the plant journal

SEB SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2009) **60**, 350–362 doi: 10.1111/j.1365-313X.2009.03958.x

Heterogeneity of Arabidopsis core promoters revealed by high-density TSS analysis

Yoshiharu Y. Yamamoto[1,†], Tomoaki Yoshitsugu[1], Tetsuya Sakurai[2], Motoaki Seki[3], Kazuo Shinozaki[2,3,4] and Junichi Obokata[1,5,*]

[1]*Center for Gene Research, Nagoya University, Nagoya 464-8602, Japan,*

[2]*Metabolomics Research Group, RIKEN Plant Science Center, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan,*

[3]*Plant Functional Genomics Research Group, RIKEN Plant Science Center, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan,*

[4]*Gene Discovery Research Group, RIKEN Plant Science Center, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan, and*

[5]*Laboratory of Plant Genome Biology, Graduate School of Life and Environmental Sciences, Kyoto Prefectural University, Kyoto 606-8552, Japan*

Received 6 January 2009; revised 26 May 2009; accepted 17 June 2009; published online 24 July 2009.

[*]For correspondence (fax +81 52 789 3083; e-mail obokata@gene.nagoya-u.ac.jp).

[†]Present address: Faculty of Applied Biological Sciences, Gifu University, Yanagido 1-1, Gifu City, 501-1193, Japan.

SUMMARY

Our limited understanding of plant promoters does not allow us to recognize any core promoter elements for the majority of plant promoters. To understand the promoter architecture of Arabidopsis, we used the combined approach of *in silico* detection of novel core promoter elements and large-scale determination of transcription start sites (TSSs). To this end, we developed a novel methodology for TSS identification, using a combination of the cap-trapper and massively parallel signature sequencing methods. This technique, CT–MPSS, allowed us to identify 158 237 Arabidopsis TSS tags corresponding to 38 311 TSS loci, which provides an opportunity for quantitative analysis of plant promoters. The expression characteristics of these promoters were analyzed with respect to core promoter elements detected by our *in silico* analyses, revealing that Arabidopsis promoters contain two main types of elements with exclusive characteristics, the TATA type and the GA type. The TATA-type promoters tend to be associated with the Y Patch and the Inr motif, and cause high expression with sharp-peak TSS clusters. By contrast, the GA type produces broad-type TSS clusters. Unlike mammalian promoters, plant promoters are not associated with CpG islands. However, plant-specific GA-type promoters share some characteristics with mammalian CpG-type promoters.

Keywords: core promoter, local distribution of short sequences analysis, massively parallel signature sequencing, TATA box, transcription start site, transcriptome.

INTRODUCTION

Recent comprehensive studies of mammalian promoters have revealed that promoters driving structural genes have a cluster of transcription start sites (TSSs), rather than a single TSS. The distribution profiles of TSSs within a cluster are related to the type of core promoter element: sharp-peak TSS clusters, which contain a high single-peak TSS accompanied by surrounding weak TSSs, are found with TATA boxes, and broad-peak TSS clusters are associated with CpG islands, two major representatives of mammalian promoters (Suzuki *et al.*, 2001; Carninci *et al.*, 2006). The TATA box tends to be absent from CpG-type promoters (Smale and Kadonaga, 2003), and is known to be the binding site of the general transcription factor TFIID, which plays a pivotal role in assembly of the transcription pre-initiation complex (Butler and Kadonaga, 2002; Smale and Kadonaga, 2003). The role of CpG islands in transcription is poorly understood (Smale and Kadonaga, 2003). CpG islands contain recognition sites for Sp1, a zinc finger-type transcription factor (Butler and Kadonaga, 2002), but many CpG-related sequences other than Sp1 binding motifs are found in the core promoter region, each with a different distribution profile (Yamamoto *et al.*, 2007a).

© 2009 The Authors
Journal compilation © 2009 Blackwell Publishing Ltd

The TATA box and Inr, a motif at the TSS (Corden *et al.*, 1980; Smale and Baltimore, 1989), are also core promoter elements in plants. Plant consensus sequences for the TATA box and the region around TSSs were first explored through pioneering work by Joshi based on 79 published gene sequences from 15 plant species (Joshi, 1987); it was only recently that these consensus sequences were updated (Shahmuradov *et al.*, 2003; Molina and Grotewold, 2005; Berendzen *et al.*, 2006; Yamamoto *et al.*, 2007b). We previously reported that the consensus sequences for the TATA box differ slightly between Arabidopsis and rice (Yamamoto *et al.*, 2007a,b). The functionality of plant TATA boxes and Inr has been confirmed by several studies (Mukumoto *et al.*, 1993; Zhu *et al.*, 1995; Nakamura *et al.*, 2002; Schmidt *et al.*, 2004; Kiran *et al.*, 2006). Plants also have TATA-less promoters (Yamamoto *et al.*, 1993; Elrouby and Bureau, 2000; Carrari *et al.*, 2001; Nakamura *et al.*, 2002). CpG islands are frequently found within mammalian TATA-less promoters, but comparative analysis of plant and mammalian promoters revealed that CpG islands are not core promoter constituents in plants (Yamamoto *et al.*, 2007a). Plant promoters have a novel sequence element called the Y Patch, but its functional characteristics are not yet known (Yamamoto *et al.*, 2007b).

Given the results of studies of mammalian core promoters, there are several unanswered questions about plant promoters. Are there also broad-peak promoters in plants? If so, which element in plants is a substitute for the CpG island in mammals? What are the roles of the newly identified plant core promoter elements?

In this report, we describe a novel method called CT–MPSS and its use for high-density mapping of Arabidopsis TSSs. This method is a combination of cap trapper, a reliable method for preparing the 5′ end of full-length cDNA (Carninci *et al.*, 1996), and massively parallel signature sequencing (MPSS), which enables ultra-parallel sequencing in a considerably shorter time and at lower cost than conventional sequencing methods (Brenner *et al.*, 2000a). As a result, the cost of the CT–MPSS method is less than one-tenth of that of CAGE (cap-analysis gene expression) analysis, which has been used for the same purpose (Shiraki *et al.*, 2003). The CT–MPSS data obtained provided the opportunity for a quantitative analysis of plant promoters, and we examined the relationship between expression characteristics and plant core promoter elements. The results revealed that Arabidopsis promoters are divided into two distinct groups, the TATA group and the GA group, and that the latter partially plays the role of mammalian CpG-type promoters.

RESULTS

Strategy for the analysis of promoter elements

We previously established a statistical methodology for *de novo* detection of promoter elements, called local distribution of short sequences (LDSS) analysis (Yamamoto *et al.*, 2007a,b); this analysis detects short sequences such as octamers that demonstrate specific localization in promoter regions. Localized distribution is a direct result of natural selection during gene evolution. We showed that the degree of localized distribution has a strong correlation with the activity of promoter elements by analyses of TATA box variants (Yamamoto *et al.*, 2007a). Our previous LDSS analysis of Arabidopsis and rice promoter regions (−1000 to −1 relative to TSS) identified about 1000 LDSS-positive octamers from each genome, and their positional profiles relative to TSS reflected their functions (Yamamoto *et al.*, 2007a,b); octamers belonging to *cis*-regulatory elements [regulatory element group, (REG)] exhibited broad distributions around −100 bp relative to TSS, and those for the TATA box showed sharply peaked distributions around −35 bp (see Figure 1).

The current study consists of three parts. First, we extended the LDSS analysis of Arabidopsis promoters in order to cover elements located not only upstream but also downstream of TSS, based on publicly available TSS information. Second, we collected a massive amount of TSS information ourselves by *wet* experiments, using novel methodology developed in this study. The TSS information was obtained from a single cDNA library without any bias such as subtraction, and therefore we were able to examine the quantitative distribution of TSSs throughout the Arabidopsis genome, as well as performing precise and fine mapping of promoters. Thirdly, utilizing the *dry* and *wet* data obtained above, we analyzed the possible roles of LDSS-positive promoter elements of Arabidopsis with respect to expression strength and the sharpness/broadness of the TSS distribution within a promoter.

LDSS detection of Arabidopsis core promoter elements

Using the publicly available TSS information (http://rarge.gsc.riken.jp/), we analyzed the Arabidopsis promoters by the LDSS method, with the downstream end of the analyzed region extended to +200 so that core promoter elements around or downstream of TSS could be detected. The LDSS-positive octamers obtained were clustered according to their positional profiles. This clustering analysis revealed the new octamer groups Inr, CA, GA and Kozak, in addition to the previously identified groups REG, TATA and Y Patch (Figure 1).

