Journal of Molecular Biology Research; Vol. 3, No. 1; 2013
ISSN 1925-430X E-ISSN 1925-4318
Published by Canadian Center of Science and Education

Characterization of Structure, Divergence and Regulation Patterns of Plant Promoters

Yingchun Liu[1*], Jiaming Yin[1*], Meili Xiao[1], Annaliese S. Mason[3], Caihua Gao[1], Honglei Liu[1], Jiana Li[1] & Donghui Fu[2]

[1] Engineering Research Center of South Upland Agriculture of Ministry of Education, College of Agronomy and Biotechnology, Southwest University, Chongqing, China

[2] Key Laboratory of Crop Physiology, Ecology and Genetic Breeding, Ministry of Education, Jiangxi Agricultural University, Nanchang, China

[3] School of Agriculture and Food Sciences and ARC Centre for Integrative Legume Research, The University of Queensland, Brisbane, Australia

[*] These authors contributed equally

Correspondence: Donghui Fu, Key Laboratory of Crop Physiology, Ecology and Genetic Breeding, Ministry of Education, Jiangxi Agricultural University, Nanchang 330045, China. Tel: 86-0791-8381-3142. E-mail: fudhui@163.com

Received: March 11, 2013 Accepted: April 18, 2013 Online Published: April 27, 2013

doi:10.5539/jmbr.v3n1p23 URL: http://dx.doi.org/10.5539/jmbr.v3n1p23

Abstract

Plant promoters have attracted increasing attention because of their irreplaceable role in modulating the spatio-temporal expression of genes interacting with transcription factors (TFs). Despite their importance, the basic characteristics of plant promoters are not well understood. In order to determine sequence diversity within promoter regions, evolutionary divergence of promoters between plant species, and the general structural characteristics of promoter sequences, we downloaded and analyzed 3922 plant promoter sequences from a wide range of plant species. The average plant promoter GC content was lower in dicotyledons than in monocotyledons, which might suggest different evolutionary pressures for promoter sequences between the two clades. Approximately 3.3% of plant promoters harbored minisatellite sequences, and 15.4% of plant promoters harbored microsatellite sequences (also called simple sequence repeats). Very few transposable elements were detected within the plant promoters. The most common transcription factor binding site (TFBS) motif was AGAGAGAGA, followed by TTAGGGTTT and then GCCGCC. Transcribed gene regions with promoters containing the corresponding TFBSs were predicted to be most commonly involved in metabolic processes, biological regulation, and stimulus response in plants. These results reveal some basic structural characteristics of plant promoters and clarify the evolutionary forces shaping plant promoters. This data might facilitate cloning of plant promoter sequences and aid in our understanding of gene spatio-temporal expression patterns in plants.

Keywords: transcription factor binding sites, minisatellite, microsatellite, transposable elements, functional annotation, GC content, evolutionary forces

1. Introduction

Promoters are sections of DNA sequence that lie upstream of the transcribed sequences and regulate their expression (Hernandez-Garcia et al., 2010). Promoters contain binding sites for transcription factors (TFs), and interact with these TFs to modulate gene expression. RNA polymerase initiates transcription at promoter sequences, and hence binding of RNA polymerase by TFs within promoter sequences regulates spatio-temporal expression of the downstream transcribed sequence (Camp et al., 2003; Halfon & Zhu, 2009; Freeman et al., 2011). Therefore, promoters are critical for priming or halting gene expression (Wolf et al., 2010; Mastroeni et al., 2011), especially in stress signaling and transcriptional activation during pathogen infection (Hwang et al., 2009; Pandey & Somssich, 2009). To date, numerous promoters have been identified in animals (Romania et al., 2011), plants (Wang et al., 2011), viruses (Smith et al., 2011b), and microorganisms (Cooper et al., 2011).

Promoters may be classified into two types according to the degree of matching between the regulatory protein and the transcription start site (TSS): Peak promoters and broad promoters. Peak promoters initiate the process of

transcription in a narrow genomic region, while broad promoters switch on transcription in a wide genomic region (Nozaki et al., 2011). Cap-analysis gene expression data can be used to indentify those two types of promoters (Carninci et al., 2006). These peak promoters generally contain TATA-boxes (except in mammals) and regulate tissue-specific transcripts in eukaryotes (Hoskins et al., 2011). For most promoters, gene transcription starts from broad regions that are usually associated with CpG islands. These broad promoters have a wide distribution of TSSs, usually over a 100-bp region, and start sites that are preferentially comprised of pyrimidine/purine dinucleotides (Carninci et al., 2006).

Promoters can be divided into prokaryotic- and eukaryotic-type promoters, which differ mainly in promoter motifs. A typical promoter sequence is thought to comprise certain motifs positioned at specific sites upstream of TSS. Two hexameric motifs centered at or near the -10 and -35 positions relative to the TSS are observed in a prokaryotic promoter, whilst a TATA box, a CCAAT box, and a GC box are usually observed in eukaryotic promoters (Bansal & Kanhere, 2005). These three types of boxes play a major role in precise initiation of transcription (Molina & Grotewold, 2005). Nevertheless, not every eukaryotic gene promoter has all three motifs (Anish et al., 2009). In addition, some novel motifs in promoter sequences, e.g. AGTTAGG (Abdullah et al., 2010), G-quadruplex (Chowdhury et al., 2010), and TATGAAAAGAATATGAGAA motifs (Wu & Huang, 2004), have been identified. Other promoter motifs, such as GATA (Obara et al., 2005) and AAAAT (Van Oers et al., 2007), are not conserved but are essential for some promoter functionality. Overall, eukaryotic promoters display more complex structures and regulation patterns than prokaryotic promoters (Bansal & Kanhere, 2005).

