtained from the Arabidopsis Biological Resource Center (Fig. S2-4). Transgenic Arabidopsis lines carrying *AtALMT1* in the *AtALMT1-KO* background used in the in planta complementation assay, and those carrying GUS in the Col-0 background for the promoter GUS-reporter assay, were generated using the Agrobacterium tumefaciens-mediated floral dip method (Clough and Bent, 1998). *AtALMT1* driven by the *AtALMT1* promoter of different lengths (–1,900, –1,200, –540, –317, –292, and –200 bp from ATG) were transformed into *AtALMT1-KO*, and *GUS* regulated by the mutated promoter of *AtALMT1* (Fig. 2-2A) was transformed into Col-0. All vectors were constructed by insertion of the DNA fragments obtained by overlap extension PCR (Horton et al., 1989) into the T-DNA of pBE2113. The fragments consisted of the *AtALMT1* promoter (deleted or mutated), the coding DNA sequence of the *GUS* or *AtALMT1* open reading frame, and 980 bp at the 39 end of AtALMT1. The sequences of the primers used are shown in Table S4. The overlapping extension PCR was carried out using PrimeSTAR Max high-fidelity Taq polymerase (Takara Bio). A hypervirulent strain of A. tumefaciens (GV3101) was used for transformation. The T2 generation of each line was used for the assays.

Growth Conditions for in Planta Complementation and Reporter Expression Assays

Arabidopsis seedlings were grown hydroponically in accordance with the method described by Kobayashi et al. (2007) in modified MGRL nutrient solution (Fujiwara et al., 1992) supplemented with 200 uM CaCl2 and one-fiftieth strength of other nutrients except inorganic phosphorus (excluded) in the presence or absence of 5 uM AlCl3 at an initial pH of 5.0 adjusted with HCl. For the in planta complementation assay of Al tolerance, about 20 seedlings were grown in the control (0 Al) and Al-toxic (5 uM Al) solutions. The solutions were refreshed every 2 d. Root length was measured on day 5, and the 10 highest values (to exclude uncontrollable late-germinated seedlings) were used for evaluation of Al tolerance. For GUS reporter expression analyses with Al treatment, seedlings were pregrown in the control solution for 10 d, and then the roots were placed in Al- toxic solution containing 10 uM (pH 5.0) for 6 or 24 h. The seedlings were incubated at 22°C 6 2°C under a 12-h-light/12-h-dark photoperiod, with light supplied at a photosynthetic photon flux density of 37 mmol m^{-2} s^{-1}. Staining of GUS was carried out with hydroponically grown seedlings as described previously (Kobayashi et al., 2013a). Briefly, 5-d-old seedlings were treated with or without Al in MGRL solution (pH 5.0) for 24 h and then stained with staining solution (1.0 mM X-glucuronide, 0.1 M sodium phosphate buffer [pH 7.0], 10 mM EDTA [pH 8.0], 0.5 mM potassium ferricyanide [pH 7.0], 0.5 mM potassium ferrocyanide [pH 7.0], 0.3% [v/v] Triton X-100, and 20% [v/v] methanol) for 30 min (Al; Fig. 2-3A) or 60 min (no Al; Fig. 2-3B) at 37°C.

Prediction of cis-Acting Elements in the AtALMT1 Promoter

The RAR of the octamer unit of the *AtALMT1* promoter was calculated using the method described by Yamamoto et al. (2011b). Briefly, Al-inducible and -suppressible genes in the stop1 mutant were identified from microarray datasets. Each of the 222 and 266 genes, respectively, were grouped as Al-inducible genes on the basis of the fold change (+Al/no Al, .3) of microarray data obtained after treatment with 10 uM Al for 6 or 24 h. Two hundred forty-nine genes were grouped as suppressed genes in the stop1 mutant on the basis of the fold change of microarray data (Col-0/stop1) after 10 uM Al treatment for 24 h. All microarray experiments were carried out using the Agilent Arabidopsis oligoDNA chip (Agilent Technologies) as described previously (Sawaki et al., 2009). The RAR was calculated as the ratio of the frequency of each octamer unit in the promoter of the grouped genes to that in the promoters of genome-wide genes. The promoter was defined as –1,000 bp from the TSS reported in the Plant Promoter Database (ppdb; http://ppdb.agr.gifu-u.ac.jp; Hieno et al., 2014). The RAR value of each octamer unit was plotted on the 0 to –540-bp region of the *AtALMT1* promoter and statistical significance ($P < 0.05$) was assessed with Fisher's exact test.

The significantly overrepresented octamer units (RAR >3, $P < 0.05$) were defined as cis-A to cis-H with collocated (>5-bp interval) octamer units with RAR greater than 3. The position of the REGs, TSS, and core promoter elements in *ALMT1* were determined from the ppdb. A consensus sequence for the same gene groups was independently computed with the Melina II tool using the Gibbs sampler method (Okumura et al., 2007). These data are shown in Fig. 2-2A.

RNA Extraction, Real-Time Quantitative Reverse Transcription-PCR, and 5' RACE

Total RNA was isolated using Sepasol-RNA I Super G (Nacalai Tesque) in accordance with the manufacturer's instructions. Total RNA was reverse transcribed with ReverTra Ace (Toyobo, Osaka). Real-time reverse transcription-PCR (except the experiment shown in Fig. 2-8) was performed with SYBR Premix Ex Taq II (Takara Bio) and the Thermal Cycler Dice Real Time System II (Takara Bio) following the manufacturer's instructions using gene-specific primer pairs (Table S4). The transcript levels were quantified with the standard curve method using a complementary DNA dilution series as described by Bustin et al. (2009). Quantification of *AtALMT1* transcripts with a different TSS (Fig. 2-8) was carried out by the standard curve method using Taqman probe with Premix Ex Taq (Probe qPCR; Takara Bio). The standard curve was developed with accurately quantified plasmid DNA (subcloned promoter in the pMD20 vector). The copy number of transcripts of each TSS was calculated arithmetically. In all experiments, transcript levels of *AtALMT1* and *GUS* were normalized against *UBQ1* (At3g52590). Contamination of genomic DNA in the sample was checked by performing the same reactions without reverse transcription, and the amplification efficiency of primers was checked for all primers. The 5' RACE of *AtALMT1* was carried out as previously described by Kihara et al. (2003). Reverse transcription was carried out with SuperScript III Reverse Transcriptase (Life Technologies) using gene-specific primers (Table S4). Amplicons derived from 5'-RACE were subcloned into pMD20 (Takara Bio) and then sequenced using the BigDye Terminator v3.1 Cycle Sequencing Kit with an ABI PRISM 3100 Genetic Analyzer (Applied Biosystems) in accordance with the manufacturer's recommended protocols.

In Vitro Protein-dsDNA Interaction Assay

The amplified luminescence proximity homogeneous assay was used to determine the interaction of AtSTOP1 and dsDNAs designed from the *AtALMT1* promoter. The FLAG (DYKDDDDK)-tagged AtSTOP1 proteins were synthesized using an in vitro transcription/translation system (BioSieg). The protein quality (i.e. efficient synthesis with the expected molecular mass) was confirmed by a western- blotting analysis using anti-FLAG (Wako Pure Chemical Industries) in accordance with the manufacturer's recommended protocols. Both biotinylated and control (nonbiotinylated) DNA oligos were obtained from supplier and used to synthesize dsDNAs. The donor and acceptor beads for the AlphaScreen detection were coated with the anti-FLAG antibody and with streptavidin, respectively. The beads were labeled with the STOP1 FLAG-tagged proteins or the biotinylated dsDNA-oligo(s) using the AlphaScreen FLAG (M2) Detection Kit (PerkinElmer) in accordance with the recommended protocols. The labeled beads were mixed in reaction buffer comprising 25 mM HEPES-KOH (pH 7.6), 40 mM KCl, 0.01% (w/v) Tween 20, and 0.1% (w/v) bovine serum albumin and incubated for 3 h at 22°C. Competitive assays to characterize the STOP1 binding sites were performed by adding mutated dsDNA-oligos to the reaction buffer containing the biotinylated dsDNA-oligo-labeled acceptor beads. The AlphaScreen signals (chemiluminescence between the donor and the acceptor beads conjugated by the binding of labeled STOP1 and dsDNA-oligo) were determined with the Enspire Multimode plate reader (PerkinElmer). The AlphaScreen signals for the control (nonbiotinylated) dsDNA- oligos in the labeling step were used for estimation of the background luminescence. Relative AlphaScreen signals were defined as the ratio of luminescence of the biotinylated dsDNA-oligos to the background.

2-3. RESULTS

Activity of the AtALMT1 Promoter in Al Tolerance and *AtALMT1* Expression

Activity of the *AtALMT1* promoter in Al tolerance was examined by means of an in planta complementation growth assay of transgenic *AtALMT1-knockout* (KO; atalmt1) lines carrying *AtALMT1* driven by a 5' deleted promoter series (from –1,900 to –200; Fig. 2-1, A and B). Growth of the transgenic line carrying *AtALMT1* driven by the –1,900 promoter was comparable to that of the wild-type ecotype Columbia (Col-0), but more extensive deletion of the 5' end of the promoter altered the degree of growth recovery. Deletion to –1,220 slightly improved growth (but not significantly; Fig. 2-1B) compared with that of the –1,900 promoter, which accounted for previous identified position of the localization of the cis-element binding with the WAKY46 repressor (Ding et al., 2013). Growth of the deletion line driven by the –540 promoter slightly decreased compared with that of the –1,900 promoter and was comparable to that of the wild type. The shorter promoters (–317, –292, and –200) than the –540 promoter could not recover Al tolerance in *AtALMT1-KO*. These results indicated that the promoter region from 0 (ATG) to –540 included critical factors that recover Al tolerance of *AtALMT1-KO*.

Expression levels of *AtALMT1* in the transgenic complemented lines were quantified by real-time quantitative PCR after Al treatment for 24 h using primer pairs that did not amplify any amplicons in the *AtALMT1-KO* lines (Fig. 2-1C). The *AtALMT1* expression level with the –540 promoter was comparable to that of the wild type and was decreased in the transgenic plants carrying the –317 promoter. Expression was negligible in the transgenic lines carrying *AtALMT1* driven by the –292 promoter. Taken together, these findings suggested that the

promoter region between –540 and 0 contained critical cis-element(s) that determine Al tolerance through *AtALMT1* expression.

Identification of Potential Promoter Regions Involved in Al-Activated and STOP1 Regulated Expression of *AtALMT1*

The RAR of octamers was plotted for the 0 (ATG) to –540 region of the *AtALMT1* promoter. A high RAR value indicated that the octamer sequence at the plotted position of the *AtALMT1* promoter was overrepresented in the promoter of Al-responsive gene groups identified by microarray experiments relative to the genome-wide promoters (Yamamoto et al., 2011b). Given that *AtALMT1* expression was highly up- regulated in response to Al treatment and was strictly regulated by the STOP1 zinc finger transcription factor, this analysis was carried out using groups of genes up- regulated by Al (after 6- and 24-h treatment) and suppressed in the stop1 mutant compared with the wild type (Fig. 2-2, A and B). In total, eight peaks (A–H; RAR . 3) were identified from the promoter scanning analysis. Except for peak G, all other peaks contained octamers that were detected under at least one condition and with statistical significance (P , 0.05, Fisher's exact test; Table S1). These peaks consisted of eight (peak G) to 15 bases (peak F). Some of the peaks (B, E, and F) contained previously identified octamers, which were predicted to be octamers related to potential cis-regulatory elements (regulatory element groups [REGs]; Yamamoto et al., 2007) based on analysis of the local distribution of octamers for the genome-wide promoters (Fig. 2-2A, blue line). Some of the peaks contained known motifs that were previously identified as cis-elements, of which some are targeted by particular transcription factors (Fig. 2-2B). Putative cis-elements in the core promoter were not detected by my method, whereas TATA boxes and a Y-patch (Y for pyrimidine) motif have been identified by other methods (Fig. 2-2A, blue and green lines).

Three transcription start sites (TSSs) were identified by 5' RACE (Fig. S2-1), which were localized at –84, –138, and –185 bp from ATG (Fig. 2-2A, orange line). Two of the TSSs were associated with putative TATA boxes. Identification of these multiple factors was consistent with the wide dynamic range of *AtALMT1* expression.

Characterization of the Predicted Promoter Regions for *AtALMT1* Expression

Eight RAR peak regions in the *AtALMT1* promoter (Fig. 2-2A) were characterized using transgenic plants carrying the GUS reporter gene driven by the mutated promoters. To inactivate these detected regions, the represented octamer (highlighted in bold in Fig. 2-2B) was mutated in the –1,110 AtALMT1 promoter (designated native promoter [NP]). Activities of the mutated promoters were evaluated by monitoring *GUS* expression by real-time quantitative PCR in the transgenic plants after Al treatment for 24 h (Fig. 2-3A). Mutation caused different expression patterns compared with the NP other than peak G position. This suggested that most of predicted positions contained functional cis-elements that regulate *AtALMT1* expression. Transcript levels of the mutated cis-B were significantly higher than that of the NP in the control treatment (no Al), whereas its transcript levels in the Al treatment showed no significant difference. This result suggested that the region may be a repressor binding site. The *GUS* transcript levels of the mutated cis-A, cis-C, and cis-H were decreased in the Al treatment, whereas they maintained similar levels of transcription in the control. This finding suggested that these regions contained cis-element(s) required for Al activation of the promoter. Mutation of cis-D, cis-E, and cis-F caused reduction of *GUS* transcript levels in both the control and Al treatment. This result suggested that these regions contained cis-binding sites that are essential for maintaining basal transcription in the control treatment and Al-activated transcription, although the degree of suppression differed. Among the cis-D, cis-E, and cis-F regions, the cis-D region was indicated to

contain the most critical factor for both transcription in the control and under Al exposure, and the mutation of this site reduced transcription of the NP less than 10^{-3}. Thus, the results indicated that the cis-D region is essential for transcription of *AtALMT1*.

Positions within the promoter associated with Al activation (i.e. corresponding to the cis-A, cis-C, cis-D, cis-E, cis-F, and cis-H regions) were further characterized by determining the relative expression level of *GUS* after 6 h of Al treatment (Fig. 2-3B). Mutation of the cis-A and cis-C regions did not cause a significant dif- ference in *GUS* transcript levels with the NP, whereas the mutated cis-H and other lines showed significantly lower *GUS* transcript levels than the NP under Al treatment. These results suggested that the cis-A and cis-C regions may be associated with a transcription factor inducible by Al after 6 h of exposure.