Y Patch, CA and GA elements (green in Figure 1b) have also been found in rice, but not in mammals (Yamamoto *et al.*, 2007a), and thus are core promoter elements specific to plants. A list of the grouped sequences is given in Table S1.

The Arabidopsis core promoter elements shown in Figure 1 are further characterized in this study, using high-density TSS data obtained as described below.

Obtaining a cap signature of cDNA using Moloney murine leukemia virus (MMLV) reverse transcriptase

To perform high-density mapping of TSSs with the aid of the MPSS method, we needed to prepare full-length cDNAs, and

© 2009 The Authors

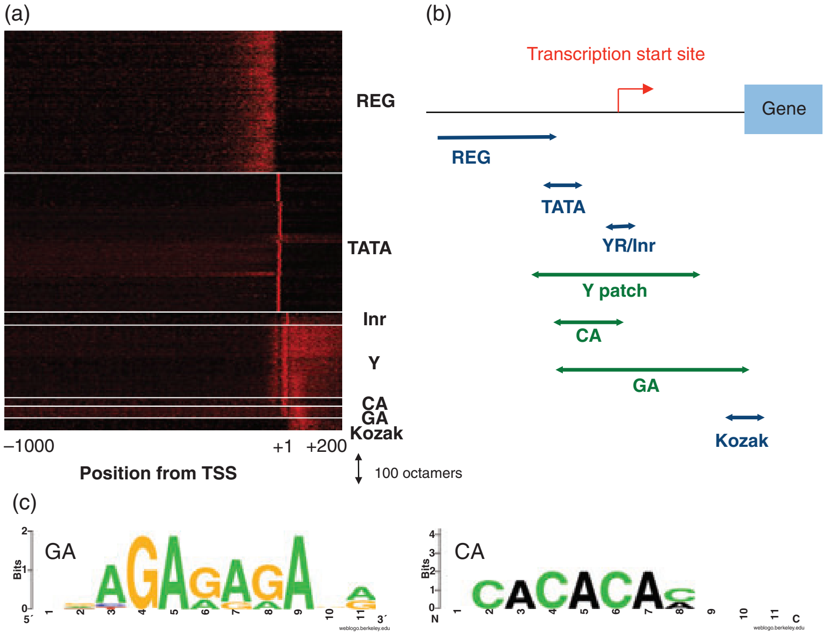

Figure 1. Arabidopsis core promoter elements identified by LDSS analysis.
LDSS analysis is a computational methodology for motif discovery of position-sensitive promoter constituents (Yamamoto *et al.*, 2007b).
(a) LDSS-positive octamer sequences were clustered according to their distribution profiles. Regions of high accumulation are shown in red and sparse regions in black. The distribution profiles of 762 octamers are bundled. The horizontal axis indicates the relative position from TSS. The double-headed arrow indicates the width for 100 octamers. REG, regulatory element group; Inr, initiator group; Y, Y (pyrimidine) Patch group; CA, CA group; GA, GA group; Kozak, Kozak group for translational initiation consensus. Octamer sequences for each group are listed in Table S1.
(b) Schematic representation of the clustered elements. The approximate positioning of each group is illustrated. Blue elements are shared with mammalian promoters, and green ones are plant-specific.
(c) Motif expression of GA and CA groups. Octamer sequences of GA and CA groups were aligned, and WebLogo was used for motif visualization. These motif expression are not the definition of these groups.

insert the 5′ ends of the cDNAs into a tag vector specifically designed for MPSS (Brenner *et al.*, 2000b). Before starting the MPSS analysis, we did pilot experiments to examine the steps described above, including cap-trapper full-length cDNA synthesis and single-strand linker ligation at the 5′ ends (cap-trapper SSLM; Shibata *et al.*, 2001), using test genes *PSAG* and *PSAH*::*LUC*. After cap-trapper SSLM, the 5′ regions of the cDNAs for *PSAG* and *PSAH*::*LUC* were specifically amplified by PCR, cloned and subjected to sequence analysis. By comparing their 5′ terminal sequences with the corresponding genomic sequences, we found that most of the clones had the addition of one or two guanine residue(s) (Figure S1a), and some contained more than one linker. These results suggest that most of the CT–MPSS tags will have additional Gs in front of the cDNA sequence, and that some tags will not map to the genome because of the insertion of additional linker(s).

Addition of Gs at the 5′ end of full-length cDNAs is also reported in CAGE analysis (Carninci *et al.*, 2006), and this is due to the cap-dependent deoxycytidyl transferase activity of MMLV reverse transcriptase (Schmidt and Mueller, 1999) rather than the template-independent terminal deoxynucleotidyl transferase (TdT) activity (Potter *et al.*, 2003). The efficiency of the activity is affected by experimental conditions and also by the version of MMLV reverse transcriptase used (Schmidt and Mueller, 1999). Figure S1(a) shows that our conditions and the use of Superscript III are very effective for deoxycytidyl transferase activity. This ‘cap signature’, the addition of Gs at the 5′ end of full-length cDNA sequence, can be used to confirm that reverse transcription reached the cap site (Schmidt and Mueller, 1999).

High-density mapping of Arabidopsis TSSs by CT–MPSS

Arabidopsis RNA from seedlings, roots, flowers, stems and etiolated seedlings was subjected to CT–MPSS analysis (Figure S2). This is a modified MPSS method that allows sequencing of the 5′ end of full-length cDNAs and that consists of the following steps: preparation of cap-trapped cDNAs (Carninci, 2002), ligation with single-strand linkers (SSLM; Shibata *et al.*, 2001), cloning of the 5′ ends of the cDNAs into a tag vector (Brenner *et al.*, 2000b), and MPSS analysis for ultrahigh-throughput parallel sequencing (Brenner *et al.*, 2000a).

Through these analyses, we obtained 372 469 tags of the 5′ end 18 bp sequence of full-length cDNAs, which consisted of 81 114 sequences (tag species) (Table 1). Subsequently, they were classified into three categories: tags with a cap signature that mapped to a single genomic locus, tags

Table 1 Statistical data for the CT–MPSS analysis

Description	Number of tags	Number of tag species or genomic loci
Successfully sequenced tags	372 469	
Tag species		81 114
Total number of tags without cap signature mapped to single locus	35 979	
Total number of cap-signature tags mapped to single locus	158 237	
Cap-signature tag species mapped to single locus		38 709
TSSs		38 311
TSSs overlapping with public data[a]		7603
Newly identified TSSs		30 708
Genes overlapping with public data[a]		7078
Newly identified genes[a]		2549
Number of TSS clusters[b]		24 453
Mean number of TSS tags per cluster	6.5	

[a]Comparison with 5′ end mapping of full-length cDNA (RAFL clones). Allows a 1 bp gap for detection of overlapping TSSs.
[b]TSS clusters are defined as groups of TSSs that are separated from nearby clusters by more than 100 bp.

without a cap signature but mapped to a single genomic locus, and the remainder, including tags mapped to multiple sites or those that failed to map to any site. For the first two categories, we compared how frequently tags mapped to genic promoter regions (−1000 to +1, relative to the translation start site). Frequency of mapping to genic promoter regions was much higher in the tags with a cap signature (61.7%) than in tags without it (18.3%), as shown in Figure S1(b). These results suggest that the tags with a cap signature are much less contaminated with uncapped cDNA ends than those without a cap signature. Given these results and those shown in Figure S1, we decided to use only tags with a cap signature for further analysis in order to obtain high-quality TSS data, but at the price of decreased data amount. Based on this policy, 158 237 TSS tags corresponding to 38 311 genomic loci were selected (Table 1). This *in silico* purification of clones that represent TSS full-length cDNAs provided higher-quality TSS information than when all the 5′ end sequences of full-length cDNA clones were used. If we apply our criterion for the selection of TSS tags to the established 92 624 clones of full-length cDNAs (Seki *et al*., 2002), only 24 155 clones remain.

The identified TSS loci were then compared with known TSSs that had been determined based on the information from RIKEN Arabidopsis full-length (RAFL) cDNAs (RIKEN Arabidopsis Genome Encyclopedia (RARGE), http://rarge.gsc.riken.jp/). This comparison revealed that, of 38 311 TSSs identified by CT–MPSS, only 7603 sites coincided with the RAFL TSSs, and 30 708 sites were newly identified TSSs (Table 1). They included additional start sites (TSSs) for structural genes whose TSSs had been identified from full-length cDNA information (7078 genes), as well as novel TSSs whose cognate TSSs had not been identified previously (2549 genes).