Promoters undergo mutations such as nucleotide substitutions, small insertions and deletions in a similar fashion to transcribed sequences (Seliverstov et al., 2009). The evolution and conservation of promoters has been scrutinized through comparative genomics studies in mammals. Previous studies include comparisons between humans and chimpanzees (Deyneko et al., 2010), and between rats, mice, rhesus monkeys, and humans for promoters of hepatic lipase genes (Van Deursen et al., 2007). GC-rich monotone gradients have been observed in eukaryotes while AT-rich monotone gradients have been observed in bacteria, along with strand biases (Calistri et al., 2011).

Each gene can have several promoters that control its spatio-temporal expression. Although promoters are important in investigating patterns of gene expression and for transgenic work, promoters are cloned far less often than transcribed gene sequences. A total of 3922 plant promoters in the Plant Promoter Database (PlantProm DB; http://linux1.softberry.com/berry.phtml) have been collected to date. Knowledge of the basic structural and evolutionary characteristics of plant promoters remain unknown, making plant promoter sequences hard to identify. To facilitate better characterization of plant promoter sequences, the 3922 available plant promoter sequences were downloaded and analyzed. Basic promoter characteristics were dissected, presence of special motifs, minisatellite sequences, microsatellite sequences, and transposable elements (TEs). We present the results of this analysis, and propose mechanisms for promoter divergence and evolution.

2. Materials and Methods

2.1 Acquisition of Plant Promoter Sequences

All plant promoter sequences from monocotyledons and dicotyledons (the latter mainly from *Arabidopsis thaliana*) were downloaded from the PlantProm DB (Release 2009.02; http://linux1.softberry.com); an annotated, non-redundant collection of proximal promoter sequences (Shahmuradov et al., 2003). These promoters could potentially be recognized by RNA polymerase II and contained experimentally determined TSSs from diverse plant species (Solovyev et al., 2003). The PlantProm DB contains both the predicted TSSs and the experimentally verified promoter TSSs, identified using approaches such as full-length cDNA/5'ESTs mapping, cap-analysis gene expression, and serial analysis of gene expression.

2.2 Detection of Microsatellite Sequences

Microsatellite sequences (also called simple sequence repeats; SSRs) are tandem repeat sequences with repeated unit lengths of 1-10 bp, present in most organisms (Morgante et al., 2002). The software SSR Locator (Da Maia et al., 2008) was used to mine SSRs with mono-, di-, tri-, tetra-, penta-, hexa-, hepta-, octa-, nova-, and decanucleotide motifs which contained a minimum of 10, 5, 4, 3, 2, 2, 2, 2, 2, and 2 repeats, respectively; only SSR sequences with a total length $\geq$ 20 bp were assigned as true SSRs (Gao et al., 2011), and subject to analysis.

2.3 Detection of Minisatellite Sequences

Minisatellites, a type of tandem repeat sequence, consist of a short series of 11-100 bp repeat units. Tandem Repeats Finder 4.04 (http://tandem.bu.edu/trf/trf.download.html) developed by Gary Benson of the Bioinformatics Program at Boston University, was used to detect minisatellite sequences (Martin, 2006). Default

parameters were used: Alignment parameters were match = 2, mismatch = 7, indel = 7, the minimum alignment score to report a repeat was 50 and the maximum period size was 100 bp.

2.4 Detection of Transposable Elements

There are two classes of transposable elements (TEs): DNA transposons and retrotransposons (Zhang et al., 2004). The Long Terminal Repeat (LTR)-Finder 1.05 (http://tlife.fudan.edu.cn/ltr_finder/) was used to detect full-length LTR retrotranspsons in genome sequences. The parameters of minimal LTR length, minimal distance between LTRs, and the output threshold score were set to 50, 100, and 3.0, respectively (Gao et al., 2012). The RepeatMasker 3.0SE-AB program (www.repeatmasker.org) was used to detect all types of transposons using the abblast (formerly known as WUBlast) search engine with *A. thaliana* set as the reference species. Since LTR-type retrotransposons detected by the LTR-Finder tool with default parameters exhibit intact retrotransposon sequence characteristics, LTR-Finder predictions were used instead of LTR retrotransposon predictions from RepeatMasker.

2.5 Prediction of Transcription Factor Binding Sites (TFBSs)

The online software NSITE-PL (http://linux1.softberry.com) with default parameters was used to predict transcription factor binding sites by recognition of regulatory motifs of plant promoters.

2.6 Functional Annotation by Blast2Go

The sequences of the transcribed gene regions with promoters containing TFBSs were in downloaded in a batch from NCBI (http://www.ncbi.nlm.nih.gov/sites/batchentrez). Blast2Go V2.6.0 (Conesa et al., 2005) (http://www.blast2go.org), a functional annotation prediction tool for unknown sequences, was used with default parameters to predict the putative functions of the transcribed gene regions with promoters containing TFBSs. Functional annotations of these genes were carried out for cellular component, biological process and molecular function.