Profiling of *AtALMT1* Expression by GUS Staining

To further characterize the peak regions other than peak G to the *AtALMT1* expression, root apices were subjected to histochemical staining for GUS activity (Fig. 2-4, A and B). After exposure of the root tip to Al for 24 h, almost all of the transgenic lines carrying *GUS* driven by the mutated *AtALMT1* promoters (mutation in the cis-A, cis-B, cis-C, cis-E, and cis-H regions) showed a similar staining profile to that of NP transgenic plants. Thus, these mutations did not notably alter the cellular speci- ficity of *GUS* expression in the root tip. Mutation in cis-F caused inactivation of expression in the root tip, which indicated that the cis-element in the F region regulated cellular-specific expression in the root tip. Mutation in the cis-D region completely inactivated expression in all root cells and thus induced severe suppression of expression (Fig. 2-4A). Mutation in the cis-B region caused positive GUS staining in the control (Fig. 2-4B), whereas the NP did not generate a positive signal. These

results further supported the hypothesis that the cis-B regions contain cis-elements that interact with a repressor.

In Vitro Binding of STOP1 Protein to the Peak cis-D Region

The mutation of the cis-D region almost completely inactivated transcription in the control and Al treatments, which was very similar to expression levels of *ALMT1* in the *stop1* mutant (Iuchi et al., 2007). In addition, this region contains a target sequence of the rice STOP1 ortholog ART1 (Tsutsui et al., 2011). This suggests that cis-D may contain STOP1 binding site(s) that are critical for *AtALMT1* expression. To test this possibility, I analyzed the capacity of STOP1 to bind to the cis-D region using an AlphaScreen system. Four overlapping double-stranded DNA (dsDNA) probes (30 bp; probes 2–5; Fig. 2-5A) were designed that covered the cis-C, cis-D, cis-E, and cis-F regions, –252 to 2331 from ATG, while the probe 1 was designed for the cis-A as the negative control. When these probes were reacted with in vitro-translated STOP1 protein, the highest signal was detected with probe 3 (Fig. 2-5C). The signal of probe 3 was competitively suppressed by the nonbiotin-labeled probe 3, but not by the nonreactive negative control probe (Fig. S2-2). These results indicated that my assay condition could detect specific binding of STOP1 to the probe 3 region.

In a competition assay using 5-bp-mutated probe 3, the STOP1 protein could bind to cis-D (Fig. S2-3). To localize the STOP1 binding position, the unique region of probe 3 (7–26 bp from the 5' end) was analyzed using individual point-mutated probes (designated M7–M26). Twenty probes were designed that included 12 probes (M8–M19) corresponding to the detected octamers at peak D (Table S2-2). The mutagenized probes (non-labeled) were mixed with the biotinylated native probe 3 in a 9:1 ratio, and then the AlphaScreen signals

were compared (Fig. 2-5C). A point mutation at 11 positions significantly increased signal intensity for native probe 3 (black bars in Fig. 2-5C), including six nucleotides in the detected peak D region (underlined; TAAGGGGAGGGC of the predicted peak D; Fig. 2-2B). These results indicated that the STOP1 protein could bind to the cis-D region, which is essential for transcription. These results indicated that the STOP1 protein can bind to a wider range of the promoter region than the cis-D region.

Characterization of Zinc Finger Domains of STOP1

STOP1 carries four Cys-2-His-2 zinc finger domains. The His-to-Tyr point mutation at the second His residue of the first domain is the probable cause of the stop1 mutant, which shows complete suppression of *AtALMT1* expression (Iuchi et al., 2007). To evaluate the impact of this mutation on the binding capacity of STOP1, I performed an AlphaScreen assay using mutagenized proteins. The second His residues were mutated to Tyr in each zinc finger domain; the mutated protein was designated MT_ZF1-4 and used for binding assays with probe 3 (Fig. 2-6A). As I inferred, MT_ZF1 (i.e. originally identified mutated position of stop1 mutant) al- most completely suppressed the binding capacity of STOP1 (less than 0.1 of native STOP1; Fig. 2-6B). MT_ZF2 and MT_ZF4 showed similar levels of suppression of the STOP1 binding capacity, suggesting that these domains are critical for binding to the AtALMT1 promoter. Mutation in ZF3 did not comparably suppress binding, which indicated that this domain may contribute less than other domains to the binding of STOP1 to the AtALMT1 promoter.

Involvement of CAMTA in Activation of the cis-C Region

The cis-C region contained the ACGCGT sequence, which is a consensus of cis-acting elements (CGCG box; [A/C]CGCG[C/G/T]) for the CALMODULIN- BINDING TRANSCRIPTION ACTIVATOR (CAMTA) transcription factor that regulates expression of stress- responsive genes carrying the CGCG box (Yang and Poovaiah, 2002). Using a previously reported microarray dataset (10 mM Al treatment for 24 h; Sawaki et al., 2009), I showed that among major stress-responsive *CAMTA* genes, *CAMTA1* to *CAMTA3* were likely responsive to Al (Table S2-3). The CAMTA genes comprise six homologous genes in Arabidopsis (Finkler et al., 2007). Time course analysis showed that *CAMTA1* and *CAMTA2* were continuously inducible by Al during treatment for 24 h (Fig. 2-7A). A transfer DNA (T-DNA) insertion mutant of *CAMTA2* significantly suppressed Al tolerance in terms of root growth (Fig. 2-7B). In addition, the expression level of *AtALMT1* decreased by about 15% in the camta2 mutant (Fig. 2-7C). I also observed binding activity of CAMTA2 to the CGCG box in the cis-C region in an AlphaScreen assay (probe 2; Fig. 2-7, D and E). Taken together, these results indicated that up-regulation of CAMTA2 is involved in the activation of *AtALMT1* expression, in particular, after 6 h of Al treatment.

Changes in AtALMT1 Transcription of TSSs

The *AtALMT1* promoter possesses two putative TATA boxes. Although mechanisms remain to be clarified, the average number of TATA boxes is significantly higher in strongly stress-responsive genes (e.g. fold change >10; Yamamoto et al., 2011a). To explore this issue in relation o *AtALMT1* transcription, I determined the TSS by 5' RACE and quantified each transcript. The 5' RACE identified three TSSs in the *AtALMT1* promoter. TSS1 and TSS2 were located in the 3' region of the putative TATA1 and TATA2 with intervals of about 20 bp (Fig. S2-1). To quantify each transcript transcribed from the

different TSSs (TSS1–TSS3), three primer pairs and a TaqMan probe were designed (Fig. 2-8A). Transcripts of TSS1 were the most abundant among the transcripts of the three TSSs, which comprised 65% of transcripts in the control and 70% to 75% after 6 and 24 h of Al treatment (Fig. 2-8B). The proportion of TSS2 transcripts was in- creased by Al treatment to about 20% after 24 h Al treatment compared with 5% in the control. By contrast, he proportion of transcripts of TSS3, which is not associated with a TATA consensus, decreased in response to Al treatment. These results suggested that the increase in shorter transcripts, which are associated with the TATA box, was associated with regulation of *AtALMT1* transcription under Al treatment.

2-4. DISCUSSION

Previous studies revealed that transcriptional regulation of *AtALMT1* plays critical roles in the protection of the sensitive root tips of Arabidopsis from Al toxicity (Hoekenga et al., 2006). This process is likely optimized to minimize carbon loss by regulation of expression levels and tissue-specific expression (Kobayashi et al., 2007, 2013a). In this study, I identified several important regions of the *AtALMT1* promoter that control expression levels based on a promoter scanning analysis. The promoter scanning analysis showed that several octamers were overrepresented in the promoter region of *AtALMT1* (Fig. 2-2). Inactivation of seven of the eight octamers altered *AtALMT1* expression under the control condition and Al treatment (Fig. 2-3, A and B). This variety of regulatory mechanisms in the promoter structure is consistent with the complex regulation of *AtALMT1* expression under Al stress. In addition, these elements likely coordinately regulate Al tolerance judged by the growth recovery by the transgenic *AtALMT1-KO* lines carrying *AtALMT1* driven by 5'-deleted promoters (Fig. 2-1). Expression of *AtALMT1* is strongly triggered by Al exposure and increases continuously during 12- to 24-h exposure to over 30 times the expression level of the control (Fig. 2-1C). The broad dynamic range of *AtALMT1* expression may be explained partly by the region (cis-B) that is likely associated with repressor (Fig. 2-3). Inactivation of the region induced expression under control conditions, which indicated that the region maintains a low expression level under the control condition. Several other regions are indicated to regulate Al activation of *AtALMT1* transcription (i.e. increase expression under Al treatment). In addition, *AtALMT1* carried another character of highly inducible genes in possessing multiple TATA boxes, which was identified by genome-wide analysis of the promoter structure in the stress-responsive genes

(Yamamoto et al., 2011a). The combination of these factors would account for the broad dynamic range of up-regulation of *AtALMT1*.

Some of these regions may regulate *AtALMT1* transcription in a time-dependent manner, suggesting that repression of the repressor proteins or induction of activator proteins occurred during Al treatment. WRKY46 was recently identified as a repressor of *AtALMT1*, whereas WRKY46 itself is repressive to Al. Thus, negatively regulated activation plays a role in Al-inducible *AtALMT1* expression (Ding et al., 2013). Conversely, in this study, I found that some cis-acting elements interact with transcription factors inducible/activated by Al (Fig. 2-3). These elements coordinately regulate the Al- responsive expression of *AtALMT1* and Al tolerance. I observed that deletion of the 5' end containing cis-A (i.e. the –317 *AtALMT1* promoter:GUS transgenic plant) resulted in decreased *AtALMT1* expression after 24 h exposure to Al (Fig. 2-1C). However, at 6 h, no change in the *GUS* expression level was observed in the transgenic line carrying the –1,110 *AtALMT1* promoter:GUS construct (Fig. 2-4; Kobayashi et al., 2013a). Conversely, some of the cis-acting elements showed no difference in Al response at both 6 and 24 h (e.g. cis-D, cis-F, and cis-H; Fig. 2-3, A and B). These factors may be activated rapidly by protein phosphorylation/dephosphorylation, which has previously been shown to be a regulatory mechanism of *AtALMT1* expression (Kobayashi et al., 2007). Combination of these mechanisms may minimize expression in the control and enhance expression in a continuously wide range.

One of the cis-acting elements cis-C contained a CGCG box, which is a binding site for the stress-inducible transcription activator CAMTA (Yang and Poovaiah, 2002). Previous studies of CAMTAs indicate that stressinducible expression of specific CAMTAs regulates expression of stress tolerance genes, such as response to pathogen attack (Galon et al., 2008), cold stress (Kim et al., 2013), and drought (Pandey et al., 2013). Combination of *in*

planta promoter:reporter assays and an in vitro protein-DNA binding assay suggested that the Al-inducible CAMTA2 activates *AtALMT1* expression by binding to the cis-C region (Fig. 2-7, D and E). The expression pattern of *CAMTA2* under Al treatment was consistent with the *AtALMT1* expression response. Expression of *CAMTA2* was induced by Al within 6 h (Fig. 2-7A), while inactivation of cis-C (binding site of CAMTA) decreased expression after 24 h, but not 6 h (Fig. 2-3, A and B). Further research on Al-inducible and Al-repressive transcription factors may identify other Al-responsive transcription factors that regulate *AtALMT1* expression.

Among the predicted cis-elements, mutation of the cis-D, cis-E, and cis-F suppressed *AtALMT1* expression to control levels in the promoter:GUS transgenic plants (Fig. 2-3, A and B). In particular, inactivation of cis-D decreased the expression level to less than 10^{-3}, which was similar to the *AtALMT1* expression level in the stop1 mutant under the control condition. An in vitro binding assay indicated that STOP1 binds to cis-D and surrounding regions of the *AtALMT1* promoter (Fig. 2-5). The cis-D sequence contained a previously identified minimum consensus of ART1 in rice (GGNVS; Tsutsui et al., 2011). However, my in vitro analysis with the *AtALMT1* promoter indicated that a wider region of the promoter interacted with STOP1, as 11 nucleotides affected the binding capacity of STOP1. Cys-2-His-2 zinc finger domains often recognize two to four nucleotides for binding (Pavletich and Pabo, 1991; Segal et al., 1999), whereas STOP1 contains four zinc finger domains (Iuchi et al., 2007). The binding assay with mutated STOP1 showed that all four zinc finger domains, including ZF1, which carries the His-to-Tyr substitution of the stop1 mutant, were functional for binding with the dsDNA of the cis-D region (Fig. 2-6). Although ZF3 showed less functionality for binding, these results strongly suggested that a broader region is required for STOP1 binding. Inactivation of the cis-acting elements severely repressed expression of *AtALMT1*, suggesting that STOP1 binding is critical for

AtALMT1 expression. In addition, the fold change (Al/control) was decreased to 5.0 from 22.3, which indicated that STOP1 binding is one factor that regulates *AtALMT1* expression in response to Al exposure.

Inactivation of cis-F altered the tissue-specific expression profile of AtALMT1 (Fig. 2-4A). GUS staining assays showed that inactivation of cis-F completely repressed expression of *AtALMT1* in the root tips and outer tissues (cortex and epidermis) of the mature root. This finding suggested that transcription factor(s) binding to cis-F play critical roles in tissue-specific expression of *AtALMT1*. In the tissues altered by mutation in cis-F tissues, an unknown factor is required for STOP1-dependent expression of *AtALMT1*. It is reported that ART1-regulating Al-responsive expression of *STAR1* in rice requires the ASR5 transcription factor, which is associated with tissue-specific expression in the root tips for binding to the GCCCA sequence in the *STAR1* promoter (Arenhart et al., 2014). Although the Arabidopsis genome does not contain an ASR homolog (Carrari et al., 2004), the same sequence was identified in the cis-F region (GCCCA; Fig. 2-2B). Interestingly, the GCCCA sequence is known to be the target cis-acting element of members of the TEOSINTE BRANCHED1, CYCLOIDEA, AND PROLIFERATING CELL FACTOR (TCP) transcription factor family, which coregulates expression of various genes in meristematic tissues together with other transcription factors (Trémousaygue et al., 2003). Although ASR5 and TCP transcription factors do not show overall similarity, a TCP-type transcription factor may play a role in tissue-specific *AtALMT1* expression in Arabidopsis. Interestingly, promoter scanning analysis using an Arabidopsis dataset (i.e. overrepresented octamers in the promoter of suppressed genes in the stop1 mutant) showed that the *TaALMT1* promoter of wheat contained a set of STOP1- binding motifs and cis-acting elements for CAMTAs and was associated with cis-acting elements for TCP domain transcription factor(s)/ASR5 (Fig. 2-9). An Al-tolerant wheat near-isogenic line (ET8)

contained three sets of STOP1/ CAMTA binding sites and expressed greater levels of *TaALMT1*, whereas an Al-sensitive near-isogenic line (ES8) carried a single set (Sasaki et al., 2006). This suggested that a similar regulatory mechanism, namely combination of STOP1-like protein/root-specific transcription factors, may be conserved in various plant species. Similar events, namely an increase in the number of STOP1/ART1 binding sites, was observed in *Holcus lanatus*, which is naturally adapted to acidic soils (Chen et al., 2013).