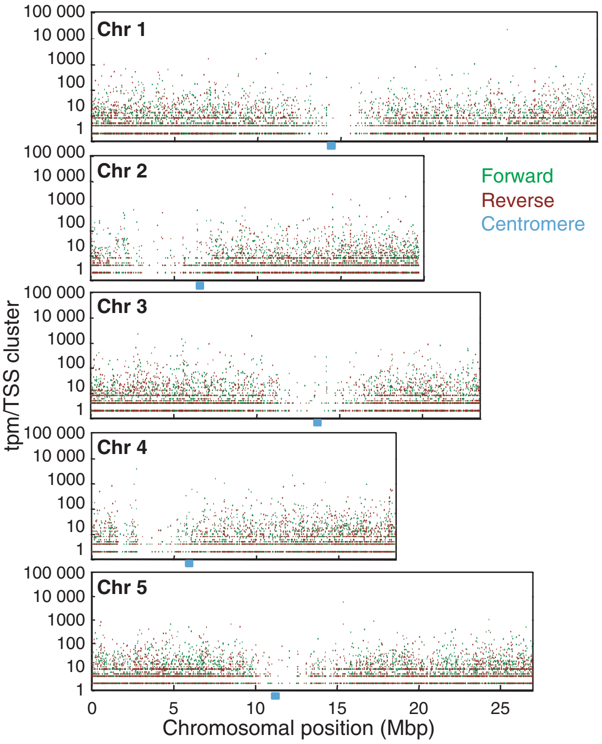

Figure 2. Global distribution of the identified promoters.
A total of 24 453 TSS clusters identified by CT–MPSS analysis were mapped to the Arabidopsis genome. The vertical axis indicates the expression level (tpm) for each TSS cluster (promoter). The direction of each promoter is indicated by its color (green for forward; red for reverse). Blue boxes indicate the position of the centromere (Arabidopsis Genome Initiative, 2000).

In many cases, multiple TSSs are clustered in a core promoter region in which the adjacent TSSs are located within 10 bp. Given this, TSS clusters were defined as those separated from a nearby TSS by more than 100 bp, and 38 311 TSS sites were grouped into 24 453 clusters. One cluster contains 6.5 TSS tags on average (Table 1). The identified TSS clusters are distributed over the genome except for pericentromeric regions (Figure 2).

Subsequently we assigned the established TSS clusters to gene models. In this study, assignment was made if a TSS cluster had the same orientation as the coding sequences and the peak TSS, the most major TSS in a cluster, was located downstream of −1000 bp relative to the translational start position. The resultant 10 285 clusters assigned to coding sequences are referred to as 'genic promoters' in this study, and subjected to further analysis. The number of genic promoters (10 285) was larger than the number of corresponding genes (9627) (Table 2), suggesting that most of the observed surplus is due to alternative promoters. The remaining TSS clusters (14 168) include

Table 2 Summary for genic promoters

Description	Number of tags, tag species or genomic loci
TSS clusters	24 453
Clusters associated with gene[a]	10 285
Genes with TSS cluster	9627
Mean number of TSS clusters per gene	1.068

[a]Allows multiple clusters per gene.

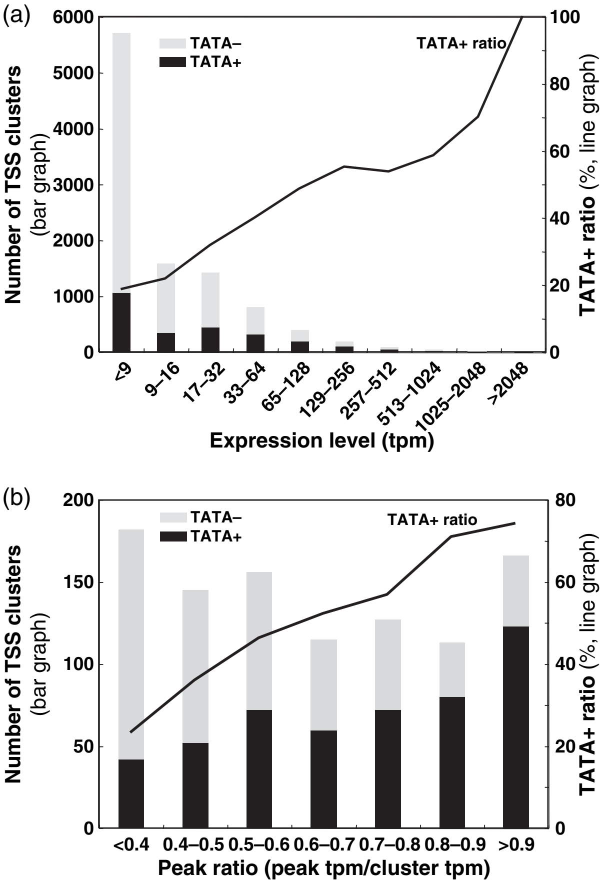

Figure 3. TATA-containing promoters and expression characteristics.Genic promoters were used for the analysis.

(a) Expression level and TATA box. The horizontal axis shows the total tpm (tags per million, relative tag abundance in a tag library) of a TSS cluster. The bar graph indicates the number of TATA-positive or -negative TSS clusters (promoters). The line graph shows the percentage of TATA-positive promoters in each fraction. The number of TATA-positive/negative promoters for '513–1024', '1025–2048' and '>2048' are 24/17, 7/3 and 7/ 0, respectively.

(b) Convergence of TSS and TATA boxes. The horizontal axis shows the peak ratio, i.e. the relative strength of a peak TSS within the TSS cluster. A cluster with a high peak ratio has a sharp peak, and one with a low peak ratio corresponds to a broad-type cluster. TSS clusters corresponding to the indicated peak ratio are divided into TATA-positive (black bar) and -negative groups (gray bar). The line graph indicates the percentage of TATA-positive promoters in each fraction. TSS clusters with expression levels of 50 tpm or more were used for calculation of the peak ratio.

those of antisense and internal promoters in various genes, and those that were not assigned to any gene models defined by TAIR (http://www.arabidopsis.org/). We tentatively denote the unidentified promoters that produce the latter class of TSS clusters, which are mostly found in intergenic regions, as orphan promoters. Antisense, internal and orphan promoters were also observed in mammalian studies (Hüttenhofer *et al.*, 2005; Johnson *et al.*, 2005; Carninci *et al.*, 2006).

Evaluation of TSS distribution

The TSS data obtained by CT–MPSS were compared with the results from conventional primer extension analysis. As shown in Figure S3(a), primer extension analysis of the transcripts from a highly expressed gene, At1g67090 (*RBCS1A*), identified one strong TSS accompanied by two minor TSSs (lower panel). The results of CT–MPSS, shown in the upper panel, match those of the primer extension analysis. In addition, CT–MPSS demonstrated 15 additional TSSs that were too weak to be detected by primer extension analysis. The results of CT–MPSS indicate that this gene has one major TSS surrounded by 16 minor TSSs; this feature is consistent with the results from conventional primer extension analysis, but much more informative.

Comparison of CT–MPSS data with publicly available full-length cDNA information also gave reasonable results (Figure S3b). From these comparisons, we concluded that the established CT–MPSS method, including the mapping process, provides reliable results.

Characterization of TATA-type promoters

We then analyzed the expression characteristics of TATA-containing genic promoters. TATA-positive promoters were identified by the presence of Arabidopsis TATA sequences (Yamamoto *et al.*, 2007b) (see also Experimental procedures) at appropriate positions relative to the peak TSS. To evaluate the expression level of the identified promoters, the abundance of individual TSS tags was summed up within each TSS cluster, and is represented as a tpm value (tpm, tags per million, is an indicator of tag abundance in MPSS analysis; Brenner *et al.*, 2000a). As shown in Figure 3(a), the occurrence rate of TATA-containing promoters in the genic promoters (line graph) greatly increases in accordance with the increase in tag abundance (tpm), i.e. mRNA expression level. Therefore, for genic promoters, there is a strong positive correlation between the expression level and possession of a TATA box.

Next, we examined the convergence of TSSs within a cluster. The peak ratio, as shown in the horizontal axis of Figure 3(b), represents the relative height of the peak TSS within a TSS cluster. A high value means the cluster has a sharp dominant peak. In contrast, a low value means that the cluster belongs to the so-called broad-type promoters (Carninci *et al.*, 2006), in which TSSs are scattered around

© 2009 The Authors

without any dominant peak. Figure 3(b) shows that the frequency of TATA-containing promoters (line graph) increases in accordance with the increase in peak ratio, which indicates sharpening of the TSS peak.