2.7 Alignment of Plant Promoter Sequences

All plant promoters underwent all-by-all BlastN analysis using the basic local alignment search tool (BLAST) (http://www.ncbi.nlm.nih.gov/blast) (Cameron & Williams, 2007) with an E value of less than e^{-10}. The alignment results were imported into Cytoscape V2.7.0 (an open source platform for complex network analysis and visualization) (http://www.cytoscape.org) to classify different groups using the 'import network from table' function.

2.8 Phylogenic Dendrogram of Plant Promoter Sequences

The Molecular Evolutionary Genetics Analysis (MEGA; http://www.megasoftware.net) 4.0 software was used to draw the phylogenetic dendrogram of different plant promoter sequence groups using the maximum composite likelihood (MCL) model with the bootstrap value set as 1000 (Kumar et al., 2007).

3. Results

3.1 Plant Promoter Sequence Sets

A total of 3922 plant promoter sequences were downloaded from the PlantProm DB: 98 from monocotyledons and 3824 from dicotyledons. Monocotyledon sequences comprised 36 plant promoters from *Zea*, 32 from *Hordeum*, and 19 from *Triticum*. Dicotyledon sequences comprised 3537 plant promoters from *Arabidopsis*, 49 from *Nicotiana*, 46 from *Solanum*, 31 from *Glycine* and 31 from *Pisum*. Another 130 plant promoters were acquired from other genera, including *Phaseolus* (13), *Brassica* (9), and *Avena* (4).

3.2 Distribution of GC Content of Plant Promoters

GC content was calculated for each plant promoter sequence. The GC content of plant promoters ranged from 13.1% to 72.6%, with an average of 34.6%. The GC content of dicotyledon promoters ranged from 13.1% to 58.6% with an average of 34.1%, whilst the GC content of monocotyledon promoters ranged from 33.0% to 72.6% with an average of 50.5%. The centre of the GC content distribution for most dicotyledon promoters was from 30% to 40%, median 34.26%, whereas the centre of the GC content distribution of most dicotyledon promoters ranged from 50% to 60%, median 51% (Figure 1).

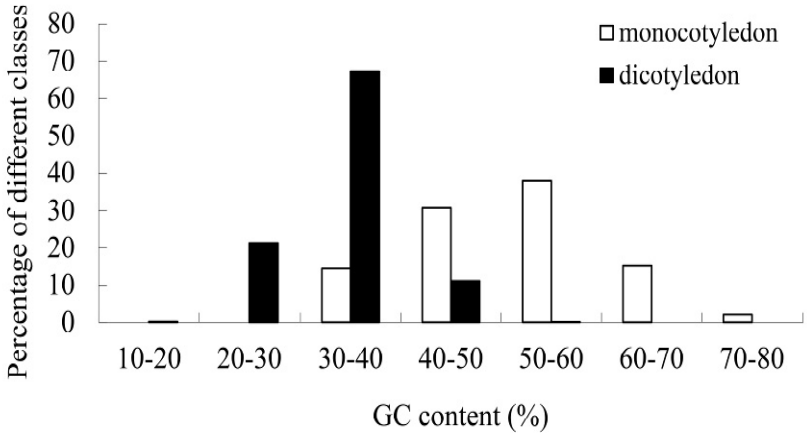

Figure 1. Proportion of plant promoter sequences with different ranges of GC content in different classes: monocotyledon and dicotyledon

3.3 Basic Characteristics of Plant Promoters

3.3.1 Detection of Microsatellites

Approximately 15% of the analyzed plant promoters (605 out of 3922) contained one or more microsatellites. Of these, 93% (563 out of 605) contained a single microsatellite, 6.5% (39 out of 605) contained two microsatellites, and 0.5% (3 out of 605) contained three microsatellites. Microsatellites with monomer motifs were by far the most common microsatellite type in the promoters (74.92%). Dimeric and trimeric microsatellite motifs were the next most common and accounted for, respectively, 15.39% and 6.14% of promoter-containing microsatellites (Table 1).

Microsatellites with monomer motifs were almost all A/T types (486 out of 487; 99.79%), with a single C monomer motif. A-motifs comprised the majority of the microsatellites with monomer repeats (345 out of 487; 70.84%) and T-motifs the minority (141 out of 487, 28.95%). AG/CT and GA/TC microsatellites comprised 71% of microsatellites with dimer motifs (Table 1).

3.3.2 Detection of Minisatellite Sequences

Approximately 2.24% of promoters (88 out of 3922) contained minisatellite sequences. No minisatellite sequences were found in monocotyledons. The length of the repeat unit ranged from 11 to 116 bp with an average of 24 bp, and the average number minisatellite repeats was 2.3, ranging from 1.9 to 3.8.

3.3.3 Analysis of TEs

No intact LTR retrotransposons were detected using LTR-finder. RepeatMasker detected 50 interspersed repeats, 6 truncated retrotransposons, and 34 DNA transposons (0.04%, 0.34%, and 0.08% of all promoters, respectively; Table 2). The most common TE types were MuDR-IS905 (0.13%), followed by hobo-Activator (0.12%), L1/CIN4 (0.02%), and Ty1/Copia (0.02%).