In this study, I efficiently identified a series of cis- elements of *AtALMT1* using RAR-based prediction of cis-elements. In planta assay of *GUS* expression validated the accuracy of prediction and indicated that regulation consisted of suppression and activation and that STOP1 binding regulates both the expression level and Al response (Fig. 2-10). In addition, I identified one of the activating transcription factors, CAMTA2, by integration of reverse genetics using T-DNA insertion lines and in vitro protein-DNA binding assays. Further molecular-level research is required to identify other transcription factors that regulate *AtALMT1* expression by the interaction with the remaining predicted cis-elements.

[Tokizawa M, Kobayashi Y, Saito T, Kobayashi M, Satoshi I, Nomoto M, Tada Y, Yamamoto YY, Koyama H (2015) STOP1, CAMTA2 and other transcription factors are involved in aluminum-inducible AtALMT1 expression. Plant Physiol 167: 991-1003]

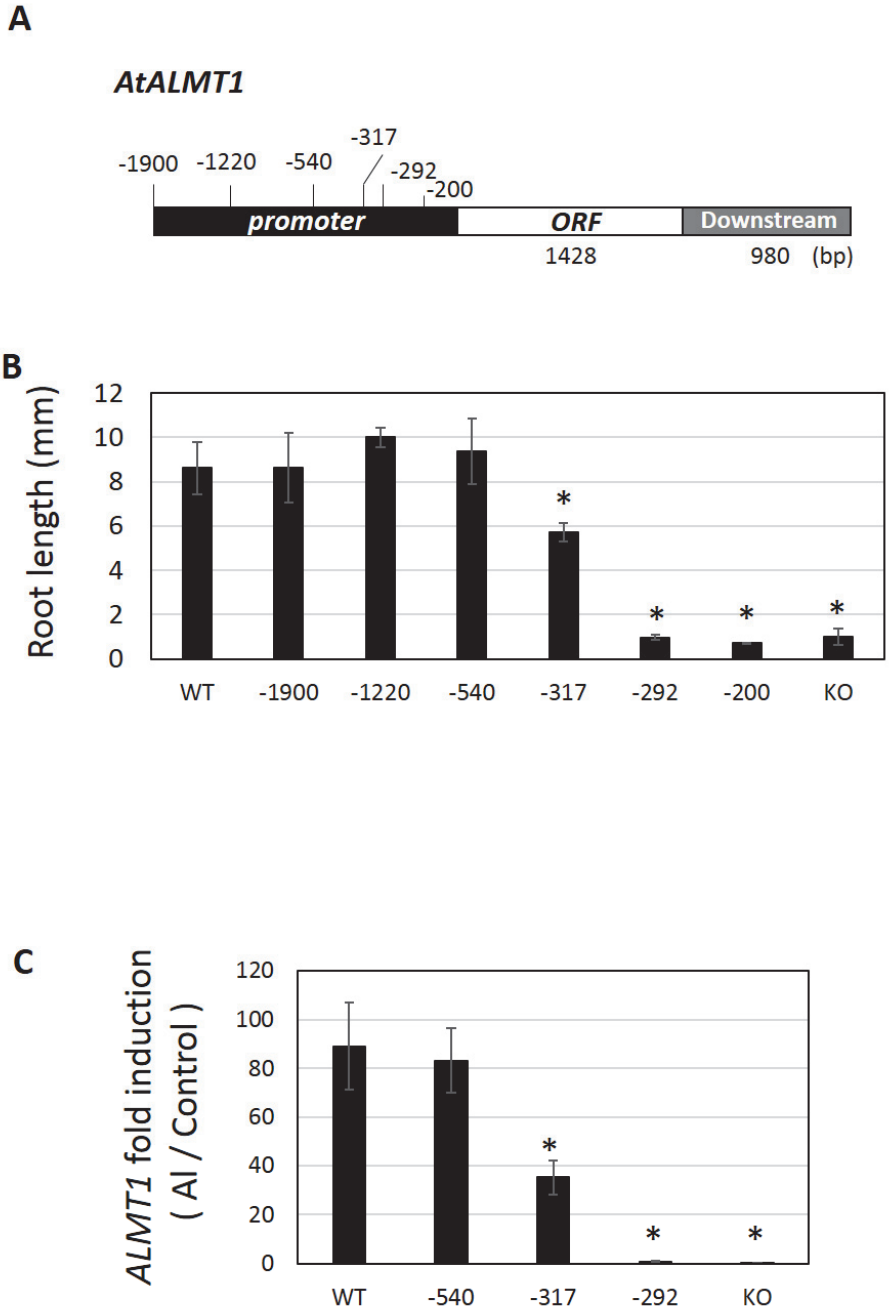
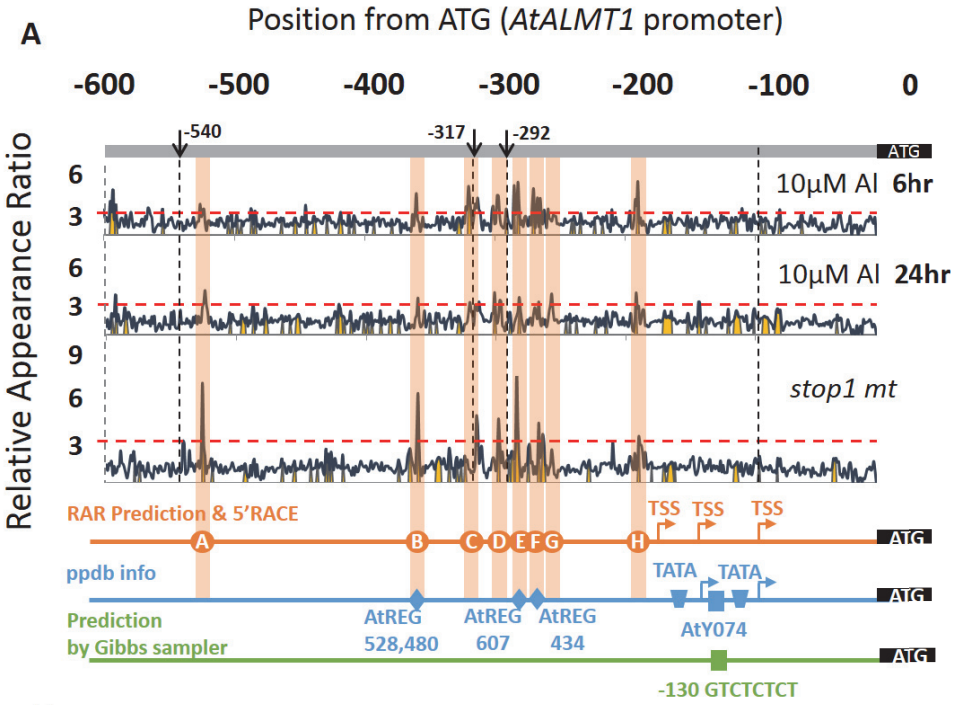


Figure 2-1. *In planta* complementation assay of *AtALMT1* driven by 5' deleted promoters of different lengths.

AtALMT1 carrying different lengths of the promoter were transformed into *AtALMT1*-knockout (KO; *atalmt1*). The position of the 5′ end of the promoter from the ORF is shown in panel A. Root length of transgenic *AtALMT1*-KO carrying *AtALMT1* driven by 5' deleted promoters, wild-type (WT) Col-0 and *AtALMT1*-KO were measured for 5d plants grown in Al toxic solution (4 µM Al, pH 5.0) or control solution (no Al, pH 5.0) (panel B: n=5, means ± SD). Transcript levels of *AtALMT1* were analyzed by real-time quantitative PCR and were normalized with the *UBQ1* expression level. Seedlings were precultured in control solution for 10 d, then the roots were placed in 10 µM $AlCl_3$ (pH 5.0) for 24 h. Fold induction of *AtALMT1* (Al treatment/control) was calculated for three lines (carrying the promoter of length −540, −317, or −292 bp), *ALMT1*-KO and WT. The mean ± SD fold induction of three replications for each line is shown in panel **(C)** Asterisks in panels B and C represent a significant difference ($P < 0.05$) compared with WT.



B

Predicted *cis*	Position From ATG	Sequence	REG *	Motif **
A	-526/-517	**GAGGGCAC**TA		
B	-361/-353	C**CTACCGGG**	ATREG528,480	Hbox
C	-320/-306	**CCTCACGC**GTCGCTC		CGCG
D	-301/-290	TAAG**GGGAGGGC**		
E	-285/-275	CTA**GTGCCCAA**	ATREG607	GCCCA
F	-270/-256	**CTGGGCTA**GGTTCGA	ATREG434	GCCCA
G	-257/-250	**GACTCCGT**		
H	-192/-183	G**TGCAACGC**A		

*ppdb http://ppdb.agr.gifu-u.ac.jp , ** Yamamoto et al., 2011b

Figure 2-2. Relative appearance ratio (RAR) scanning plot for the *AtALMT1* promoter based on the relative appearance frequency calculated from microarray datasets.
(A) The RAR of each octamer was plotted to its 3′-end position in the *AtALMT1* promoter. The Al-inducible genes (fold change [Al/control] > 3) at different time points (treatment for 6 or 24 h with 10 μM Al, pH 5) and the genes suppressed in the *stop1* mutant after 24 h Al treatment (fold change [WT/stop1] < 2.5) were grouped from the microarray data set. The RAR was calculated from the frequency of the octamer in the promoter of the grouped genes relative to that of the 24,956 genome-wide genes. The black lines represent the RAR plots, and yellow-shaded regions represent significantly overrepresented octamers ($P < 0.05$, Fisher's exact test). Promoter regions detected by significantly overrepresented octamers (RAR > 3, $P < 0.05$) are highlighted with vertical bars (designated A to G). Closely associated regulatory element groups (REGs) (predicted from ppdb), octamers of the A to H regions, and the TSS predicted from ppdb are shown below the plots. Positions of TATA boxes and a Y-patch motif predicted by ppdb and by Gibbs sampling using suppressed genes in the *stop1* mutant are shown. **(B)**The position within the promoter of each peak detected in A. Octamers used for mutation analysis in Figure 3 (underlined), the corresponding REG (obtained from ppdb), and the putative motif of *cis*-acting elements are shown.

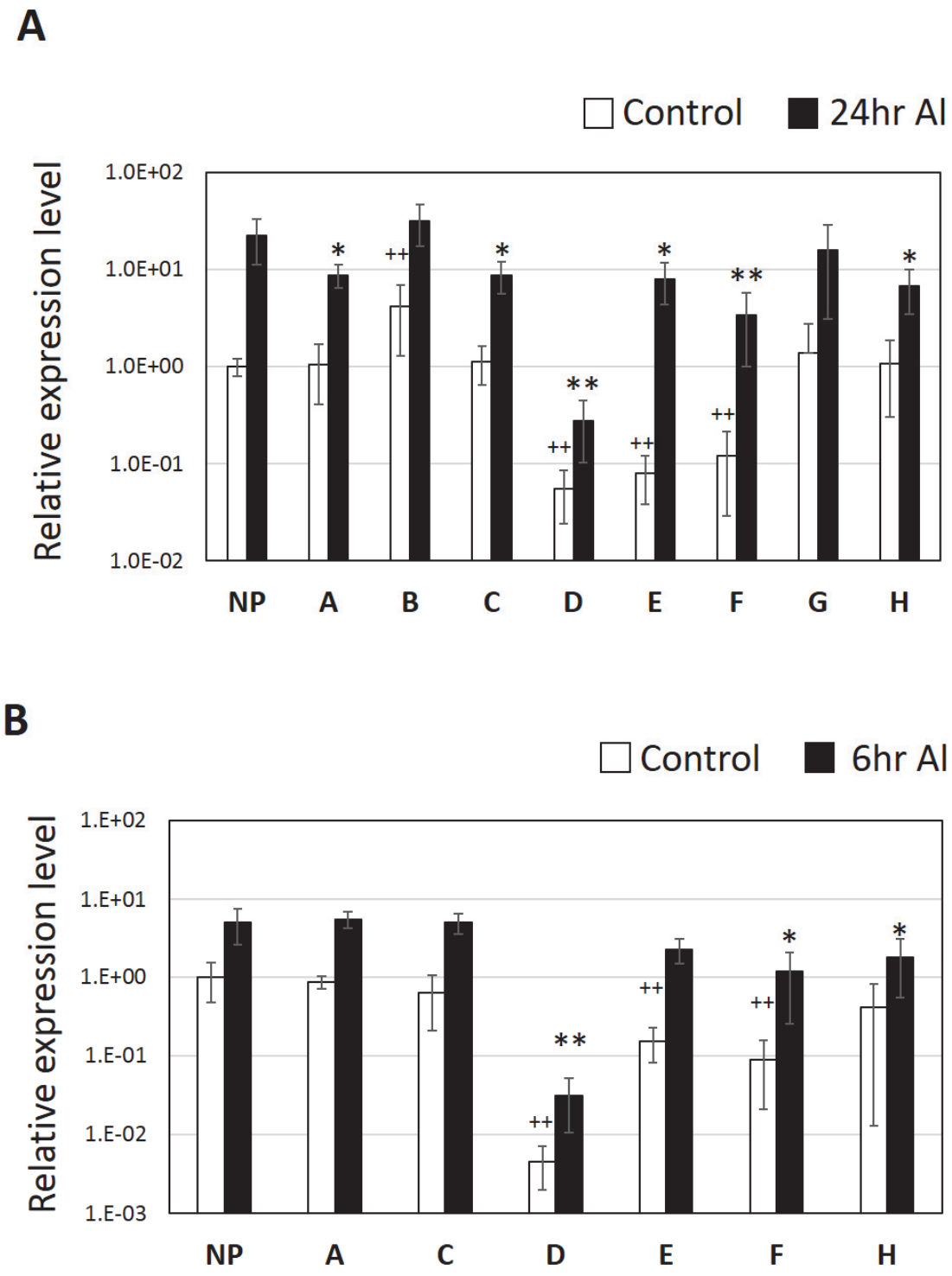


Figure 2-3. Changes in activity of *AtALMT1* promoters carrying substitutions of nucleotides at the position of overrepresented octamers.

Representative octamers in the A to H regions were substituted (see Figure 2b), and the promoter activity was evaluated using transgenic plants carrying the *GUS* reporter gene driven by the substituted promoter. The *GUS* reporter expression was quantified in the control (−1,100 from ATG) and the substituted promoter lines by real-time quantitative PCR. NP indicates the non-mutated promoter. Relative expression levels (*GUS/UBQ1*) in the control (no Al) solution (white bars) and in 10 μM Al solution (black bars) are shown after treatment for 24 h (A) and 6 h (B). The mean ± SD values of three replications are shown. Asterisks indicate a significant difference from the relative expression level of the control transgenic lines (Student's *t* test; * or +, $P < 0.05$; ** or ++, $P < 0.01$).