These results suggest that the TATA element makes plant promoters with high expression activity and highly convergent TSSs, representing the promoters with a sharp dominant peak TSS, consistent with mammalian studies. This conservation with characteristics of the mammalian TATA box has been revealed.

Effect of local sequences around TSS

We analyzed how local sequences around the peak TSS affect promoter characteristics. In our previous study, most plant promoters were found to have a consensus sequence where the −1 position is pyrimidine (Y:C or T) and +1 is purine (R:A or G); this consensus is called the YR rule (Yamamoto *et al.*, 2007b). This is a relaxed form of the Inr motif, which have longer consensus sequences (Butler and Kadonaga, 2002; Nakamura *et al.*, 2002). In this study, we considered several Inr-like TSS consensus sequences, YR, YYR, TCA, TCAY and YTCAY (TSS underlined). The latter sequences are more stringent and closer to Inr.

The relationship between these Inr-like motifs and the expression level is shown in Figure 4(a). First, we noticed that all five sequences are overrepresented at the TSSs. For example, the observed frequency of YR is higher than 70% (Figure 4a), whereas the expected frequency for random dimer sequences is 25%. Such enrichment is also observed for other Inr-like motifs. Figure 4(a) shows that the occurrence rates of all the Inr-like sequences increase in accordance with the increase in expression level. In particular, the rate of YTCAY increases about sixfold from the fraction '<5 tpm' to the fraction '513–1024 tpm'. These results indicate that a positive relationship exists between the expression level and the presence of Inr-like motifs. This feature is similar to that of the TATA box (Figure 3a).

Figure 4(b) shows the relationship between Inr-like motifs and the sharpness/broadness of the TSS distribution. Although the frequency of YR is saturated at the peak ratio of 0.5–0.6, the frequencies of the other Inr-like elements increase in accordance with the increase in peak ratio, indicating that a positive correlation exists between these Inr-like elements and the sharpness of TSS peak. These results support the view that Inr is the site of pol II interaction, rather than a *cis*-element for transcription factors as reported previously (Butler and Kadonaga, 2002). According to this view, Inr interacts with pol II and stringent Inr-like elements make sharp-peak TSS clusters, whereas unrefined promoters that do not match the YR rule fail to initiate effective transcription, resulting in low expression levels. This agrees with the results shown in Figure 4(a,b).

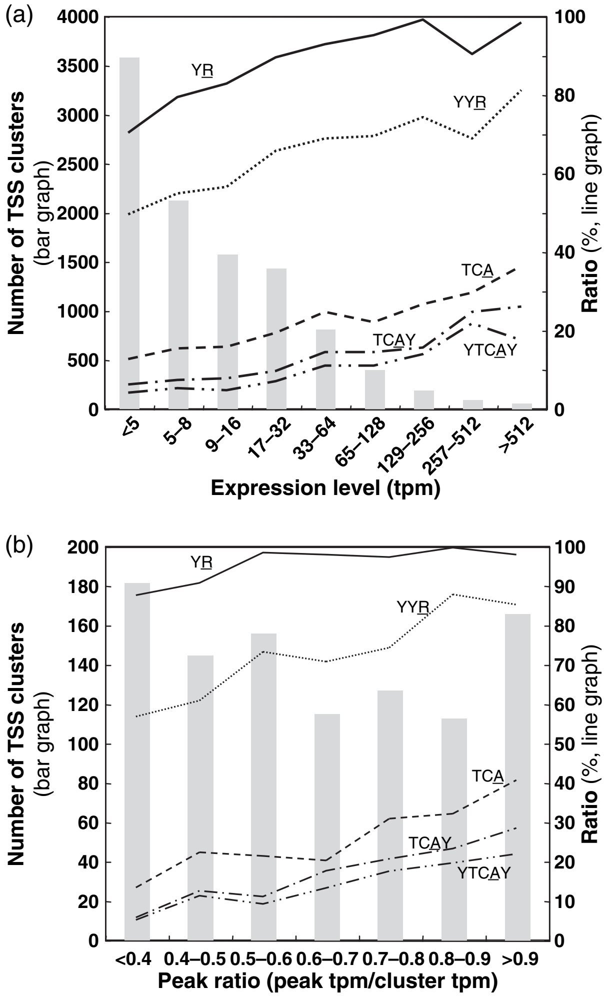

Figure 4. Sequence around TSSs and expression characteristics.
The bar graph indicates the number of promoters with indicated expression levels, and the line graph shows the percentage of promoters corresponding to each TSS type. Y, pyrimidine (C or T); R, purine (A or G). The underlined residue in the motifs is +1. The results for genic promoters are shown.
(a) Expression level and TSS sequence. Promoters were sorted by expression level as shown in the horizontal axis. For all Inr-like motifs, the increase in the ratio for '>512' compared to '<5' is significant (<5%).
(b) Shape of a TSS cluster and TSS sequence. Promoters were sorted by peak ratio, i.e. the tpm of the peak TSS/tpm of the cluster.

Combination of the TATA box and Inr-like motifs

We subsequently examined the relationship between the TATA box and Inr-like elements. Statistical analysis of co-occurrence revealed that the TATA box and all the Inr-like motifs tend to be found together (Table 3). In addition, the frequency of TATA-containing promoters increases as the Inr-like motif becomes more stringent (Figure 5). As seen in Figure 5, the overall frequency of TATA-positive promoters is 25.1%, and this increases to 41.0% in YTCAY-type promoters. Therefore, the TATA box and Inr-like elements appear to have a co-evolutionary relationship, suggesting

© 2009 The Authors

Table 3 Co-existence of TATA boxes and Inr-like motifs

Motif[a]	TATA-positive[b]	TATA-negative[b]	TATA-positive rate (%)	*P* value[c]	Correlation[d]
YR+	2322	5930	28.1	5.92e–53	Positive
YR−	256	1777	12.6		
YYR+	1696	4261	28.2	3.49e–18	Positive
YYR−	902	3446	20.7		
TCA+	600	1104	35.2	9.38e–25	Positive
TCA−	1978	6603	23.1		
TCAY+	360	511	41.3	3.23e–28	Positive
TCAY−	2218	7196	23.6		
YTCAY+	250	360	41.0	6.80e–19	Positive
YTCAY−	2328	7346	24.1		
Total	2578	7707	25.1	–	–

[a] Motif at peak TSS in a cluster.

[b] Number of promoters (=TSS clusters) that have or do not have the core elements. The total number of analyzed promoters is 10 285.

[c] Probability based on the assumption of independent distributions calculated by Fisher's exact probability test.

[d] A *P* value >0.01 is considered to indicate no correlation, and is expressed as 'none'. Both 'positive' and 'negative' mean the presence of significant correlation. 'Negative' correlation means a mutually exclusive relationship.

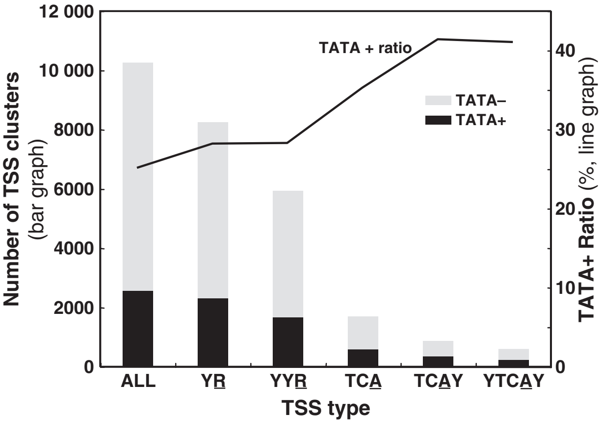

Figure 5. Concentration of TATA boxes in Inr-like TSSs.
Genic promoters were classified according to the local sequence around the peak TSS in a cluster, as shown on the horizontal axis. Y, C or T; R, A or G. The peak TSS in the motifs is underlined. From left to right, the TSS type becomes more strict and closer to the Inr motif; promoters of the right-hand TSS types are subsets of the left-hand TSS types. 'ALL' represents all the genic promoters. The bar graph indicates the number of TATA-positive (black) and -negative (gray) promoters among the indicated TSS types, and the line graph shows the percentage of TATA-positive promoters in each fraction.

that TATA-positive promoters are exposed to more selection pressure towards creation of a stringent Inr motif around TSSs than TATA-less promoters are, or, conversely, that Inr motifs attract the TATA box during evolution. Enrichment of both the TATA box and Inr-related motifs with increasing convergence of TSSs (Figures 3b and 4b) implies that they act synergistically on TSS to bring about greater convergence, a scenario that is consistent with the suggested co-evolutionary relationship.