Table 1. Type and distribution of microsatellites in the collected plant promoter sequences

Group	Type	Type 1[a]		Type 2[b]		Subtotal	The subtotal/ The group total [%]	Overall [%]
		Number	Percentage[%]	Number	Percentage [%]			
Monomers	A/T	345	70.99	141	29.01	486	99.79	74.77
	C	1	100.00	-	-	1	0.21	0.15
Dimers	GA/TC	5	10.64	42	89.36	47	47.00	7.23
	AG/CT	8	33.33	16	66.67	24	24.00	3.69
	TA	12	100.00	-	-	12	12.00	1.85
	AT	10	100.00	-	-	10	10.00	1.54
	AC	5	100.00	-	-	5	5.00	0.77
	CA	2	100.00	-	-	2	2.00	0.31
Trimers	AGA/TCT	4	40.00	6	60.00	10	25.00	1.54
	AAG/CTT	2	22.22	7	77.78	9	22.50	1.38
	GAA/TTC	5	62.50	3	37.50	8	20.00	1.23
	AAC	3	100.00	-	-	3	7.50	0.46
	ACA	3	100.00	-	-	3	7.50	0.46
	ATC/GAT	1	50.00	1	50.00	2	5.00	0.31
	CCA	2	100.00	-	-	2	5.00	0.31
	ATT	1	100.00			1	2.50	0.15
	GTC	1	100.00			1	2.50	0.15
	TCG	1	100.00	-	-	1	2.50	0.15
Other		23				23		3.46

[a] the left hand side motif

[b] the right hand side motif (reverse complement of [a]).

The percentage of Type 1 and Type 2 motifs was derived by the number of Type 1 or Type 2 motifs divided by the subtotal.

Table 2. Predictions of presence of different types of transposable elements (TEs) in plant promoter sequences

TEs	Number of TEs	Average length of TE harbored in promoter sequence[bp]	Percentage of plant promoter sequences containing TEs in all promoters [%]
DNA transposons	34	97.2	0.34
Retroelements	6	57.8	0.04
Unclassified	10	83.3	0.08

3.3.4 Analysis of TFBSs

We used the online software NSITE-PL to predict 31259 TFBS motifs from 3922 plant promoter sequences. On average, one promoter contained eight TFBS motifs. Motif lengths ranged from 4 to 51 bp (predominantly ≤ 30 bp; 99.9%) with an average length of 11 bp.

TFBS with 10-bp motifs comprised the highest proportion of TFBS in the promoters (25.5%), followed by TFBS with 12-bp motifs (14.2%), then TFBS with 9-bp motifs (14.18%) (Figure 2). TFBS motif length mostly ranged from 6 to 17 bp (97.8% of all promoters). Up to 50% of the motifs were classified into 545 motif types, demonstrating that some key TFBS motifs are widely distributed in promoters. Most TFBS motifs possessed the characteristics of simple repeat sequences.

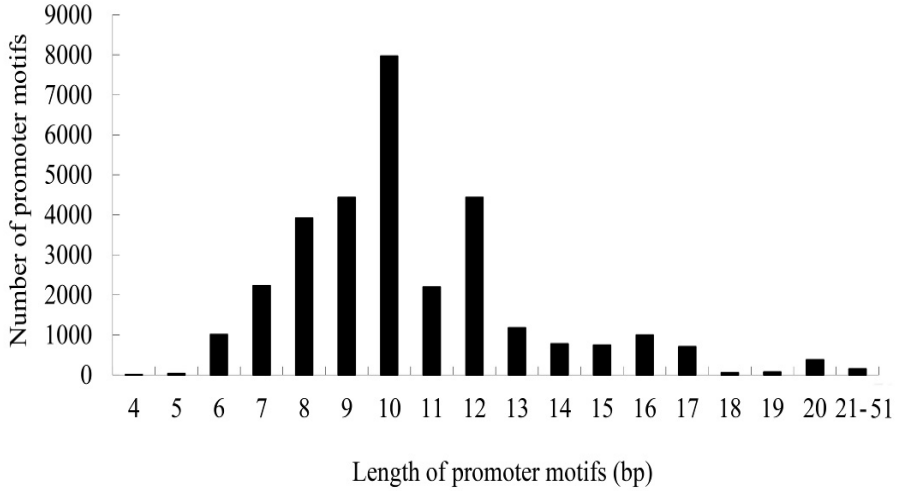

Figure 2. Distribution of motif lengths of transcription factor binding site (TFBS) in plant promoters. Promoter TFBS motifs were predicted using the software NSITE-PL to process plant promoter sequences

The TFBS motif with the highest frequency was AGAGAGAGA (1.6%; 495 out of 31259), which has previously been suggested to be a regulatory element for light responsive photo-transduction regulation in plants (Parida et al., 2009). The second most common TFBS motif was TTAGGGTTT (1.3%; 392 out of 31259); this motif has been shown to interact directly with MYB2-box-like elements in the promoters of osmotic, drought, and ABA-induced genes (Yun et al., 2010). The next most common TFBS motif was GCCGCC (1.1%; 336 out of 31259), involved in the cell cycle, jasmonic acid (JA) responsiveness and sugar signaling (Hu et al., 2011) (Figure 3). The three most common motifs comprised 4.0% of the total motif types, with the remaining motifs present at lower frequencies. The G+C content of TFBS motifs varied from 0.0% to 100.0%, with an average of 43.35%. Motifs with G+C content ranging from 0.0% to 50.0% accounted for 74.7% of all motifs, suggesting that critical promoter motifs exist in AT-rich regions.