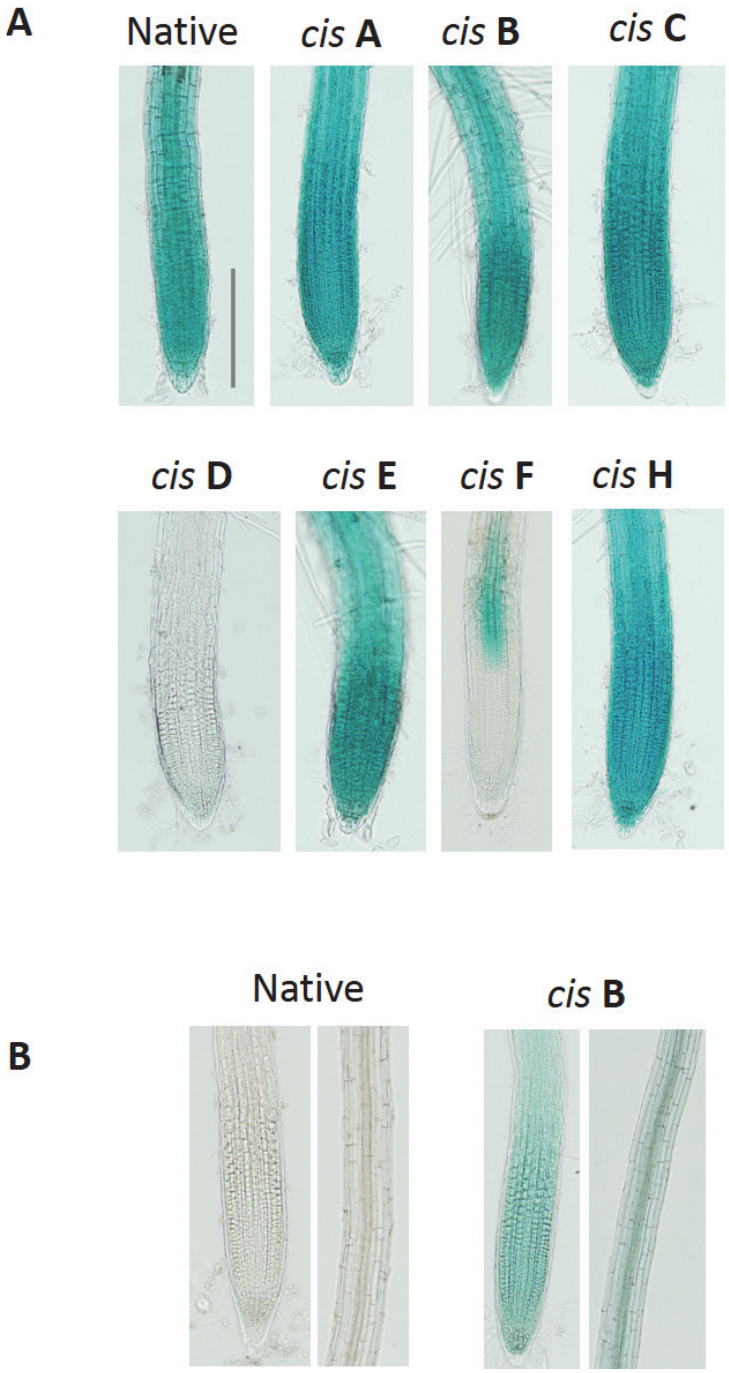
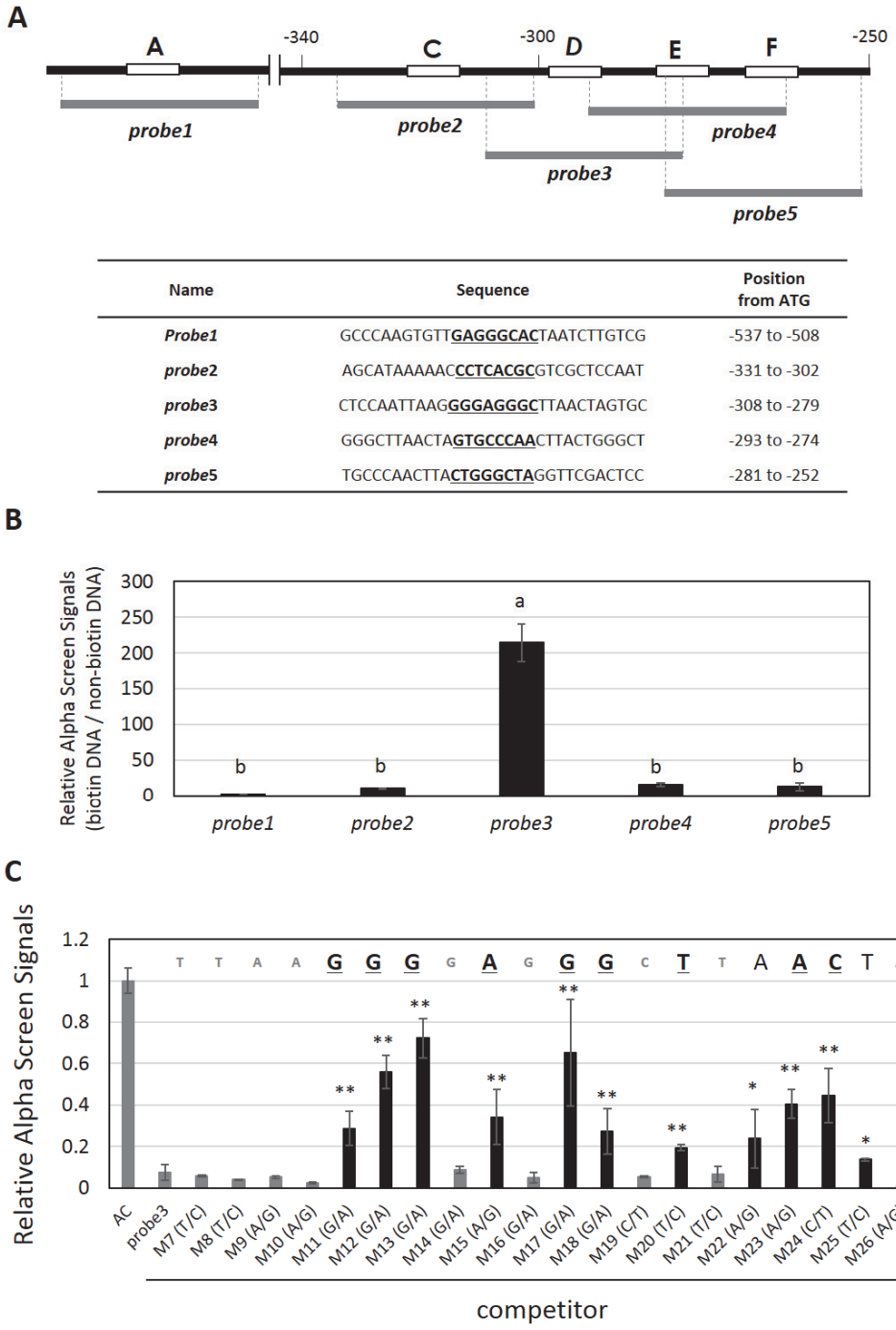


Figure 2-4. Histochemical analysis of GUS expression in the transgenic plants carrying *AtALMT1* promoter:GUS.
GUS staining was carried out 30–60 min after incubation in 10 μM Al solution (pH 5.6) for 24 h (A) or control solution (no Al, pH 5.6; B). Native and *cis*-A to -H (mutated in the regions *cis*-A to -H) were identical to the transgenic lines used in Figure 3. Identical results were confirmed in at least three independent experiments. Bar indicates 20 μm.

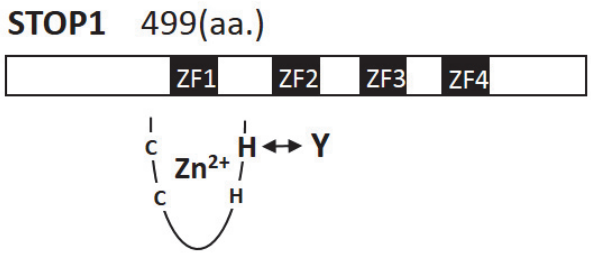


Name	Sequence	Position from ATG
Probe1	GCCCAAGTGTT**GAGGGCAC**TAATCTTGTCG	-537 to -508
probe2	AGCATAAAAAC**CCTCACGC**GTCGCTCCAAT	-331 to -302
probe3	CTCCAATTAAG**GGGAGGGC**TTAACTAGTGC	-308 to -279
probe4	GGGCTTAACTA**GTGCCCAA**CTTACTGGGCT	-293 to -274
probe5	TGCCCAACTTA**CTGGGCTA**GGTTCGACTCC	-281 to -252

Figure 2-5. *In vitro* binding assay of double-stranded DNA and the STOP1 protein using an AlphaScreen system.

A. *In vitro* translated STOP1 protein labeled with the accepter beads of the AlphaScreen system was incubated with the 30 bp double-stranded DNA. B. Relative AlphaScreen signals were calculated as the ratio of AlphaScreen signals of the reactive probe (biotin labeled) to those of the non-reactive probe (non-biotin labeled) in the presence of the labeled STOP1 protein and streptavidin-coated donor beads. Values are the mean ± SD (n = 3). Different letters above the bars indicate a significant difference ($P < 0.05$, Tukey's test). C. Competitive assays of the probe3 region with the single nucleotide mutagenized probes. The reactive probe3 (see B) was incubated with the labeled STOP1 protein in the presence of non-labeled probe3 or the probe that carried a single-nucleotide substitution. Relative values ± SD ($n > 3$) were calculated as the ratio of the value obtained in the absence of the competitor (AC). Asterisks indicate a significant higher than the relative AlphaScreen signals of non-reactive probe3 (Student's t test; *$P < 0.05$, **$P < 0.01$).

A



	Position (aa.)	Amino acid substitution		
MT_ZF1	266	H (CAT)	→	Y (TAT)
MT_ZF2	328	H (CAC)	→	Y (TAC)
MT_ZF3	355	H (CAC)	→	Y (TAC)
MT_ZF4	385	H (CAC)	→	Y (TAC)

B

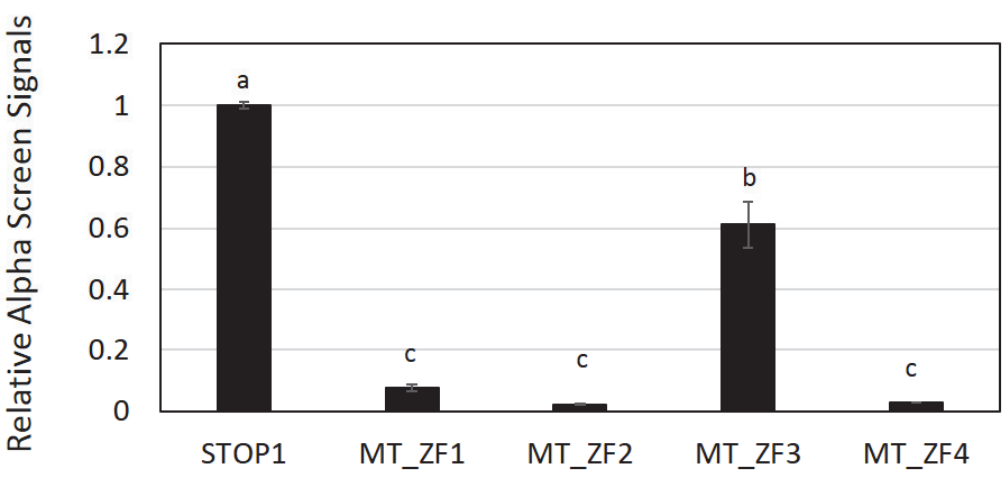


Figure 2-6. Characterization of the capacity of zinc-finger domains of STOP1 to bind to the *AtALMT1* promoter.

A. His (H) to Tyr (Y) mutations were introduced to four Cys2Hys2 zinc finger domains. The capacity to bind to probe 3 (see Figure 5) was analyzed with an AlphaScreen system. B, Relative luminescence intensity of the labeled probe3 and STOP1 proteins (native STOP1 and mutated proteins, MT_ZF1 to 4). Values are the mean ± SD (n = 3) relative to native STOP1 protein. Different letters above the bars indicate a significant difference ($P < 0.05$, Tukey's test).

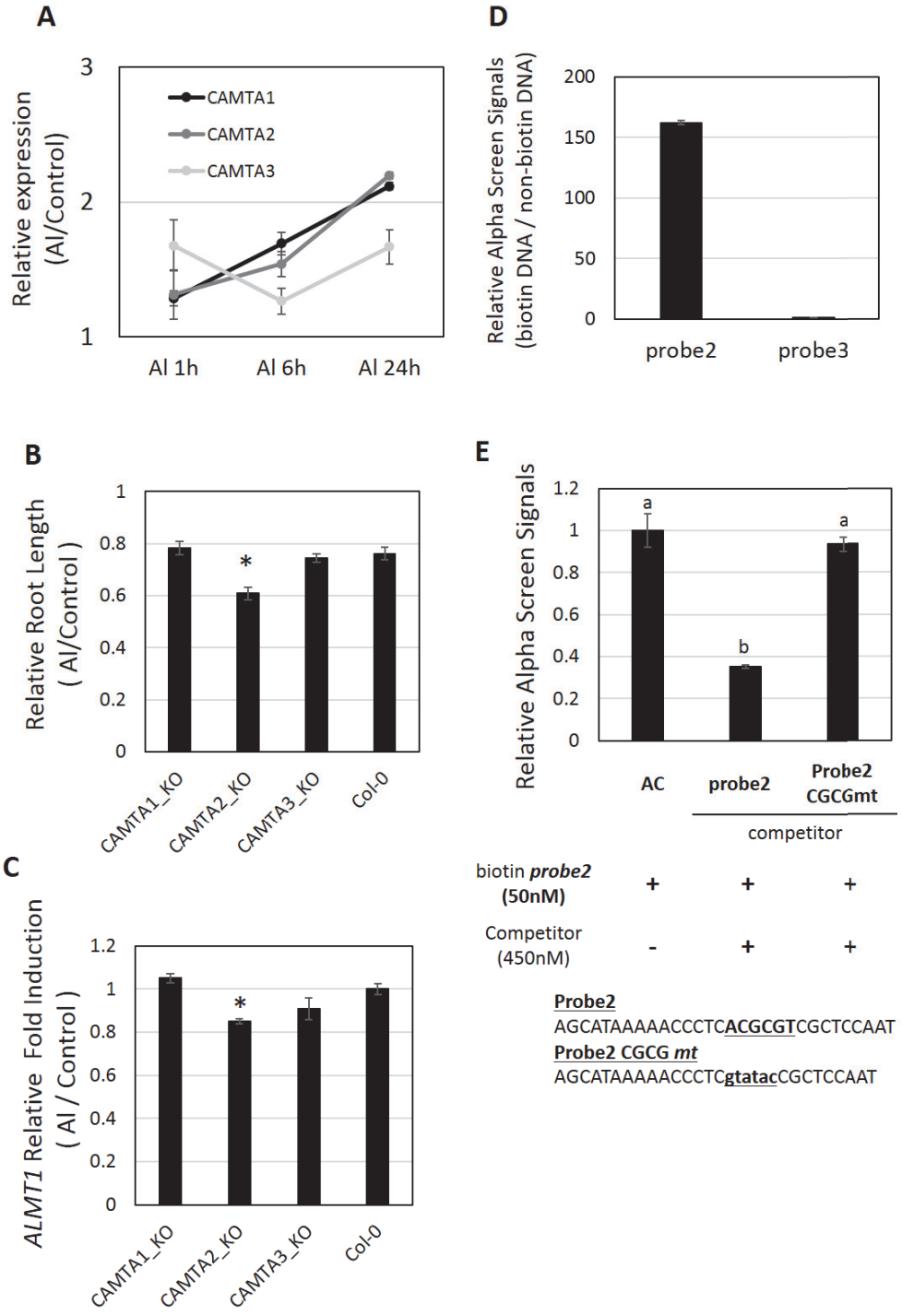


Figure 2-7. Characteristics of Al-responsive CAMTAs in *AtALMT1* expression and Al tolerance of arabidopsis.