Characterization of Y Patch-containing promoters

Y Patches are uniquely found in plant promoters, but not in mammals (Yamamoto *et al.*, 2007a,b). As functional information on Y Patches is lacking, we analyzed the effect of this element on the promoter characteristics.

The relationship between Y Patches and mRNA expression level is shown in Figure 6(a). Y Patches show a weak positive correlation with mRNA expression level. This tendency is still observed after TATA-positive promoters are removed; therefore, this weak correlation is a characteristic of Y Patches (data not shown). Y Patches have no obvious contribution to the convergence of TSS (Figure 6b). Statistical analysis of co-occurrence revealed that Y Patches occur independently of the TATA box, but have a tendency to be accompanied by Inr-like motifs (Table 4). In addition, the frequency of Y Patch-containing promoters increases in accordance with the increased stringency of Inr-like motifs (Figure 6c), as in the case of the TATA box.

Unique characteristics of GA elements

GA and CA elements are both newly identified promoter constituents in this study (Figure 1), but they differ in abundance. GA-positive promoters amount to 21.6% of the total genic promoters, while CA-containing promoters constitute as little as 1.1%. The number of CA-containing promoters was not high enough to allow statistical analysis in this study.

Figure 7 presents a characterization of GA-containing promoters. The presence of GA elements shows a positive correlation with mRNA expression level (Figure 7a), a tendency that is essentially similar for the TATA box (Figure 3a) and Y Patch (Figure 6a). However, GA-positive promoters

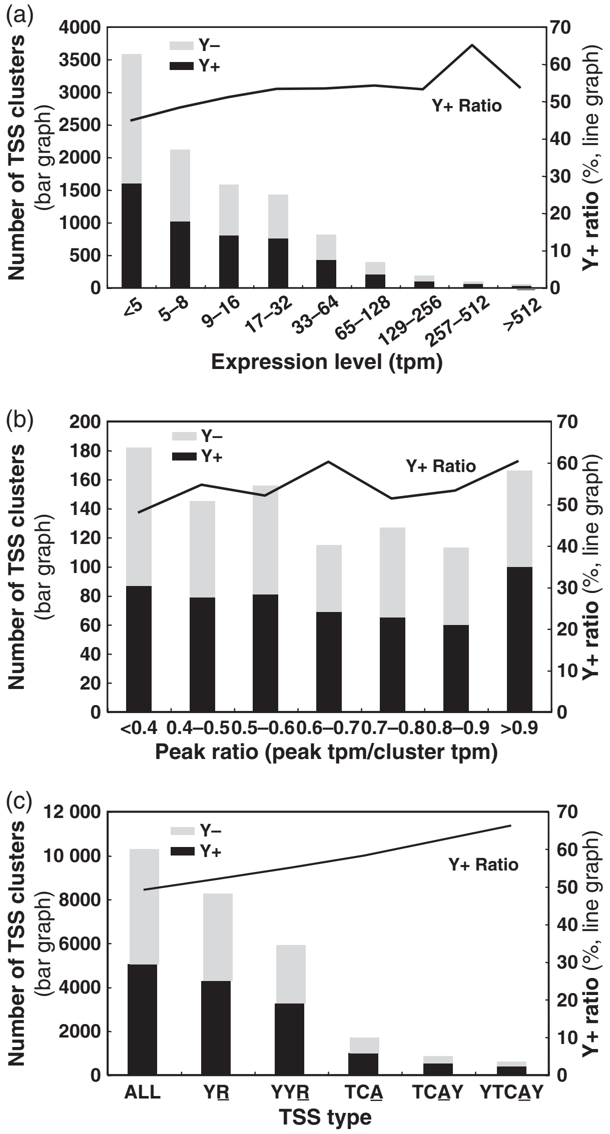

Figure 6. Characterization of Y Patch.
The bar graph indicates the number of Y Patch-positive (black) and -negative (gray) promoters in the promoter fractions as shown on the horizontal axis, and the line graph shows the percentage of Y Patch-positive promoters. The results for genic promoters are shown.
(a) Expression level and Y Patch. Promoters were sorted by expression level as shown on the horizontal axis. The significance of the increase in the Y-positive ratio of '257–512' compared with '<5' is positive, but that of '>512' is negative, using Fisher's exact test.
(b) Peak shape and Y Patch. Promoters were sorted by peak ratio, i.e. tpm for the peak TSS/tpm of the cluster.
(c) TSS sequence and Y Patch. Promoters were sorted by TSS type as indicated. Y, C or T; R, A or G. The peak TSS in the motifs is underlined.

exhibited quite different features from TATA-containing promoters with regard to the shape of TSS clusters. As shown in Figure 7(b), the abundance of GA-positive promoters tends to increase with the decrease in peak ratio, a tendency opposite to that for the TATA box (Figure 3b). Therefore, GA elements tend to occur in broad-type promoters, while TATA boxes associate with highly convergent TSS. A contrasting characteristic between GA elements and TATA boxes was also found with regard to the affinity to for Inr-like motifs. As shown in Figure 7(c), the abundance of GA elements decreases when Inr-like motifs become closer to YTCAY, a tendency opposite to that for TATA boxes (Figure 5).

A statistical co-occurrence analysis of the GA elements with TATA, Y Patch and Inr-like elements is presented in Table 5. Occurrence of the GA elements bears no relationship to TATA boxes, but, interestingly, the GA element is mutually exclusive of Y Patch and Inr-like motifs.

Taken together, these observations reveal that the Arabidopsis genome contains two opposing core promoter groups, namely the TATA group and the GA group. TATA-group promoters contain the TATA box, Y Patch and Inr elements, and produce sharp peak-type TSS clusters. The GA group tends to show broad-type TSS clusters.

TATA-type promoters are over-represented in environmental response genes

Next, we analyzed the occurrence of core promoter elements with respect to gene ontology (GO). GO at TAIR (http://www.arabidopsis.org/) allows functional annotation of genes based on several classifications, and we used the classification 'biological process' to characterize individual promoters.

As shown in Figure S4, TATA-positive promoters are relatively enriched in the GO categories 'response to stress' and 'response to abiotic or biotic stimulus'. This result is a reflection of a general tendency that promoters with 'regulated' expression, including those with tissue-specific expression, have a higher frequency of TATA-positive promoters than those with 'constitutive' expression, as reported in mammalian studies (Suzuki *et al.*, 2001; Schug *et al.*, 2005; Carninci *et al.*, 2006).

In contrast to the TATA box, the Y Patch and the GA element do not show a notable preference for gene categories. These results suggest that the TATA promoter group is somewhat specialized for specific types of transcription, while the GA group is rather neutral.

These characteristics of the TATA and GA promoter groups were further confirmed using microarray data (Yamamoto *et al.*, 2004). Table 6 summarizes the respective occurrence of TATA, GA and coreless (absence of either TATA, GA, Y or CA octamers) promoter groups in promoters that are responsive or non-responsive to several abiotic stresses. These abiotic stress-responsive promoters are significantly rich in TATA-positive promoters, compared to the average occurrence of the TATA group for total genes. It is also noted that occurrence of the TATA group is significantly low in promoters that are not responsive to these stresses. Therefore, the difference in TATA frequency between abiotic stress-responsive and non-responsive

Table 4 Co-existence of Y Patches

Element type[a]	Y Patch-positive[b]	Y Patch-negative[b]	Y-positive rate	*P* value[c]	Correlation[d]
TATA+	1227	1351	47.6%	0.111	None
TATA−	3809	3898	49.3%		
YR+	4289	3963	52.0%	4.39e–35	Positive
YR−	747	1286	36.7%		
YYR+	3257	2680	54.9%	9.48e–186	Positive
YYR−	1779	4369	28.9%		
TCA+	991	713	58.2%	1.78e–16	Positive
TCA−	4045	4536	8581		
TCAY+	541	330	62.1%	4.71e–16	Positive
TCAY−	4495	4919	47.7%		
YTCAY+	404	206	66.2%	8.98e–19	Positive
YTCAY−	4632	5043	47.9%		
Total	5036	5249	49.0%		

[a–d]See footnote to Table 3.

promoters is as much as threefold. Compared to this striking characteristic of TATA group, the tendency of the GA group is not so clear (Table 6), as in the case of the analysis with GO categorization (Figure S4). However, statistically significant enrichment was detected for GA and coreless groups in non-responsive promoters (Table 6).