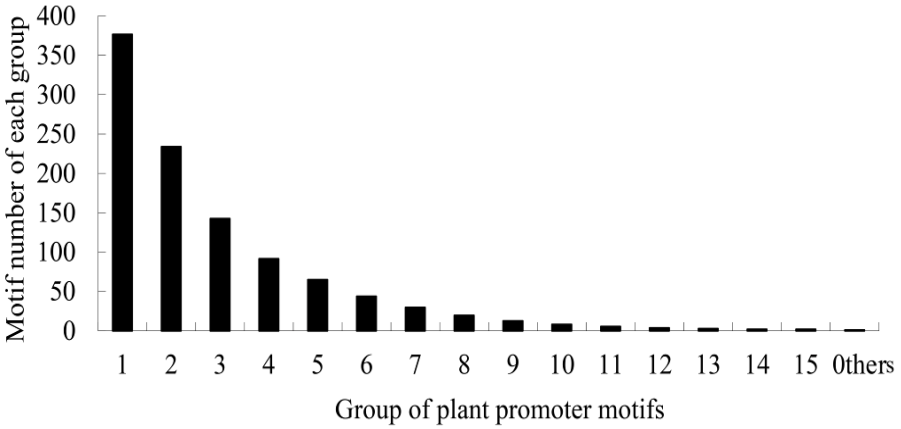

Figure 3. Mean number distribution of transcription factor binding site (TFBS) motifs in each plant promoter group. A total of 31259 motifs could be classified into 16 groups with identical motifs whole length. These groups are arranged by number of members for each motif, from greatest to least

Aside from the conserved motifs in the TFBS mentioned above, different cis-regulatory elements were also found in promoter sequences: 29201 cis-regulatory elements were identified in total. Although the frequency of most regulatory element types was low, some regulatory elements (G-box, GA-box, and ABRE motifs) were found at considerably higher frequencies (Figure 4). Among those three, G-box regulatory elements were the most common, accounting for 7.07% (2065 out of 29201) of the total regulatory elements. GA-box regulatory

elements were the second most common at 5.00% (1460 out of 29201), and ABRE regulatory elements were the third most common at 4.81% (1405 out of 29201). Our results show that a small number of motifs with high affinities for binding proteins are widely distributed in promoter sequences.

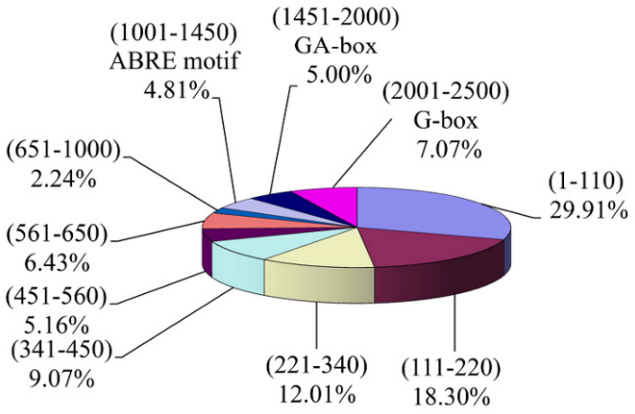

Figure 4. Number distribution of plant promoter regulatory elements detected in promoters. The values in the brackets represent the number range of all kinds of regulatory elements detected, and the percentages denote the total percentage of promoters in each regulatory element group

3.3.5 Putative Functional Annotation of the Transcribed Gene Regions with Promoters Containing the Corresponding TFBSs

NSITE (Version 2.2004; Softberry Inc.) was used to recognize TFBSs and provide information for the transcribed gene regions with promoters containing the corresponding TFBSs. Blast2Go was used to predict the functional annotation of the transcribed gene regions with promoters containing the corresponding TFBSs for biological process, molecular function and cellular component. For biological process annotation, the most common involvement was in metabolic processes (27.81%), followed by biological regulation (27.54%), and response to stimulus (17.77%) (Figure 5a). With respect to molecular functionality, the transcribed gene regions with the promoters containing the corresponding TFBSs mainly played a role in binding function (45.57%), followed by catalytic activity (23.86%) and other unknown molecular functions (17.57%) (Figure 5b). The transcribed gene regions with promoters containing the corresponding TFBSs most commonly functioned in the organelles (42.75%), followed by the intracellular (22.46%), and cellular components (20.53%) (Figure 5c). Hence, for biological process annotation, the transcribed gene regions with promoters containing the corresponding TFBSs were mainly involved in metabolic processes; with respect to molecular functionality, the most common function was binding; and with regard to cellular component annotation, transcribed gene regions most commonly functioned in the organelles.