A. Expression of Al-responsive *CAMTAs* (*1*, *2*, and *3*) were quantified by reverse-transcription real-time quantitative PCR after exposure to 10 μM Al solution (pH 5.0). Values are the mean ± SD expression level relative to the control (no Al, pH 5.0). B., C. Relative root growth (Al/control) in 5-day-old seedlings (with or without 5 μM Al, pH 5.0, $n = 10$) (B) and expression of *AtALMT1* quantified after incubation in 10 μM Al (pH 5.0) for 24 h ($n = 3$) (C). Values are the mean ± SE (B) and SD (C), and asterisks indicate a significant difference relative to Col-0 (Student's t test, $P < 0.05$). D., E. AlphaScreen signals in the binding assay for probe2 (containing CGCG-box) and probe3 (see Figure 5) with the CAMTA2 protein (D) and those in the competitive assay using the mutagenized probe2 (E). Different letters above the bars indicate a significant difference ($P < 0.05$, Tukey's test).

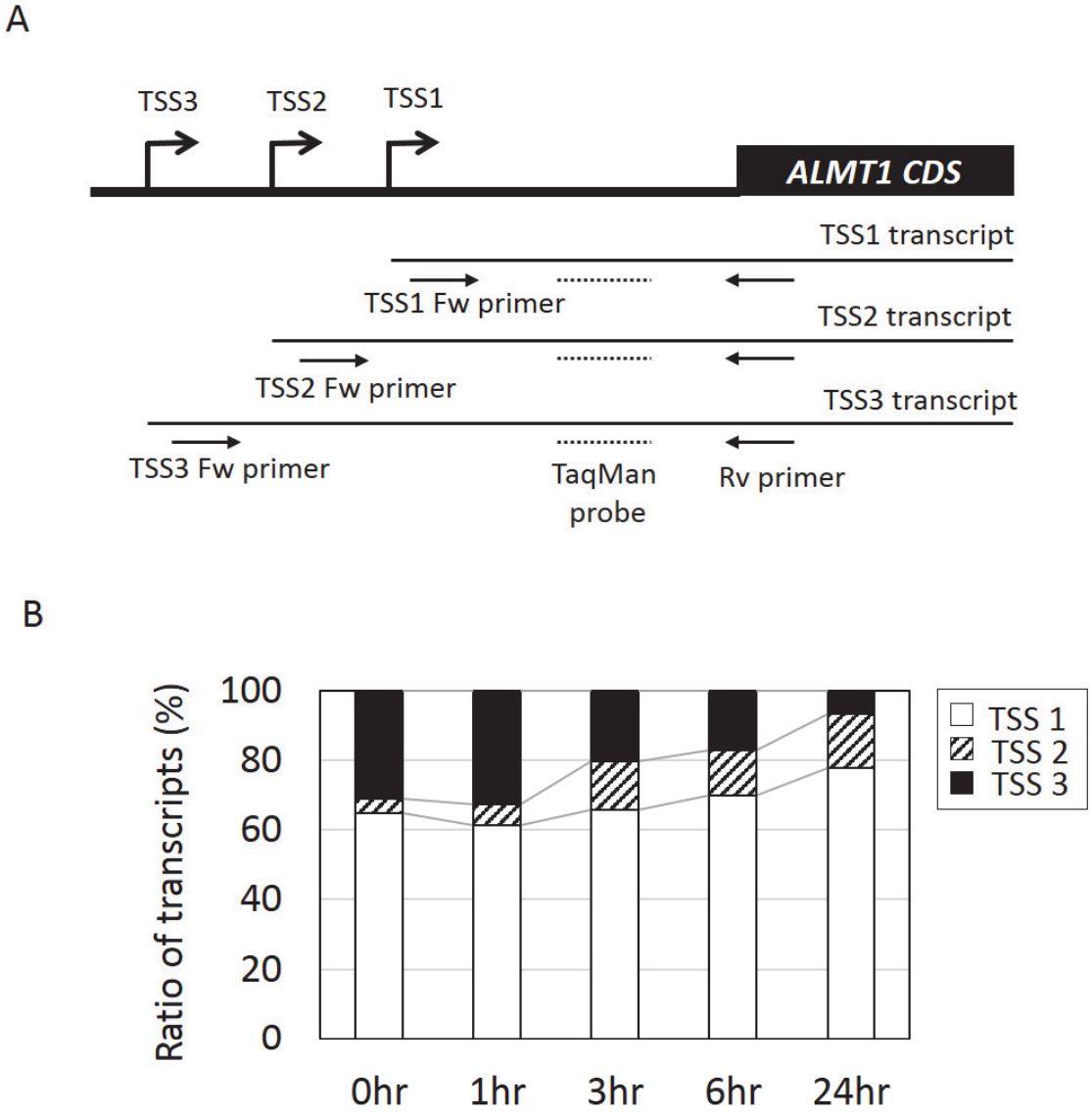


Figure 2-8. Relative amounts of *AtALMT1* transcripts that carried different lengths of the 5′ untranslated region.

A. Transcripts of *AtALMT1* were quantified by quantitative reverse-transcription PCR using different primer pairs and the TaqMan probe to quantify *TSS1-3* (*TSS1* primer pair), *TSS2* and *3* (*TSS2* primer pair), and *TSS3* (*TSS3* primer pair). B, Relative proportions of *TSS1*, *2* and *3* transcripts at different time points during treatment with 10 μM Al (pH 5.0) for 24 h.

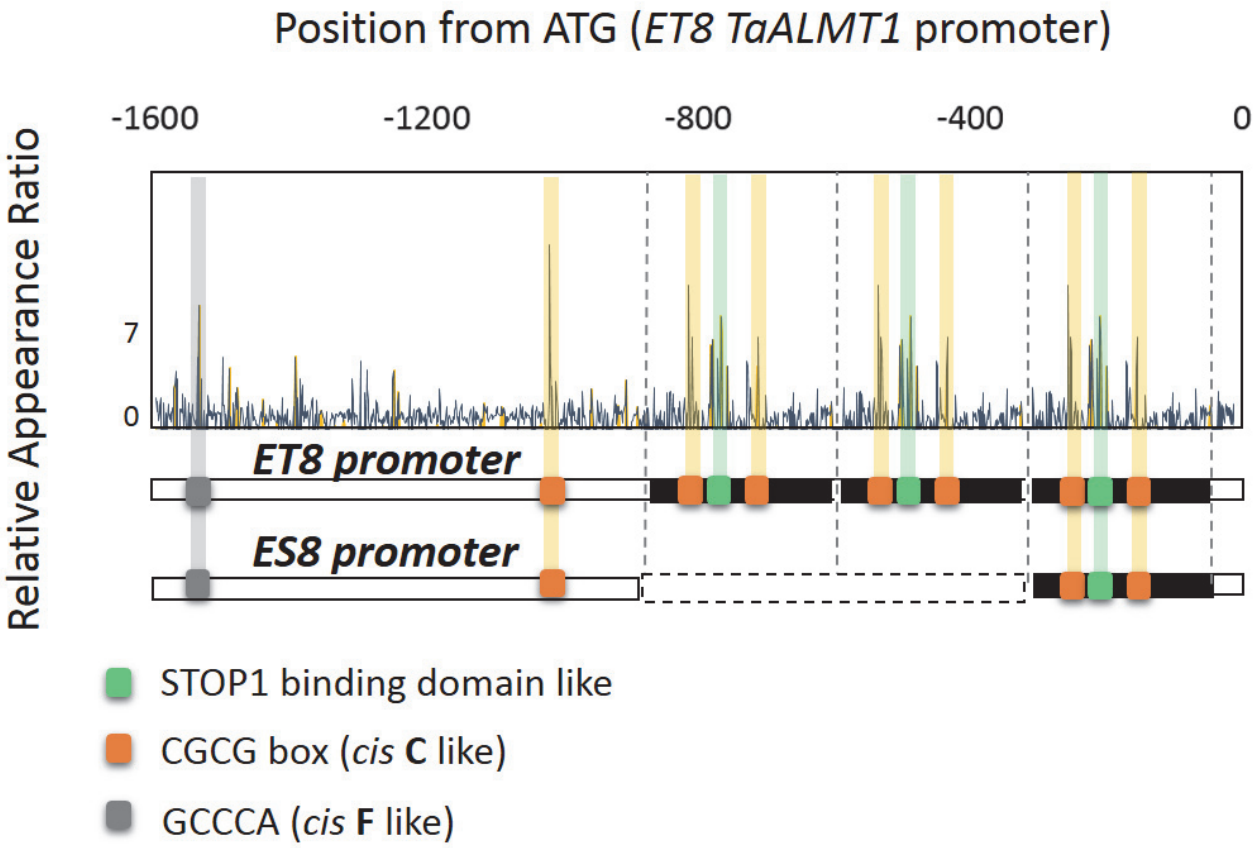


Figure 2-9. Promoter scanning analysis of the *ALMT1* promoter of wheat (*TaALMT1*) near-isogenic lines that carried different levels of *ALMT1* expression (ET8 and ES8).
RAR values calculated from the Arabidopsis data (suppressed genes in the *stop1* mutant in response to Al treatment) were plotted for the promoters of ET8 and ES8. Putative STOP1-binding (green) and peaks *cis*C-like (CGCG-box, orange) and *cis*F-like (GCCCA, gray) sequences are indicated.

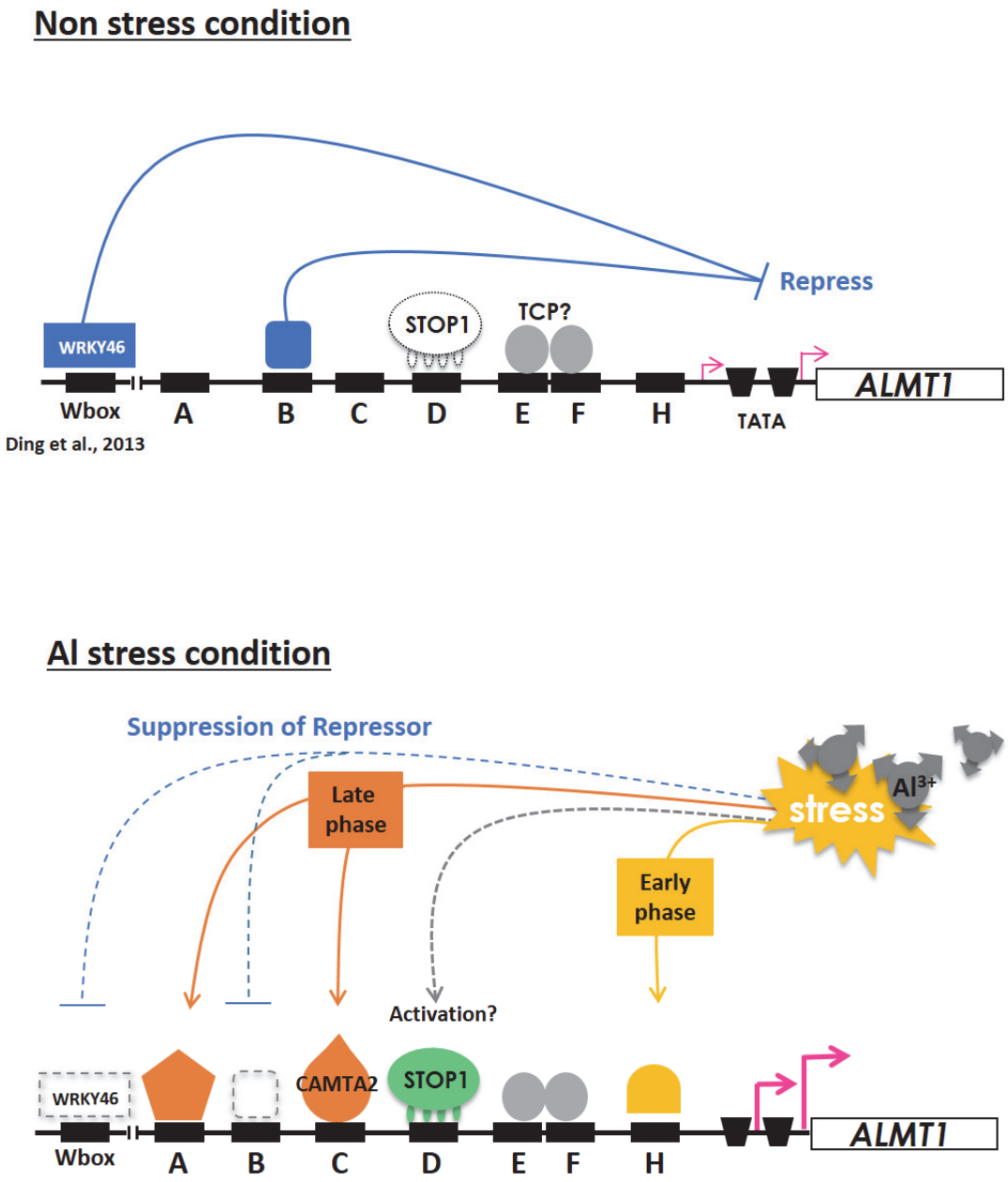
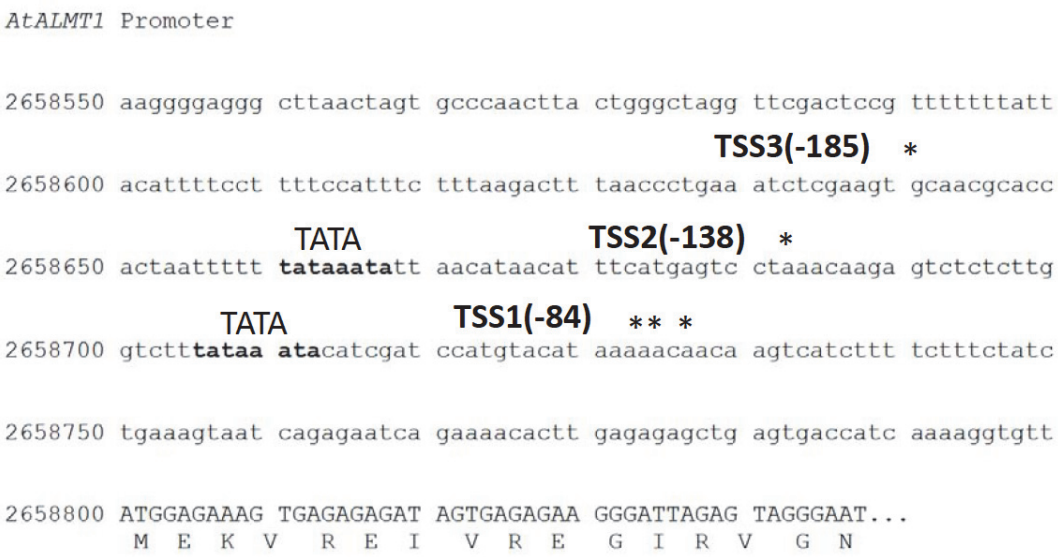