DISCUSSION

In this study, we have obtained genome-wide high-density TSS information for Arabidopsis*,* and the dataset obtained enables quantitative analysis of plant promoters. Selection of the cap-trapper TSS tags with the cap signature at their 5′ termini allowed us to obtain high-quality TSS information. In parallel, we comprehensively identified core promoter elements using a genome-wide bioinformatics approach. The core elements were characterized based on (i) their expression level in a promoter-based manner, distinct from conventional gene-based way, (ii) the shape of the TSS clusters (peak ratio), and (iii) the position of the peak TSS within a promoter. The CT–MPSS analysis enabled us to obtain all this information. This report describes successful characterization of the TATA box and Inr-like elements, as well as recently recognized core promoter elements such as Y Patches and the GA element.

In mammals, sharp-peak TSS clusters are associated with a TATA box, and broad-type clusters are associated with CpG islands (Suzuki *et al.*, 2001; Carninci *et al.*, 2006). TATA boxes are conserved in plant promoters (Joshi, 1987), but the lack of CpG islands in plant promoters (Yamamoto *et al.*, 2007a) raises the question of what their counterparts are in plants. Our combined bioinformatics and functional genomics analyses suggest that the GA group partially plays the role of mammalian CpG islands. Two lines of evidence support this idea: (i) absence of the GA element in the TATA group, and (ii) preference for broad-type TSS clusters in GA-positive promoters. However, there is a difference between CpG and GA: CpG islands are under the control of DNA methylation (Bird and Wolffe, 1999; Bird, 2002), but there are no methylation targets in the GA element, suggesting that methylation does not occur on the GA element. Another difference is that human CpG-type promoters are preferred by housekeeping genes (Schug *et al.*, 2005; Saxonov *et al.*, 2006), but Arabidopsis GA-type promoters showed no such preference. This might mean that the plant GA-type promoters are not as differentiated as the mammalian CpG type.

There are far fewer members (octamers) in the GA group than there are TATA boxes and Y Patches. However, as many as 21.6% of the genic promoters have the GA element (deduced from Table 5), which is comparable to the proportion possessing a TATA box (25.1%, Table 3). Hence, the GA element should not be overlooked in elucidating the structure–function relationship of plant promoters. On the other hand, a minority of promoters (1.12%) have the CA element, suggesting that this is a specialized element rather than a general core promoter constituent. Conservation of this element between Arabidopsis and rice suggests it has biological importance, but the paucity of identified promoters in this group has made it difficult to analyze their expression characteristics statistically.

Figure 8 summarizes the relationship between the core promoter elements examined in this study. Plant promoters can be divided into those of the TATA group (gray oval in the figure) and those of the GA group, which are mutually exclusive (Table 5). Promoters of the TATA group have TSS clusters with a sharp peak and high expression (Figure 3), and are associated with TATA boxes, Inr-like elements and Y Patches (Tables 3 and 4). Promoters of the GA group have broad-type TSS clusters, in contrast to the TATA group. Both groups show high expression (Figures 3a, 4a, 6a and 7a). The characteristics of these two groups, as well as their different distribution profiles among GO categories (Figure S4) and different expression profiles (Table 6), suggest that they have distinct roles.

In both plants and mammals, the genic promoters consist not only of TATA-type promoters, but also of broad-type

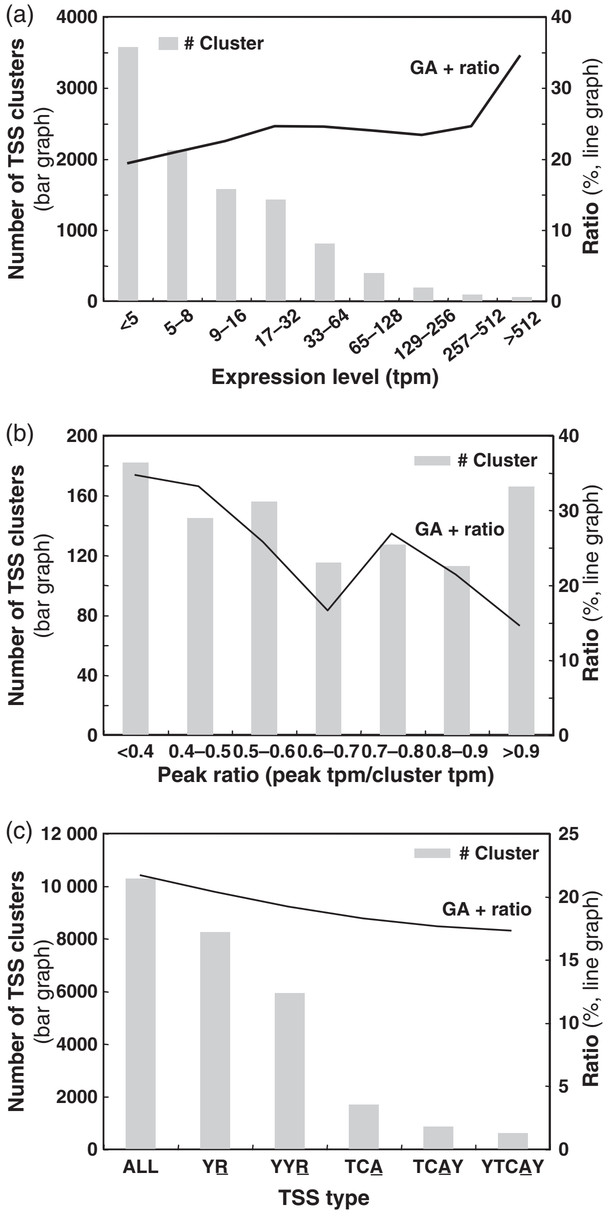

Figure 7. Characterization of GA and CA elements.

The bar graph indicates the number of promoters in each fraction as shown on the horizontal axis, and the line graph shows the percentage of GA-positive promoters in each fraction. The results for genic promoters are shown.

(a) Relationship between expression level and the indicated core elements. Promoters were sorted by expression level as shown in the horizontal axis. The significance of the increase in GA-positive ratios from '257–512' to '<5' is positive.

(b) Relationship between peak shape and indicated core elements. Promoters were sorted by peak ratio, i.e. tpm for the peak TSS/tpm of the cluster.

(c) Relationship between TSS type and the indicated core elements. Promoters were sorted by TSS type as indicated. Y, C or T; R, A or G. The peak TSS in the motifs is underlined.

TSSs: the GA group in plants and the CpG type in mammals. What is the advantage of possessing alternative core promoter groups in these organisms? One hypothesis is that the TATA type has some disadvantages in some situations, e.g. constitutive or embryonic expression is avoided by TATA-type promoters (Schug *et al.*, 2005). In this case, the presence of a complementary type of promoter would provide benefits to organisms. Further studies are required in order to reveal the biological importance of core promoter heterogeneity.

EXPERIMENTAL PROCEDURES

Extraction of core promoter elements

Extraction of LDSS-positive octamer sequences based on public TSS information (http://rarge.gsc.riken.jp/) was performed as described previously (Yamamoto *et al.*, 2007b). Clustering of the extracted octamers according to their distribution profiles was performed using CLUSTER software (http://rana.lbl.gov/EisenSoftware.htm) with the k-means method, and visualized by TREEVIEW (http://rana.lbl.gov/EisenSoftware.htm). The overlapping distributions of the Y, CA, GA and Kozak groups (Figure 1) made it difficult to separate them completely using their positional information alone. Therefore, clustered groups of TATA, Y Patches and GA and CA octamers were each subjected to further selection to exclude contaminating sequences as described in Appendix S1. Selected octamer sequences are shown in Table S1. Motif expression of GA and CA groups was achieved using WEBLOGO (Crooks *et al.*, 2004) (http://weblogo.berkeley.edu/) after alignment of the octamers by CLUSTALW (http://align.genome.jp/clustalw/). Identification of promoter elements was performed by searching for respective LDSS-positive octamers at an appropriate position relative to the peak TSS: TATA, −45 to −18; Y Patch, −50 to +50; CA, −35 to −1; GA, −35 to +75. These positional restrictions were determined according to their positional characteristics as shown in Figure 1.