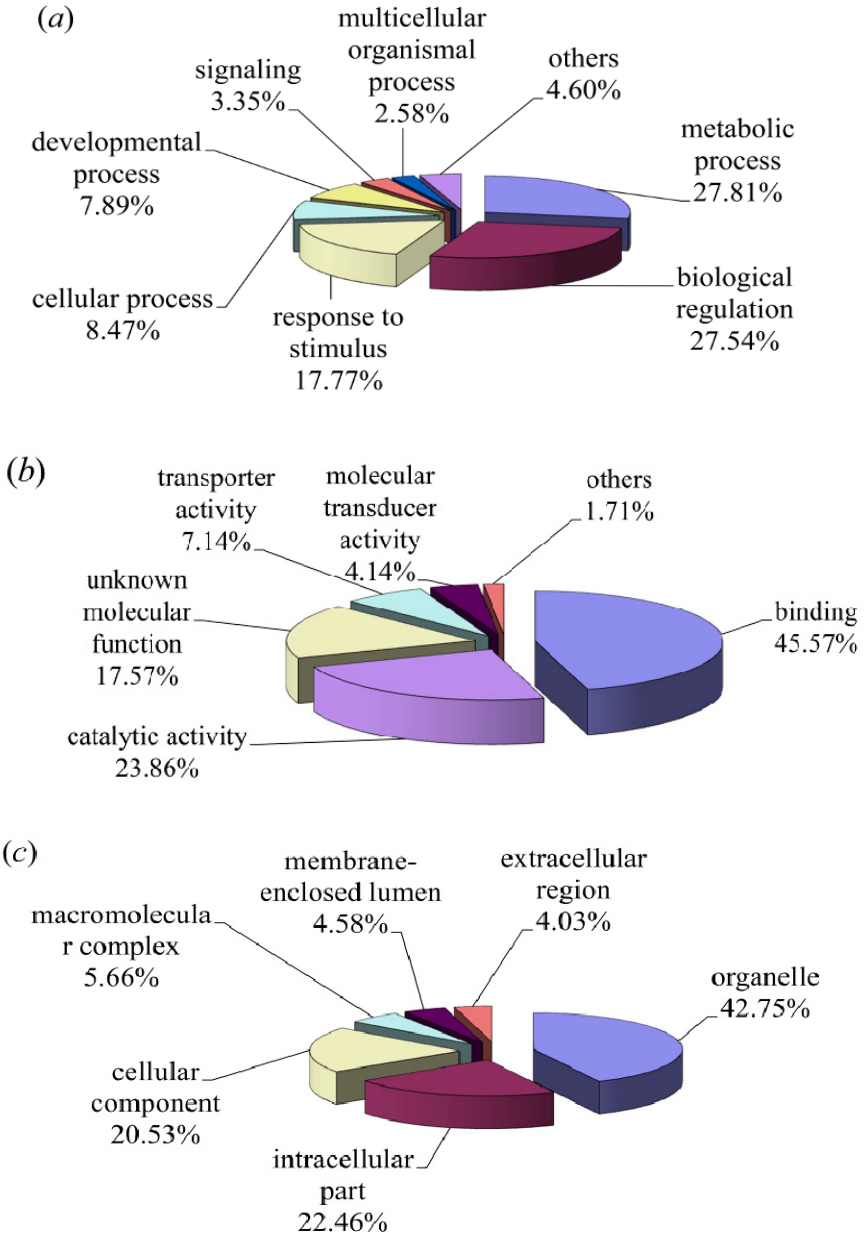

Figure 5. Functional annotation of the transcribed gene regions with promoters containing the corresponding transcription factor binding sites (TFBSs)

(a) Biological process; (b) Molecular function; (c) Cellular components.

3.3.6 Analysis of Alignment and Phylogenetic Dendrogram of Plant Promoter Sequences

All-by-all BlastN analysis of the plant promoters did not allow clear classification into different subclasses (Figure 6), indicating that the homology of these plant promoter sequences was relatively low. Nevertheless, according to the structure of the phylogenetic dendrogram, the ancestral lineages produced in MEGA 4 (Figure 7) and the species taxonomy, the plant promoter sequences could be classified into 8 groups containing 1172, 791, 60, 24, 136, 59, 287, and 1393 sequences, respectively (Figure 8). The genetic distance between the 8 groups was 0.19 on average, indicating greater divergence within the plant promoter sequence groups than between groups.

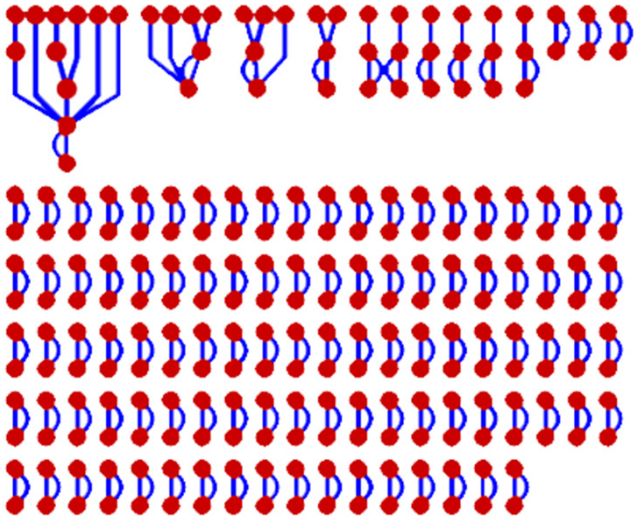

Figure 6. Classification of different plant promoter sequences

All-by-all BlastN analysis was used to classify different plant promoter sequences into different subclasses. The circles represent different plant promoter sequences, and the lines between the circles denote the homology between the two plant promoter sequences distributed in the two circles.