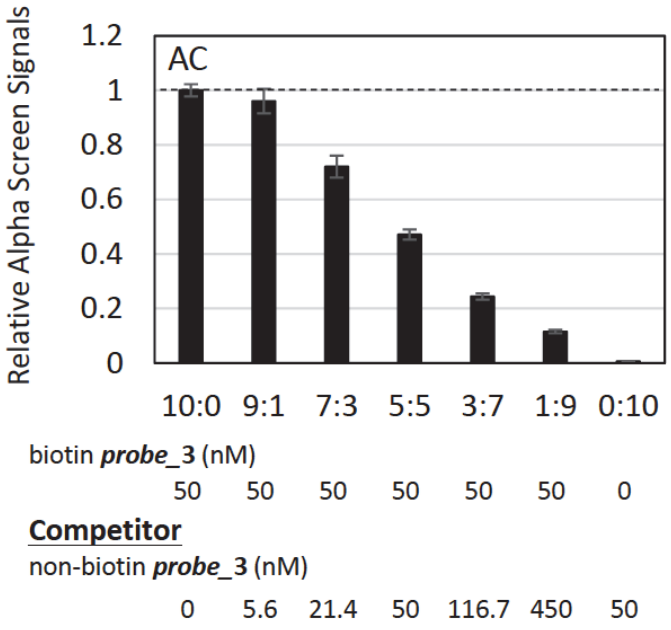
Figure 2-10. Schematic representation of Al-inducible expression of *AtALMT1*.
Black rectangles indicate *cis*-acting elements predicted by promoter scanning in Figure 2 and confirmed by mutated promoter-reporter assays (Figure 3). Putative functions of transcription factors (e.g. suppressor or activator) are indicated for the experimentally validated transcription factors (STOP1 and CAMTA2, this study; WRKY46, Ding et al., 2013).



	Position from ATG	5'End Sequence	Number of transcripts	
	-84	AACAAGTCAT...	17	
TSS1	-83	ACAAGTCATC...	5	/34
	-81	AAGTCATCTT...	12	
TSS2	-138	AAACAAGAGT...	21	/21
TSS3	-185	GCACCACTAA...	6	/6

Figure S2-1. The 5' end of AtALMT1 transcripts determined by 5' RACE.
5' RACEanalyses were performed RNA samples isolated from Al treated roots. Asterisks indicate 5'ends of transcripts and were categorized as TSS1, 2 and 3. Number of transcripts were shown below the sequence.

A



B

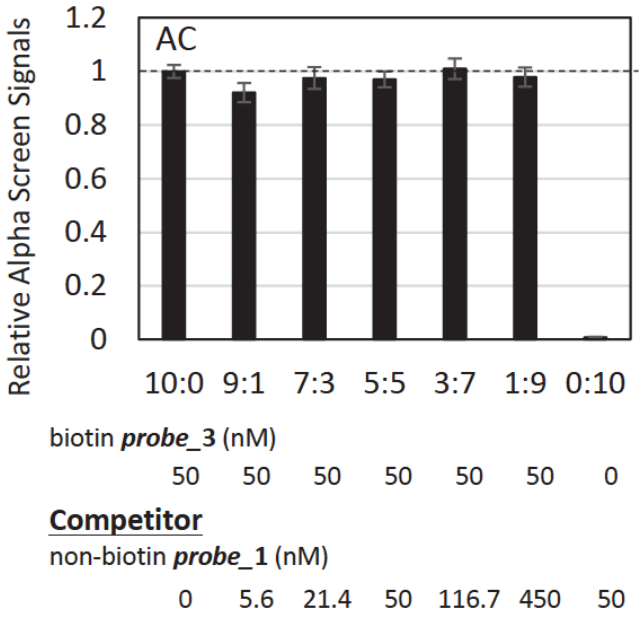


Figure S2-2. In vitro binding assay of STOP1 protein to the dsDNA probe 3 containing putative STOP1 binding sites of AtALMT1 promoter.
Competitive binding assay of biotinylated probe 3 in the presence of non-labeled probe 3 (A) or probe 1 (negative control) (B). Values are the mean ± SD (n = 3).

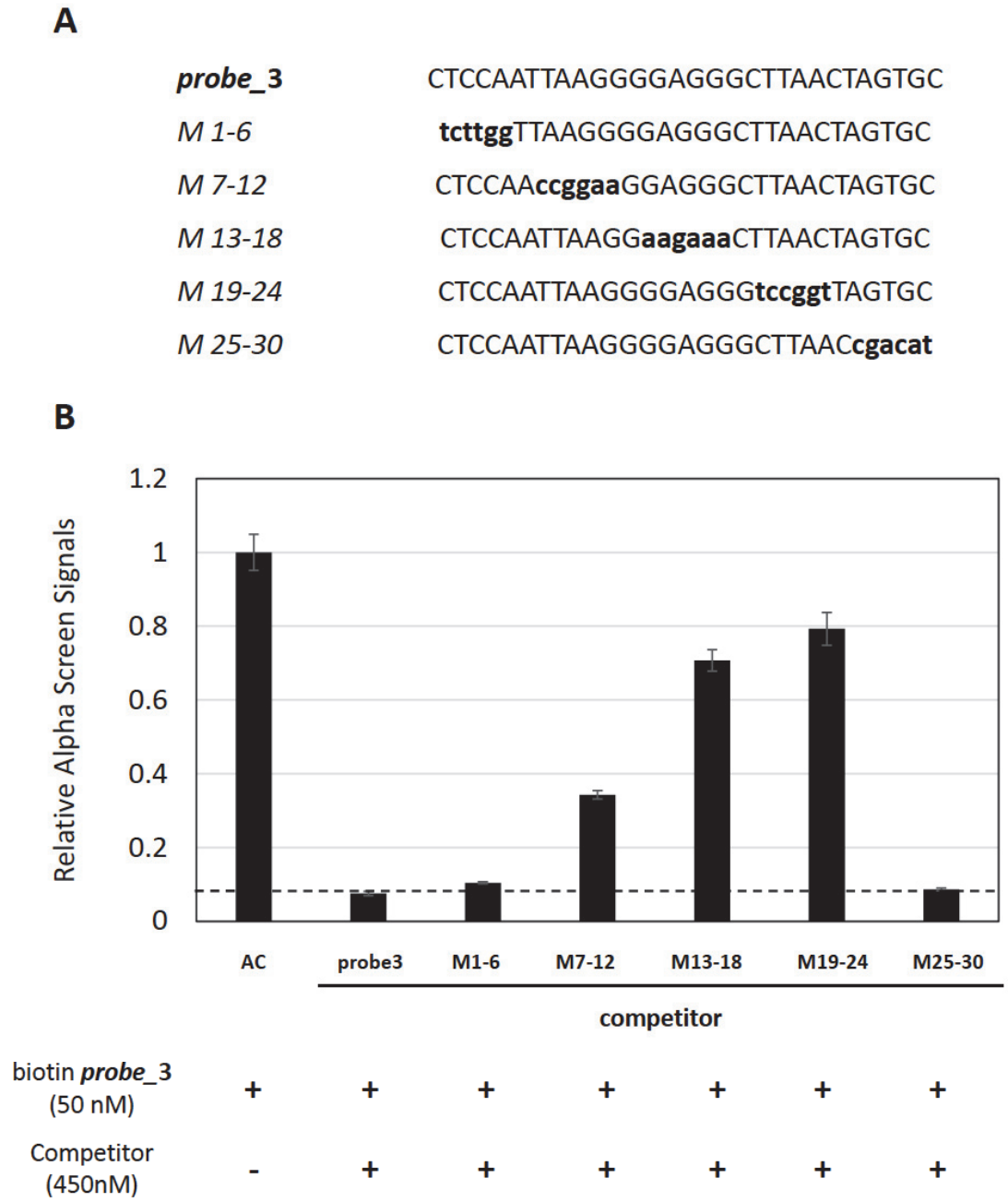
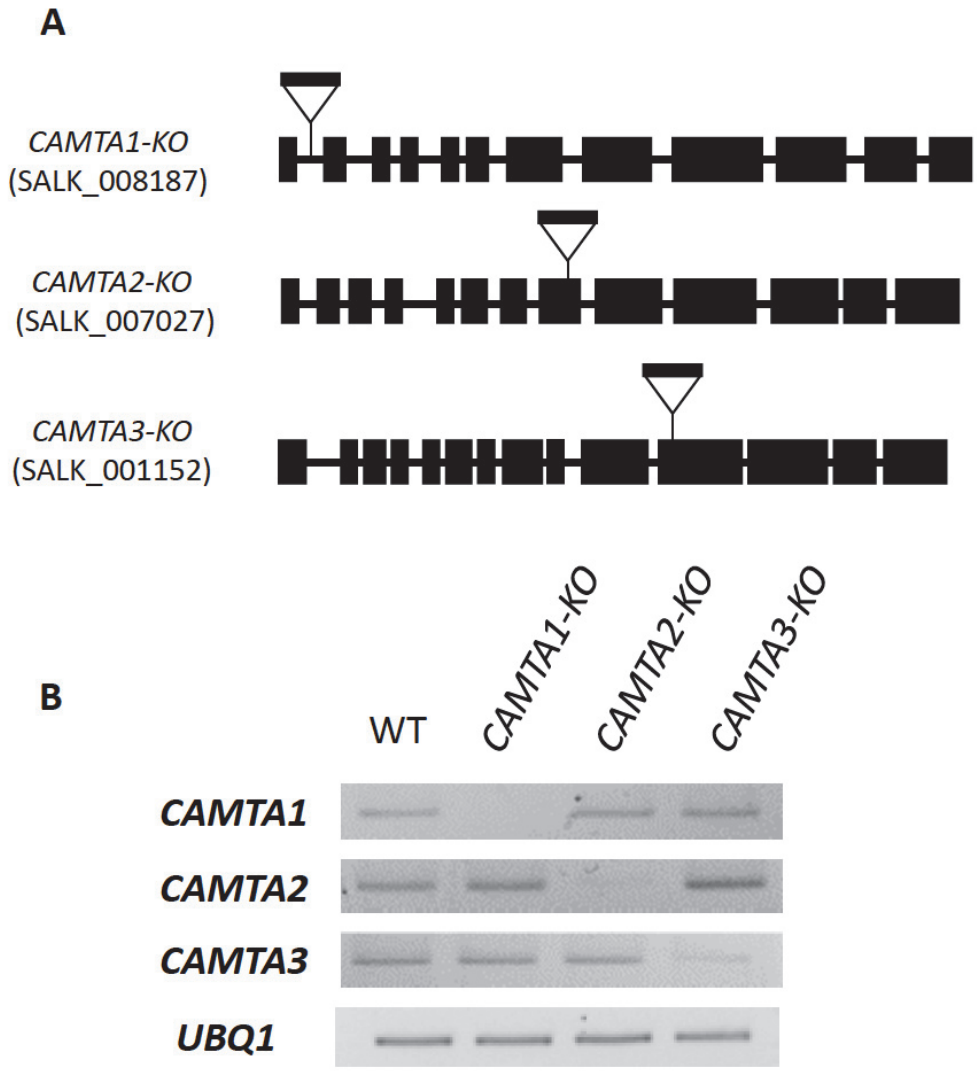


Figure S2-3. In vitro binding assay of STOP1 protein to the mutated dsDNA probe3 (see Fig 5). Relative alpha screen signals of STOP1 protein to the biotinylated probe3 with absence (AC) or presence of non-biotinylated mutated dsDNA in panel A are shown in the panel B. Values ± SD are shown (n = 3).



Supplemental Figure S4. Position of T-DNA insertion in the knockout lines of CAMTA1, 2 and 3, and whose expression levels in Al stressed conditions.
Panel A shows position of T-DNA insertion, while the panel B shows the gel image of RT-PCR.

Table S2-1.List of overrepresented octamer units in the AtALMT1 promoter based on the relative appearance rate calculated from microarray datasets.