It should be noted that our list of TATA sequences is rather stringent based on the policy of our previous study (Yamamoto *et al.*, 2007b). Therefore, the TATA-positive promoters all have strong TATA elements; the promoters referred to as TATA-negative in this study could be a mixture of absolute TATA-less promoters and promoters containing potential or weak TATA boxes.

CT–MPSS analysis

An outline of the experiments is shown in Figure S2. Total RNA of *Arabidopsis thaliana* ecotype Columbia was extracted from the aerial parts of 5-day-old etiolated seedlings (Yamamoto *et al.*, 1998), the aerial parts of 9-day-old light-grown seedlings (Kimura *et al.*, 2001), roots grown in Gamborg's B5 liquid medium (Hauge and Goodman, 1992) under dim light and mild rotation, and inflorescences including flowers, flower buds, siliques, premature seeds and stems from 1 to 2-month-old soil-grown plants.

Total RNA (250 μg) from each preparation was mixed to give a total of 1 mg RNA, and poly(A)+ RNA was extracted using an mRNA purification kit (Absolutely mRNA purification kit, Stratagene, http://www.stratagene.com/) according to the manufacturer's instructions. For the cap-trapper method, 19 μg of poly(A)+ RNA was used, and single-stranded cDNA was prepared (Carninci, 2002). Reverse transcription was achieved using random hexamers as primers and Superscript III (Invitrogen, http://www.invitrogen.com).

Details of the method of linker ligation to single-stranded cDNA based on the CAGE method (Shiraki *et al.*, 2003) are given in Appendix S1. Cloning into the tag vector, preparation of Megaclone microbeads (Takara Bio Inc., http://www.takara-bio.com), and MPSS reading analysis have been described previously (Brenner *et al.*, 2000b; Stolovitzky *et al.*, 2005). Sequencing by MPSS is described in Appendix S1.

© 2009 The Authors

Table 5 No partner of GA

Element type[a]	GA-positive[b]	GA-negative[b]	GA-positive rate (%)	*P* value[c]	Correlation[d]
TATA+	530	2048	20.6	0.142	None
TATA–	1692	6015	22.0		
Y+	847	4189	16.8	4.80e–31	Negative
Y–	1375	3874	26.2		
YR+	1676	6576	20.3	3.01e–10	Negative
YR–	546	1487	26.9		
YTCAY+	105	505	17.2	0.00613	Negative
YTCAY–	2117	7558	21.9		
Total	2222	8063	21.6	–	–

[a–d]See footnote to Table 3.

Table 6 TATA-type promoters are rich in stress responsive genes

	Promoters[a]	TATA-positive	GA-positive	Coreless[b]
Genic promoters[c]		25.1%	21.6%	27.6%
		NA	NA	NA
	10 285	2587	2222	2834
High light induction[d]		**46.9%**	23.5%	19.8%
		P = 2.03E-5	*P* = 0.385	*P* = 0.0714
	81	38	19	16
Drought induction[e]		**51.4%**	22.9%	17.1%
		P = 7.71E-4	*P* = 0.495	*P* = 0.115
	35	18	8	6
Induction by H_2O_2 treatment[f]		**46.3%**	25.3%	**19.8%**
		P = 6.46E-9	*P* = 0.149	*P* = 0.0149
	162	75	41	32
No response to high		**16.5%**	**23.2%**	**29.3%**
light/drought/H_2O_2		*P* = 1.75E-22	*P* = 0.0452	*P* = 0.0424
	2669	440	618	781

The core types of stress-inducible promoters and those with no response to the stresses were analyzed. In each of the last three columns, the percentage of the corresponding core type of promoters among a gene group identified from microarray data, the probability of this core type based on the assumption of neutral distribution, and the number of promoters, are shown from the top to the bottom, respectively. When the difference of a ratio to the global control is significant, as determined using Fisher's exact test ($P < 0.05$, one-sided), the percentage is shown in bold. Responses to high light stress, drought stress and H_2O_2 application were identified by microarray analysis (Yamamoto *et al.*, 2004). Genes with more than threefold activation in response to the indicated abiotic stresses were selected as inducible genes, and those with 0.75–1.5-fold induction by any of the three stresses were chosen as 'no response' genes. Detection of core elements was achieved with the aid of the peak TSSs for detection of position-dependent promoter elements determined by CT–MPSS and LDSS-positive octamers.

[a]Promoters for which TSS information is available from CT–MPSS data were included in the analysis.

[b]Promoters with neither TATA, GA, Y, or CA octamers were considered as 'coreless'.

[c]Control set to show the global average.

[d]High light treatment was performed at 150 W m^{-2} (750 μE m^{-2} sec^{-1}) for 3 h.

[e]Drought treatment was performed for 3 h.

[f]A 3% H_2O_2 solution was sprayed onto the plants 3 h before harvesting.

Mapping of TSS tags to the Arabidopsis genome

All the tag species were mapped to the Arabidopsis genome sequence (TAIR6, http://www.arabidopsis.org/) using 18 bp or shortened sequences after deletion of one or two Gs at the 5′ end. Mapping was performed using custom Perl scripts under the condition of no mismatch. Tag species that mapped to multiple loci were ignored and not used in this study. The relationship between the mapped positions and the coding sequence was determined based on the TAIR6 annotation information. Contamination of tags by organellar RNA and nuclear-coded RNA gene products such as rRNA was quite rare (2.5×10^{-3} and 6.7×10^{-5}, respectively), indicating that cap-containing RNA species were strictly selected. Statistical information on mapping is summarized in Table 1.

TSS tags were considered to be clustered when separated from a nearby TSS by more than 100 bp, and clusters were considered to be associated with genes when located <1 kb upstream from a translation initiation codon.

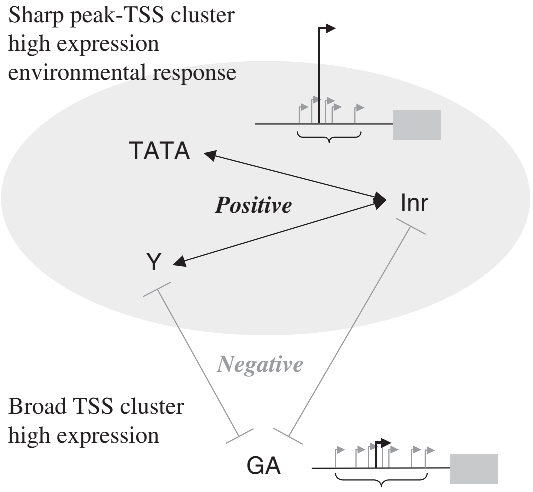

Figure 8. Relationship between core promoter elements.
The black arrow indicates co-localization in a promoter ('positive'). By contrast, a gray bar with flat ends indicates absence of co-localization in a promoter ('negative'). Given the co-localization and shared expression characteristics, elements in the gray oval are considered to form a single group, acting in a cooperative manner. This model is further supported by the results shown in Figures 4, 5c and 6c.

The 5′ end of the sequence of RAFL cDNA (http://rarge.gsc.riken.jp/) was mapped to the genome, and 36 266 distinct TSSs were identified. TSSs from RAFL and CT–MPSS information were compared, and coinciding sites were identified. TSSs at a distance of 1 bp were considered the same in this comparison, given the technical limitations of mapping of RAFL clones.

A searchable database for TSS information described in this article is available at http://ppdb.gene.nagoya-u.ac.jp. The ppdb also integrates the LDSS information (Yamamoto *et al*., 2007b) and has been described previously (Yamamoto and Obokata, 2008). The TSS information can be downloaded from the website.

Evaluation of statistical features

After mapping of MPSS tag species to the genome sequence, the relative abundance of TSS tags in the TSS tag library (tpm) was associated with each genomic locus at 1 bp resolution. The bar graph in Figure S3(a) shows an example of mapped information. The expression level of each promoter was calculated as the sum of the tpm for each TSS cluster. Clusters with 50 tpm or more were used for evaluation of the sharpness of TSS clusters. The peak ratio was calculated as the tpm of the peak TSS/total tpm of the corresponding TSS cluster.

The other sequence analyses were achieved using Perl and C^{++} programs and Excel software (Microsoft, http://www.microsoft.com). The independence of each promoter element was examined using Fisher's exact probability test.

GO and microarray analysis

TSS clusters were associated with AGI codes, and corresponding information for GO classification was obtained from TAIR (http://www.arabidopsis.org/). The results of GeneChip (ATH1) analyses for high light, drought and H_2O_2 responses were obtained from previous work (Yamamoto *et al*., 2004). Using the GeneChip data, AGI numbers for the stress-induced and non-induced genes were identified. Their peak TSSs were searched for in CT–MPSS data. Genes with the identified peak TSSs were then subjected to detection of promoter elements as described above.