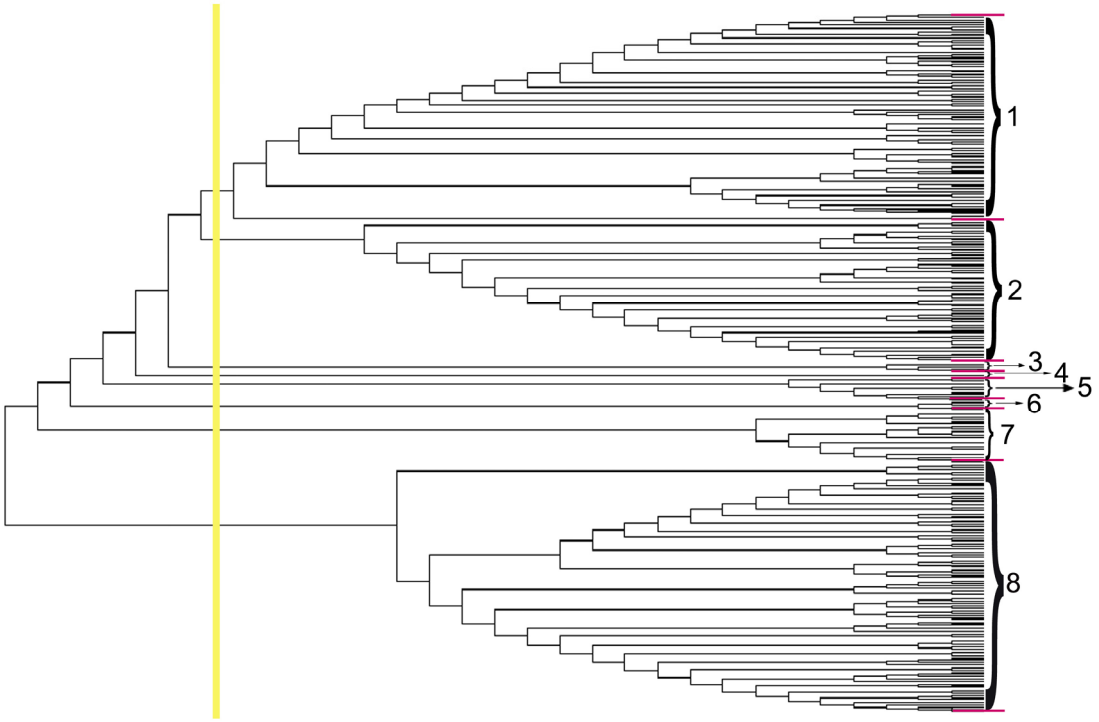

Figure 7. Phylogenetic dendrogram of the plant promoter sequences of 288 species

All plant promoter sequences were classified into 8 classes. The yellow line represents the demarcation of different classes. The numbers on the right represent the plant promoter sequence groups, with groups marked out using two pink lines.

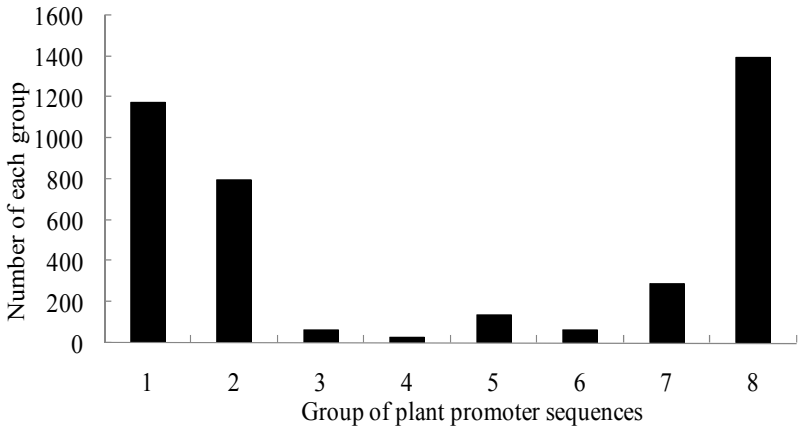

Figure 8. Distribution of the number of plant promoter sequences in each group

We divided the whole plant promoter sequences into eight groups according to the standard of uniform ancestral lineage in their phylogenetic dendrogram, and then the sequence number per group was counted and labeled in Y-axis.

4. Discussion

4.1 GC Content and Mutability of Plant Promoter Sequences

In the current study, the GC content of plant promoters was between 30% and 40% in most dicotyledon species, but was between 50% and 60% in most monocotyledon species, indicating that the GC content of plant promoters in monocotyledons is generally higher than that in dicotyledons. AT-rich regions are prone to mutate to generate diversity more often than GC-rich regions, and are inserted by exogenous gene fragments such as transposons (Gupta et al., 2005). Hence, more complex gene regulation may be required in dicotyledons compared to monocotyledons. AT-rich microsatellite sequences were also very common in the plant promoter sequences, suggesting that the mutability of plant promoters may have an important evolutionary adaptive role in diversification of gene expression. Nevertheless, some transcribed gene regions with GC-rich promoters are expressed more efficiently (Singh et al., 2012) suggesting that balancing selective pressure may exist for retention of GC-rich promoter sequences for genome stability.