Position	Octamer	RAR			P-value			Motif*	Predicted *Cis* region
		6h	24h	stop1	6h	24h	stop1		
-526	GAGGGCAC	2.88	2.41	7.71	0.299	0.346	0.008		
-525	AGGGCACT	1.68	2.8	2.99	0.453	0.166	0.15		A
-524	GGGCACTA	2.34	3.91	2.09	0.353	0.099	0.386		
-361	TCCTACCG	2.08	1.74	3.71	0.387	0.443	0.107	Hbox	B
-360	CCTACCGG	3.88	3.24	6.91	0.234	0.273	0.038		
-320	CCTCACGC	4.54	2.84	1.01	0.014	0.095	0.63		
-319	CTCACGCG	3.31	2.07	0.74	0.037	0.184	0.608	CGCG	
-318	TCACGCGT	2.04	1.71	0.91	0.261	0.332	0.7	CGCG	
-317	CACGCGTC	0.91	0.76	0.81	0.701	0.624	0.653	CGCG	C
-316	ACGCGTCG	1.56	2.61	1.39	0.477	0.185	0.516	CGCG	
-315	CGCGTCGC	2.88	2.41	5.14	0.299	0.346	0.063	CGCG	
-314	GCGTCGCT	2.5	2.08	4.45	0.336	0.387	0.079		
-313	CGTCGCTC	3.46	2.89	1.54	0.119	0.159	0.482		
-301	TAAGGGGA	2.68	3.72	1.59	0.108	0.013	0.363		
-300	AAGGGGAG	1.01	1.69	0.9	0.63	0.336	0.697		
-299	AGGGGAGG	1.14	0.95	2.02	0.588	0.716	0.265		D
-298	GGGGAGGG	3.69	1.54	4.93	0.108	0.483	0.026		
-297	GGGAGGGC	3.63	3.03	3.23	0.248	0.288	0.273		
-285	CTAGTGCC	4.62	1.29	2.75	0.031	0.545	0.171		
-284	TAGTGCCC	1.94	1.62	8.64	0.408	0.466	0		E
-283	AGTGCCCA	2.47	2.06	4.41	0.2	0.259	0.015	GCCCA	
-282	GTGCCCAA	4.93	3.29	1.76	0.004	0.038	0.32	GCCCA	
-270	CTGGGCTA	4.27	2.38	1.27	0.037	0.212	0.549	GCCCA	
-269	TGGGCTAG	0.74	1.85	0.66	0.609	0.227	0.554	GCCCA	
-268	GGGCTAGG	1.65	1.38	1.47	0.458	0.52	0.497		
-267	GGCTAGGT	3.46	2.89	4.63	0.119	0.159	0.031		F
-266	GCTAGGTT	0.73	1.22	1.95	0.603	0.492	0.205		
-265	CTAGGTTC	3.41	0.71	2.28	0.034	0.591	0.152		
-264	TAGGTTCG	1.68	1.4	3.74	0.339	0.422	0.013		
-263	AGGTTCGA	1.01	0.84	3.15	0.592	0.577	0.009		
-257	GACTCCGT	1.44	3.61	2.57	0.504	0.055	0.189		G
-210	TAACCCTG	2.37	1.98	3.16	0.212	0.274	0.075		
-192	GTGCAACG	3.31	3.68	0.98	0.067	0.027	0.729		
-191	TGCAACGC	2.74	2.29	1.22	0.171	0.223	0.562		H
-190	GCAACGCA	4.97	2.49	3.55	0.004	0.126	0.03		

Relative appearance rate in the Al-inducible (fold change >3) and repressed in the stop1 mutant in Al treatment (fold change <2.5) were listed with the RAR values and P value of student's t-test. Red colar indicates that the RAR values are above threshhold, and the yellow filled coloms indecate $p<0.05$. Predicted cis-regions shown in Fig 5 were also shown.

Table S2-2. Sequence of mutated probes used for in vitro binding assay of STOP1 protein to the *AtALMT1* promoter region.

Mutation	Sequence
probe3	CTCCAA**TTAAGGGGAGGGCTTAACTA** GTGC
M7 (T/C)	CTCCAA**c**TAAGGGGAGGGCTTAACTAGTGC
M8 (T/C)	CTCCAAT**c**AAGGGGAGGGCTTAACTAGTGC
M9 (A/G)	CTCCAATT**g**AGGGGAGGGCTTAACTAGTGC
M10 (A/G)	CTCCAATTA**g**GGGGAGGGCTTAACTAGTGC
M11 (G/A)	CTCCAATTAA**a**GGGAGGGCTTAACTAGTGC
M12 (G/A)	CTCCAATTAAG**a**GGAGGGCTTAACTAGTGC
M13 (G/A)	CTCCAATTAAGG**a**GAGGGCTTAACTAGTGC
M14 (G/A)	CTCCAATTAAGGG**a**AGGGCTTAACTAGTGC
M15 (A/G)	CTCCAATTAAGGGG**g**GGGCTTAACTAGTGC
M16 (G/A)	CTCCAATTAAGGGGA**a**GGCTTAACTAGTGC
M17 (G/A)	CTCCAATTAAGGGGAG**a**GCTTAACTAGTGC
M18 (G/A)	CTCCAATTAAGGGGAGG**a**CTTAACTAGTGC
M19 (C/T)	CTCCAATTAAGGGGAGGG**t**TTAACTAGTGC
M20 (T/C)	CTCCAATTAAGGGGAGGGC**c**TAACTAGTGC
M21 (T/C)	CTCCAATTAAGGGGAGGGCT**c**AACTAGTGC
M22 (A/G)	CTCCAATTAAGGGGAGGGCTT**g**ACTAGTGC
M23 (A/G)	CTCCAATTAAGGGGAGGGCTTA**g**CTAGTGC
M24 (C/T)	CTCCAATTAAGGGGAGGGCTTAA**t**TAGTGC
M25 (T/C)	CTCCAATTAAGGGGAGGGCTTAAC**c**AGTGC
M26 (A/G)	CTCCAATTAAGGGGAGGGCTTAACT**g**GTGC

Red color indicates mutated nucleotide.

Table S2-3. Fold change (10 μM Al/control; pH 5, 24 hours) of CAMTA families in Al-treated roots.

Name	AGI code	FC(Al/Control)	Description
CAMTA1	AT5G09410	1.27	ethylene induced calmodulin binding protein
CAMTA2	AT5G64220	1.93	Calmodulin-binding transcription activator protein with CG-1 and Ankyrin domains
CAMTA3	AT2G22300	1.26	signal responsive 1
CAMTA4	AT1G67310	0.95	Calmodulin-binding transcription activator protein with CG-1 and Ankyrin domains
CAMTA5	AT4G16150	0.94	calmodulin binding;transcription regulators
CAMTA6	AT3G16940	0.90	calmodulin binding;transcription regulators

Table S2-4. Sequence information of PCR primers

PCR category		Sequence (5' to 3')	
		Forward primer sequence	Reverse primer sequence
ALMT1 mutation primer	cisA Mutation	GCCCAAGTGTTAAAAAAAATAATCTTGTCG	CGACAAGATTATTTTTTTTAACACTTGGGC
	cisB Mutation	CATGATCAACTAAAAAAAATTCATCTAAC	GTTAGATGAATTTTTTTTAGTTGATCATG
	cisC Mutation	GCATAAAAACAAAAAAAAGTCGCTCCAA	TTGGAGCGACTTTTTTTTGTTTTTATGC
	cisD Mutation	CTCCAATTAAGTTTCTTTATTAACTAGTGCC	GGCACTAGTTAATAAAGAAACTTAATTGGAG
	cisE Mutation	GGGCTTAACTAAAAAAAAACTTACTGGGC	GCCCAGTAAGTTTTTTTTTAGTTAAGCCC
	cisF Mutation	TGCCCAACTTAAAAAAAAAGGTTCGACTCC	GGAGTCGAACCTTTTTTTTTAAGTTGGGCA
	cisG Mutation	GGGCTAGGTTCAAAAAAAATTTTTTATTAC	GTAATAAAAAATTTTTTTTGAACCTAGCCC
	cisH Mutation	CTGAAATCTCGAAGAAAAAAAAACCACTAA	TTAGTGGTTTTTTTTTCTTCGAGATTTCAG
Real-time RT PCR	*GUS*	GTGTGAGCGTCGCAGAACATT	CGGAAGCAACGCGTAAACTC
	AtALMT1 (At1g08430)	TCTTCATGTTTTTCATGGTTTGAGTT	CACAGTTTTACATGACGTTGATAATGAT
	AtCAMTA1	ACGTTGCTACTGGATGCTTGAACA	TGGTCCGGTTACCCTTAACCTC CA
	AtCAMTA2	CACCCAGTGGTTCACTCTTTCTC	TGCCCGTCTTTTCGGAAATA
	AtCAMTA3	CTGGGCCTTAGAACCAACAATAA	ACCATTTACATCGCGAAAATCA
	UBQ1 (At3g52590)	TCGTAAGTACAATCAGGATAAGATG	CACTGAAACAAGAAAAACAAACCCT
Real-time RT PCR for Quantification of *AtALMT1* transcripts	*AtALMT1* _TSS1	AATATTAACATAACATTTCATGAGTCCTAA	CGGGTCTTCATTCCCTACTCTA
	*AtALMT1*_ TSS2	CATCGATCCATGTACATAAAAACA	CGGGTCTTCATTCCCTACTCTA
	*AtALMT1*_ TSS3	CATCTTTTCTTTCTATCTGAAAGTAATCAG	CGGGTCTTCATTCCCTACTCTA

REFERENCES (Chapter 1**)**

Alexandrov NN, Troukhan ME, Brover VV, Tatarinova T, Flavell RB, Feldmann KA (2006) Features of Arabidopsis genes and genome discovered using full-length cDNAs. Plant Mol Biol **60:** 69-85

Alkhateeb RS, Vorholter FJ, Ruckert C, Mentz A, Wibberg D, Hublik G, Niehaus K, Puhler A (2016) Genome wide transcription start sites analysis of Xanthomonas campestris pv. campestris B100 with insights into the gum gene cluster directing the biosynthesis of the exopolysaccharide xanthan. J Biotechnol **225:** 18-28

Basehoar AD, Zanton SJ, Pugh BF (2004) Identification and distinct regulation of yeast TATA box-containing genes. Cell **116:** 699-709

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30:** 2114-2120

Bracht J, Hunter S, Eachus R, Weeks P, Pasquinelli AE (2004) Trans-splicing and polyadenylation of let-7 microRNA primary transcripts. RNA **10:** 1586-1594

Cai X, Hagedorn CH, Cullen BR (2004) Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. RNA **10:** 1957-1966

Cao J, Schneeberger K, Ossowski S, Gunther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C, Wang X, Ott F, Muller J, Alonso-Blanco C, Borgwardt K, Schmid KJ, Weigel D (2011) Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nature Genetics **43:** 956-963

Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, Engstrom PG, Frith MC, Forrest AR, Alkema WB, Tan SL, Plessy C, Kodzius R, Ravasi T, Kasukawa T, Fukuda S, Kanamori-Katayama M, Kitazume Y, Kawaji H, Kai C, Nakamura M, Konno H, Nakano K, Mottagui-Tabar S, Arner P, Chesi A, Gustincich S, Persichetti F, Suzuki H, Grimmond SM, Wells CA, Orlando V, Wahlestedt C, Liu ET, Harbers M, Kawai J, Bajic VB, Hume DA, Hayashizaki Y (2006) Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet **38:** 626-635

Collins DW, Jukes TH (1994) Rates of transition and transversion in coding sequences since the human-rodent divergence. Genomics **20:** 386-396

Dvir S, Velten L, Sharon E, Zeevi D, Carey LB, Weinberger A, Segal E (2013) Deciphering the rules by which 5'-UTR sequences affect protein expression in yeast. Proc Natl Acad Sci U S A **110:** E2792-2801

FitzGerald PC, Shlyakhtenko A, Mir AA, Vinson C (2004) Clustering of DNA sequences in human promoters. Genome Res **14:** 1562-1574

Gingold H, Pilpel Y (2011) Determinants of translation efficiency and accuracy. Mol Syst Biol **7:** 481

Hüttenhofer A, Schattner P, Polacek N (2005) Non-coding RNAs: hope or hype? Trends Genet **21:** 289-297

Kawaguchi R, Bailey-Serres J (2005) mRNA sequence features that contribute to translational regulation in Arabidopsis. Nucleic Acids Res **33:** 955-965

Kimura M, Yoshizumi T, Manabe T, Yamamoto YY, Matsui M (2001) *Arabidopsis* transcriptional regulation by light stress *via* hydrogen peroxide-dependent and -independent pathways. Genes to Cells **6:** 607-617

Kodzius R, Kojima M, Nishiyori H, Nakamura M, Fukuda S, Tagami M, Sasaki D, Imamura K, Kai C, Harbers M, Hayashizaki Y, Carninci P (2006) CAGE: cap analysis of gene expression. Nat Methods **3:** 211-222

Kozak M (1981) Possible role of flanking nucleotides in recognition of the AUG initiator codon by eukaryotic ribosomes. Nucleic Acids Res **9:** 5233-5252

Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, S. KA, Lee CH, Nelson WD, Ploetz L, Singh S, Wensel A, Huala E (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Research **40:** D1202-1210.

Lee Y, Kim M, Han J, Yeom KH, Lee S, Baek SH, Kim VN (2004) MicroRNA genes are transcribed by RNA polymerase II. EMBO J **23:** 4051-4060

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25:** 1754-1760

Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res **18:** 1851-1858

Lin S, Zhang L, Luo W, Zhang X (2016) Characteristics of Antisense Transcript Promoters and the Regulation of Their Activity. Int J Mol Sci **17**

Lin Z, Li WH (2012) Evolution of 5' untranslated region length and gene expression reprogramming in yeasts. Mol Biol Evol **29:** 81-89

Martin W, Rujan T, Richly E, Hansen A, Cornelsen S, Lins T, Leister D, Stoebe B, Hasegawa M, Penny D (2002) Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc Natl Acad Sci U S A **99:** 12246-12251

Mejia-Guerra MK, Li W, Galeano NF, Vidal M, Gray J, Doseff AI, Grotewold E (2015) Core Promoter Plasticity Between Maize Tissues and Genotypes Contrasts with Predominance of Sharp Transcription Initiation Sites. Plant Cell **27:** 3309-3320

Morton T, Petricka J, Corcoran DL, Li S, Winter CM, Carda A, Benfey PN, Ohler U, Megraw M (2014) Paired-end analysis of transcription start sites in Arabidopsis reveals plant-specific promoter signatures. Plant Cell **26:** 2746-2760

Moshonov S, Elfakess R, Golan-Mashiach M, Sinvani H, Dikstein R (2008) Links between core promoter and basic gene features influence gene expression. BMC Genomics **9:** 92

Nakamura M, Tsunoda T, Obokata J (2002) Photosynthesis nuclear genes generally lack TATA-boxes: a tobacco photosystem I gene responds to light through an initiator. Plant J **29:** 1-10

Ni T, Corcoran DL, Rach EA, Song S, Spana EP, Gao Y, Ohler U, Zhu J (2010) A paired-end sequencing strategy to map the complex landscape of transcription initiation. Nat Methods **7:** 521-527

Onodera Y, Haag JR, Ream T, Costa Nunes P, Pontes O, Pikaard CS (2005) Plant nuclear RNA polymerase IV mediates siRNA and DNA methylation-dependent heterochromatin formation. Cell **120:** 613-622

Orekhova AS, Rubtsov PM (2013) Bidirectional promoters in the transcription of mammalian genomes. Biochemistry (Mosc) **78:** 335-341

Ossowski S, Schneeberger K, Lucas-Lledo JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M (2010) The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. Science **327:** 92-94

Otsuki T, Ota T, Nishikawa T, Hayashi K, Suzuki Y, Yamamoto J, Wakamatsu A, Kimura K, Sakamoto K, Hatano N, Kawai Y, Ishii S, Saito K, Kojima S, Sugiyama T, Ono T, Okano K, Yoshikawa Y, Aotsuka S, Sasaki N, Hattori A, Okumura K, Nagai K, Sugano S, Isogai T (2005) Signal sequence and keyword trap in silico for selection of full-length human cDNAs encoding secretion or membrane proteins from oligo-capped cDNA libraries. DNA Res **12:** 117-126