ACKNOWLEDGEMENTS

We acknowledge Dr Piero Carninci for providing information about the cap-trapper method and critical reading of the manuscript, Dr Albrecht von Arnim and Dr Minami Matsui for critical reading of the manuscript, and Mr Yoshitsugu Yamamoto and Mr Takehito Ito for technical assistance with tag mapping. This work was partially supported by Grants-in-Aid for Scientific Research in Priority Areas 'Comparative Genomics' (18017014 to J.O.), Scientific Research B (17370017 to J.O.) and C (19570035 to Y.Y.Y.), and Exploratory Research (19657016 to J.O.) from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:
Figure S1. Cap signature of the cap-trapped cDNA.
Figure S2. Outline of CT–MPSS.
Figure S3. Validation of CT–MPSS data.
Figure S4. GO and core promoter type.
Table S1. Arabidopsis octamers identified by LDSS analysis.
Appendix S1. Supplementary methods.
Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

REFERENCES

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.

Berendzen, K.W., Stuber, K., Harter, K. and Wanke, D. (2006) *Cis*-motifs upstream of the transcription and translation initiation sites are effectively revealed by their positional disequilibrium in eukaryote genomes using frequency distribution curves. *BMC Bioinformatics*, **7**, 522.

Bird, A. (2002) DNA methylation patterns and epigenetic memory. *Genes Dev.* **16**, 6–21.

Bird, A.P. and Wolffe, A.P. (1999) Methylation-induced repression – belts, braces, and chromatin. *Cell*, **99**, 451–454.

Brenner, S., Johnson, M., Bridgham, J. *et al.* (2000a) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.* **18**, 630–634.

Brenner, S., Williams, S.R., Vermaas, E.H. *et al.* (2000b) *In vitro* cloning of complex mixtures of DNA on microbeads: physical separation of differentially expressed cDNAs. *Proc. Natl Acad. Sci. USA*, **97**, 1665–1670.

Butler, J.E. and Kadonaga, J.T. (2002) The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes Dev.* **16**, 2583–2592.

Carninci, P.. (2002) Generation of full-length libraries. In *DNA Microarrays: A Molecular Cloning Manual* (Bowtell, D. and Sambrook, J., eds). Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press, pp. 647–660.

Carninci, P., Kvam, C., Kitamura, A. *et al.* (1996) High-efficiency full-length cDNA cloning by biotinylated CAP trapper. *Genomics*, **37**, 327–336.

Carninci, P., Sandelin, A., Lenhard, B. *et al.* (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.* **38**, 626–635.

Carrari, F., Frankel, N., Lijavetzky, D., Benech-Arnold, R., Sanchez, R. and Iusem, N.D. (2001) The TATA-less promoter of VP1, a plant gene controlling seed germination. *DNA Seq.* **12**, 107–114.

Corden, J., Wasylyk, B., Buchwalder, A., Sassone-Corsi, P., Kedinger, C. and Chambon, P. (1980) Promoter sequences of eukaryotic protein-coding genes. *Science*, **209**, 1406–1414.

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

Elrouby, N. and Bureau, T.E. (2000) Molecular characterization of the Abp1 5′-flanking region in maize and the teosintes. *Plant Physiol.* **124**, 369–377.

© 2009 The Authors

Hauge, B.M. and Goodman, H.M. (1992) Genome mapping in *Arabidopsis*. In *Methods in Arabidopsis Research* (Koncz, C., Chua, N.-H. and Schell, J., eds). Singapore: World Scientific Publishing, pp. 191–223.

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

Johnson, J.M., Edwards, S., Shoemaker, D. and Schadt, E.E. (2005) Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet.* **21**, 93–102.

Joshi, C.P. (1987) An inspection of the domain between putative TATA box and translation start site in 79 plant genes. *Nucleic Acids Res.* **15**, 6643–6653.

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

Kiran, K., Ansari, S.A., Srivastava, R., Lodhi, N., Chaturvedi, C.P., Sawant, S.V. and Tuli, R. (2006) The TATA-box sequence in the basal promoter contributes to determining light-dependent gene expression in plants. *Plant Physiol.* **142**, 364–376.

Molina, C. and Grotewold, E. (2005) Genome wide analysis of Arabidopsis core promoters. *BMC Genomics*, **6**, 25.

Mukumoto, F., Hirose, S., Imaseki, H. and Yamazaki, K. (1993) DNA sequence requirement of a TATA element-binding protein from Arabidopsis for transcription in vitro. *Plant Mol. Biol.* **23**, 995–1003.

Nakamura, M., Tsunoda, T. and 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.

Potter, J., Zheng, W. and Lee, J. (2003) Thermal stability and cDNA synthesis capacity of SuperScript III reverse transcriptase. *Focus*, **25**, 1.

Saxonov, S., Berg, P. and Brutlag, D.L. (2006) A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl Acad. Sci. USA*, **103**, 1412–1417.

Schmidt, W.M. and Mueller, M.W. (1999) CapSelect: a highly sensitive method for 5′ CAP-dependent enrichment of full-length cDNA in PCR-mediated analysis of mRNAs. *Nucleic Acids Res.* **27**, e31.

Schmidt, K., Heberle, B., Kurrasch, J., Nehls, R. and Stahl, D.J. (2004) Suppression of phenylalanine ammonia lyase expression in sugar beet by the fungal pathogen *Cercospora beticola* is mediated at the core promoter of the gene. *Plant Mol. Biol.* **55**, 835–852.

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

Seki, M., Narusaka, M., Kamiya, A. *et al.* (2002) Functional annotation of a full-length Arabidopsis cDNA collection. *Science*, **296**, 141–145.

Shahmuradov, I.A., Gammerman, A.J., Hancock, J.M., Bramley, P.M. and Solovyev, V.V. (2003) PlantProm: a database of plant promoter sequences. *Nucleic Acids Res.* **31**, 114–117.

Shibata, Y., Carninci, P., Watahiki, A., Shiraki, T., Konno, H., Muramatsu, M. and Hayashizaki, Y. (2001) Cloning full-length, cap-trapper-selected cDNAs by using the single-strand linker ligation method. *BioTechniques*, **30**, 1250–1254.

Shiraki, T., Kondo, S., Katayama, S. *et al.* (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl Acad. Sci. USA*, **100**, 15776–15781.

Smale, S.T. and Baltimore, D. (1989) The 'initiator' as a transcription control element. *Cell*, **57**, 103–113.

Smale, S.T. and Kadonaga, J.T. (2003) The RNA polymerase II core promoter. *Annu. Rev. Biochem.* **72**, 449–479.

Stolovitzky, G.A., Kundaje, A., Held, G.A., Duggar, K.H., Haudenschild, C.D., Zhou, D., Vasicek, T.J., Smith, K.D., Aderem, A. and Roach, J.C. (2005) Statistical analysis of MPSS measurements: application to the study of LPS-activated macrophage gene expression. *Proc. Natl Acad. Sci. USA*, **102**, 1402–1407.

Suzuki, Y., Tsunoda, T., Sese, J. *et al.* (2001) Identification and characterization of the potential promoter regions of 1031 kinds of human genes. *Genome Res.* **11**, 677–684.

Yamamoto, Y.Y. and Obokata, J. (2008) ppdb: a plant promoter database. *Nucleic Acids Res.* **36**, D977–D981.

Yamamoto, Y., Tsuji, H. and Obokata, J. (1993) Structure and expression of a nuclear gene for the PSI-D subunit of photosystem I in *Nicotiana sylvestris*. *Plant Mol. Biol.* **22**, 985–994.

Yamamoto, Y.Y., Matsui, M., Ang, L.-H. and Deng, X.-W. (1998) Role of COP1 interactive protein in mediating light-regulated gene expression in Arabidopsis. *Plant Cell*, **10**, 1083–1094.

Yamamoto, Y.Y., Shimada, Y., Kimura, M., Manabe, K., Sekine, Y., Matsui, M., Ryuto, H., Fukunishi, N., Abe, T. and Yoshida, S. (2004) Global classification of transcriptional responses to light stress in *Arabidopsis thaliana*. *Endocytobiosis Cell Res.* **15**, 438–452.

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

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

Zhu, Q., Dabi, T. and Lamb, C. (1995) TATA box and initiator functions in the accurate transcription of a plant minimal promoter in vitro. *Plant Cell*, **7**, 1681–1689.