Potter J, Zheng W, Lee J (2003) Thermal stability and cDNA synthesis capacity of SuperScript III reverse trnascriptase. Focus **25.1:** 19-24

Rao YS, Wang ZF, Chai XW, Nie QH, Zhang XQ (2013) Relationship between 5' UTR length and gene expression pattern in chicken. Genetica **141:** 311-318

Schug J, Schuller WP, Kappen C, Salbaum JM, Bucan M, Stoeckert CJ, Jr. (2005) Promoter features related to tissue specificity as measured by Shannon entropy. Genome Biol **6:** R33

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, 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:** 15776-15781

Smith DR (2014) Mitochondrion-to-plastid DNA transfer: it happens. New Phytol **202:** 736-738

Sunkar R, Li YF, Jagadeeswaran G (2012) Functions of microRNAs in plant stress responses. Trends Plant Sci **17:** 196-203

Suzuki Y, Tsunoda T, Sese J, Taira H, Mizushima-Sugano J, Hata H, Ota T, Isogai T, Tanaka T, Nakamura Y, Suyama A, Sakaki Y, Morishita S, Okubo K, Sugano S (2001) Identification and characterization of the potential promoter regions of 1031 kinds of human genes. Genome Res **11:** 677-684

Takahashi H, Kato S, Murata M, Carninci P (2012) CAGE (cap analysis of gene expression): a protocol for the detection of promoter and transcriptional networks. Methods Mol Biol **786:** 181-200

Taylor MS, Kai C, Kawai J, Carninci P, Hayashizaki Y, Semple CA (2006) Heterotachy in mammalian promoter evolution. PLoS Genet **2:** e30

The_Arabidopsis_Genome_Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408:** 796-815

Tokizawa M, Kobayashi Y, Saito T, Kobayashi M, Satoshi I, Nomoto M, Tada Y, Yamamoto YY, Koyama H (2015) STOP1, CAMTA2 and other transcription factors are involved in aluminum-inducible AtALMT1 expression. Plant Physiol **167:** 991-1003

Tsuchihara K, Suzuki Y, Wakaguri H, Irie T, Tanimoto K, Hashimoto S, Matsushima K, Mizushima-Sugano J, Yamashita R, Nakai K, Bentley D, Esumi H, Sugano S (2009) Massive transcriptional start site analysis of human genes in hypoxia cells. Nucleic Acids Res **37:** 2249-2263

Xie Z, Allen E, Fahlgren N, Calamar A, Givan SA, Carrington JC (2005) Expression of Arabidopsis MIRNA genes. Plant Physiol **138:** 2145-2154

Yamamoto YY, Ichida H, Abe T, Suzuki Y, Sugano S, Obokata J (2007) Differentiation of core promoter architecture between plants and mammals revealed by LDSS analysis. Nucleic Acids Res **35:** 6219-6226

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

Yamamoto YY, Yoshioka Y, Hyakumachi M, Maruyama K, Yamaguchi-Shinozaki K, Tokizawa M, Koyama H (2011) Prediction of transcriptional regulatory elements for plant hormone responses based on microarray data. BMC Plant Biol **11:** 39

Yamamoto YY, Yoshioka Y, Hyakumachi M, Obokata J (2011) Characterization of core promoter types with respect to gene structure and expression in *Arabidopsis thaliana*. DNA Research **18:** 333-342

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 Journal **60:** 350-362

Yamashita R, Sathira NP, Kanai A, Tanimoto K, Arauchi T, Tanaka Y, Hashimoto S, Sugano S, Nakai K, Suzuki Y (2011) Genome-wide characterization of transcriptional start sites in humans by integrative transcriptome analysis. Genome research **21:** 775-789

Yang H (2009) In plants, expression breadth and expression level distinctly and non-linearly correlate with gene structure. Biol Direct **4:** 45; discussion 45

REFERENCES (Chapter 2)

Arenhart RA, Bai Y, Valter de Oliveira LF, Bucker Neto L, Schunemann M, Maraschin FD, Mariath J, Silverio A, Sachetto-Martins G, Margis R, Wang ZY, and Margis-Pinheiro M (2014) New insights into aluminum tolerance in rice: The ASR5 protein binds the *STAR1* promoter and other aluminum-responsive genes. Mol Plant **7**: 709-721

Baetz U and Martinoia E (2014) Root exudates: the hidden part of plant defense. Trends Plant Sci **19**: 90-98

Birnbaum K, Shasha DE, Wang JY, Jung JY, Lambert GM, Galbraith DW and Benfey PN (2003) A gene expression map of the Arabidopsis root. Science **302**:1956-1960

Bustin SA, Benes V, Garson JA, Hellemans J, Huggett J, Kubista M, Mueller R, Nolan T, Pfaffl MW, Shipley GL, Vandesompele J, and Wittwer CT (2009) The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. Clin Chem **55**: 611-622

Carrari F, Fernie AR, Iusem ND (2004) Heard it through the grapevine? ABA and sugar cross-talk: the ASR story. Trends Plant Sci **9**: 2-4

Chen ZC, Yokosho K, Kashino M, Zhao F, Yamaji N, and Ma JF (2013) Adaptation to acidic soil is achieved by increased numbers of *cis*-acting elements regulating *ALMT1* expression in *Holcus lanatus*. Plant J **76**: 10-23

Clough SJ and Bent AF (1998) Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. Plant J **16**: 735-743

Ding ZJ, Yan JY, Xu XY, Li GX, and Zheng SJ (2013) WRKY46 functions as a transcriptional repressor of *ALMT1*, regulating aluminum-induced malate secretion in Arabidopsis. Plant J **76**: 825-835

Finkler A, Ashery-Padan R, and Fromm H (2007) CAMTAs: calmodulin-binding transcription activators from plants to human. FEBS Lett **581:** 3893-3898

Fujiwara T, Hirai MY, Chino M, Komeda Y, Naito S (1992) Effects of Sulfur Nutrition on Expression of the Soybean Seed Storage Protein Genes in Transgenic Petunia . Plant Physiol **99**: 263-268

Galon Y, Nave R, Boyce JM, Nachmias D, Knight MR, Fromm H (2008) Calmodulin-binding transcription activator (CAMTA) 3 mediates biotic defense responses in *Arabidopsis*. FEBS Lett **582**: 943-8

Hoekenga OA, Maron LG, Pineros MA, Cancado GM, Shaff J, Kobayashi Y, Ryan PR, Dong B, Delhaize E, Sasaki T, Matsumoto H, Yamamoto Y, Koyama H, and Kochian LV (2006) *AtALMT1*, which encodes a malate transporter, is identified as one of several genes critical for aluminum tolerance in *Arabidopsis*. Proc Natl Acad Sci USA **103**: 9738-9743

Hieno A, Naznin HA, Hyakumachi M, Sakurai T, Tokizawa M, Koyama H, Sato N, Nishiyama T, Hasebe M, Zimmer AD, Dang D, Reski R, Rensing S, Obokata J, Yamamoto YY (2013) ppdb: Plant Promoter Database Version 3.0. Nucleic Acids Res **42**: 1188-1192

Horton RM, Hunt HD, Ho SN, Pullen JK, and Pease LR (1989) Engineering hybrid genes without the use of restriction enzymes: gene splicing by overlap extension. Gene **77:** 61-68

Huang CF, Yamaji N, Mitani N, Yano M, Nagamura Y and Ma JF (2009) A bacterial-type ABC transporter is involved in aluminum tolerance in rice. Plant Cell **21:** 655-667

Iuchi S, Koyama H, Iuchi A, Kobayashi Y, Kitabayashi S, Kobayashi Y, Ikka T, Hirayama T, Shinozaki K, and Kobayashi M (2007) Zinc finger protein STOP1 is critical for proton tolerance in *Arabidopsis* and coregulates a key gene in aluminum tolerance. Proc Natl Acad Sci USA **104**: 9900-9905

Kim Y, Park S, Gilmour SJ, Thomashow MF (2013) Roles of CAMTA transcription factors and salicylic acid in configuring the low-temperature transcriptome and freezing tolerance of Arabidopsis. Plant J **75**: 364-76

Kobayashi Y, Hoekenga OA, Itoh H, Nakashima M, Saito S, Shaff JE, Maron LG, Pineros MA, Kochian LV, and Koyama H (2007) Characterization of *AtALMT1* expression in aluminum-inducible malate release and its role for rhizotoxic stress tolerance in Arabidopsis. Plant Physiol **145**: 843-852

Kobayashi T and Nishizawa NK (2012) Iron uptake, translocation, and regulation in higher plants. Annu Rev Plant Biol **63**: 131-152

Kobayashi Y, Kobayashi Y, Sugimoto M, Lakshmanan V, Iuchi S, Kobayashi M, Bais HP, and Koyama H (2013a) Characterization of the complex regulation of *AtALMT1* expression in response to phytohormones and other inducers. Plant Physiol **162**: 732-740

Kobayashi Y, Kobayashi Y, Watanabe T, Shaff JE, Ohta H, Kochian L, Wagatsuma T, Kinraide TB, and Koyama H (2013b) Molecular and physiological analysis of Al^{3+} and H^{+} rhizotoxicities at moderately acidic conditions. Plant Physiol **163**: 180-192

Kochian LV, Hoekenga OA, and Piñeros MA (2004) How do crop plants tolerate acid soils? Mechanisms of aluminum tolerance and phosphorous efficiency. Annu Rev Plant Biol **55:** 459-493

Lakshmanan V, Kitto SL, Caplan JL, Hsueh YH, Kearns DB, Wu YS, and Bais HP (2012) Microbe-associated molecular patterns-triggered root responses mediate beneficial rhizobacterial recruitment in Arabidopsis. Plant Physiol **160**: 1642-1661

Liang C, Pineros MA, Tian J, Yao Z, Sun L, Liu J, Shaff J, Coluccio A, Kochian LV, and Liao H (2013) Low pH, aluminum, and phosphorus coordinately regulate malate exudation through GmALMT1 to improve soybean adaptation to acid soils. Plant Physiol **161**: 1347-1361

Liu J, Magalhaes JV, Shaff J, Kochian LV (2009) Aluminum-activated citrate and malate transporters from the MATE and ALMT families function independently to confer Arabidopsis aluminum tolerance. Plant J **57:** 389-399

Liu J, Piñeros MA, and Kochian LV (2014) The role of aluminum sensing and signaling in plant aluminum resistance. J Integr Plant Biol **56:** 221-230

Narusaka Y, Nakashima K, Shinwari ZK, Sakuma Y, Furihata T, Abe H, Narusaka M, Shinozaki K, and Yamaguchi-Shinozaki K (2003) Interaction between two cis-acting elements, ABRE and DRE, in ABA-dependent expression of *Arabidopsis rd29A* gene in response to dehydration and high-salinity stresses. Plant J **34:** 137-148

Neumann G, Massonneau A, Martinoia E, and Römheld V (1999) Physiological adaptations to phosphorus deficiency during proteoid root development in white lupin. Planta **208**: 373-382

Okumura T, Makiguchi H, Makita Y, Yamashita R, Nakai K (2007) Melina II: a web tool for comparisons among several predictive algorithms to find potential motifs from promoter regions. Nucleic Acids Res **35**: 227-231

Pavletich NP and Pabo CO (1991) Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. Science **252**: 809-817

Pandey N, Ranjan A, Pant P, Tripathi RK, Ateek F, Pandey HP, Patre U V, Sawant S V (2013) CAMTA 1 regulates drought responses in *Arabidopsis thaliana*. BMC Genomics **14**: 216

Rudrappa T, Czymmek KJ, Pare PW, and Bais HP (2008) Root-secreted malic acid recruits beneficial soil bacteria. Plant Physiol. **148**: 1547-1556

Sasaki T, Yamamoto Y, Ezaki B, Katsuhara M, Ahn SJ, Ryan PR, Delhaize E, and Matsumoto H (2004) A wheat gene encoding an aluminum-activated malate transporter. Plant J **37**: 645-653

Sasaki T, Ryan PR, Delhaize E, Hebb DM, Ogihara Y, Kawaura K, Noda K, Kojima T, Toyoda A, Matsumoto H (2006) Sequence upstream of the wheat (*Triticum aestivum L.*) *ALMT1* gene and its relationship to aluminum resistance. Plant Cell Physiol **47**: 1343-54

Sawaki Y, Iuchi S, Kobayashi Y, Kobayashi Y, Ikka T, Sakurai N, Fujita M, Shinozaki K, Shibata D, Kobayashi M, and Koyama H (2009) STOP1 regulates multiple genes that protect Arabidopsis from proton and aluminum toxicities. Plant Physiol **150**: 281-294

Segal DJ, Dreier B, Beerli RR, and Barbas CF (1999) Toward controlling gene expression at will: selection and design of zinc finger domains recognizing each of the 5'-GNN-3' DNA target sequences. Proc Natl Acad Sci USA **96**: 2758-2763

Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov A V, Frith MC, Fu Y, Kent WJ, et al (2005) Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol **23**: 137-44

Trémousaygue D, Garnier L, Bardet C, Dabos P, Hervé C, and Lescure B (2003) Internal telomeric repeats and 'TCP domain'protein-binding sites co-operate to regulate gene expression in *Arabidopsis thaliana* cycling cells. Plant J **33**: 957-966

Tsutsui T, Yamaji N, and Ma JF (2011) Identification of a cis-acting element of ART1, a C2H2-type zinc-finger transcription factor for aluminum tolerance in rice. Plant Physiol **156**: 925-931

Yamaji N, Huang CF, Nagao S, Yano M, Sato Y, Nagamura Y, and Ma JF (2009) A Zinc finger transcription factor ART1 regulates multiple genes implicated in Aluminum tolerance in Rice. Plant Cell **21**: 3339-3349

Yamamoto YY, 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

Yamamoto YY, Yoshioka Y, Hyakumachi M, Obokata J (2011a) Characteristics of core promoter types with respect to gene structure and expression in *Arabidopsis thaliana*. DNA Res **18**: 333-342

Yamamoto YY, Yoshioka Y, Hyakumachi M, Maruyama K, Yamaguchi-Shinozaki K, Tokizawa M, and Koyama H (2011b) Prediction of transcriptional regulatory elements for plant hormone responses based on microarray data. BMC Plant Biology **11:** 39

Yang T, Poovaiah BW (2002) A calmodulin-binding/CGCG box DNA-binding protein family involved in multiple signaling pathways in plants. J Biol Chem **277**: 45049-45058

Zou C, Sun K, Mackaluso JD, Seddon AE, Jin R, Thomashow MF, and Shiu SH

(2011) Cis-regulatory code of stress-responsive transcription in *Arabidopsis thaliana*. Proc Natl Acad Sci USA **108**: 14992-14997