4.2 Frequency and Possible Functionality of Key Promoter Motifs

According to the results, the length of most plant promoter TFBSs ranged from 6 to 17 bp, with AGAGAGAGA (1.6%; 495 out of 31259), TTAGGGTTT (1.3%; 392 out of 31259), and GCCGCC (1.1%; 336 out of 31259), being the most common. These high-frequency motifs may represent cis-regulatory elements which enhance the expression of sets of related genes. These common motifs detected may exist in the promoters of genes which have been highly conserved in species evolution, such as genes that play basic roles in plant growth and development. For example, the motif AGAGAGAGA is a known regulatory element participating in light-responsive regulation of phototransduction in plants (Parida et al., 2009). This motif is also present in the promoter of the *WRKY* gene which encodes the WRKY protein, one of the largest families of TFs, regulating processes such as response to biotic and abiotic stresses in plants (Zhang & Wang, 2005; Rushton et al., 2010). In rice, the WRKY gene family contains over 100 members (Pandey & Somssich, 2009). Likewise, the second most common TFBS motif (TTAGGGTTT) can directly interact with MYB2-box-like elements in the promoters of osmotic, drought, and ABA-induced genes (Yun et al., 2010). In contrast, different organisms may also have organism-specific but genome-wide TFBS motifs. For example, in Actinobacteria, the most significant TFBS motif is TCGAACA (Janky & van Helden, 2008). Similarly, the octamer AAAATTGA motif exists in the predicted core promoters of almost half the Mimivirus genes (Suhre et al., 2005). Therefore, high-frequency TFBS motifs may play multiple and comprehensive roles in many processes occurring in different organisms.

In addition, plant promoter motifs play important roles in accurate initiation of transcription. TFs can combine with DNA to orchestrate transcription of specific cis-regulatory elements (Rombauts et al., 2003). Only small numbers of TFs also combine with special promoter motifs to regulate expression of large numbers of genes (Smith et al., 2011a). Identification of such broad promoters may be useful for transgenic breeding, because the combination between these critical motifs and just a few TFs may allow for more effectively controlled

expression of a batch of downstream transcribed gene regions. Critical promoter motifs with important roles can also be used to construct regulatory sequences which contribute to the spatio-temporal expression of transgenic plants. Thus, recombined regulatory sequences could not only accelerate the speed of breeding but also help in obtaining special gene products.

4.3 Functional Annotation of the Transcribed Gene Regions With Promoters Containing TFBSs

In this study, 31259 motifs of TFBS were detected from 3922 plant promoter sequences. On average, one promoter contained eight TFBS motifs. What are functions of these transcribed gene regions with promoters containing TFBSs? Blast2GO annotation revealed that the transcribed gene regions with TFBS-containing promoters commonly controlled metabolic processes during plant development, mainly had molecular binding functionality, and were operative in the organelles. We may characterize and mine critical TFBS and promoters from these transcribed gene regions to serve breeding purposes. Promoter cloning and subsequent manipulation of spatio-temporal gene expression offers significant promise as a developing research field in transgenic breeding. Promoter-based transgenic technologies have already been applied to great effect in wheat, where a heat-inducible promoter in transgenic wheat effectively controlled the spatio-temporal expression of a transgene (Freeman et al., 2011).

4.4 Some Microsatellites are Universally Distributed in Plant Promoters

Different species share common, prevalent motifs in promoters. The current study observed that $(A)_n$, $(T)_n$, $(AG)_n$, $(GA)_n$, $(CT)_n$, and $(TC)_n$ were the predominant mononucleotide and dinucleotide microsatellite motifs, respectively. This result suggests that microsatellites with specific motifs survived during natural selection due to positive selective advantages. The monomer microsatellites (almost all A and T motifs) accounted for the highest proportion of the microsatellite-containing promoter sequences. As the A/T-motif microsatellites are easily mutated (Gao et al., 2011), this may indicate a positive selection pressure due to the advantage provided by the extra diversity of gene expression in adapting to the environment and evolving into more complex higher organisms.

In summary, the GC content of plant promoters in monocotyledons appeared to be higher than that in dicotyledons. Most microsatellites and TEs were quite rare in promoter sequences, whereas microsatellites with A and T monomers were very commonly observed and may provide adaptive mutability potential in plant promoter sequences. Motifs of particular lengths occurred mainly on the TFBSs, and regulatory elements occurring with high frequency were mostly G-box, GA-box, and ABRE motifs. For biological process annotation, the transcribed gene regions with promoters containing the corresponding TFBSs were mainly involved in metabolic processes; with respect to molecular functionality, the most common function was binding; and with regards to cellular component annotation, the most common functional location was the organelles

The characteristics of higher A/T content, more microsatellites and a small quantity of TEs in plant promoters may play a role in evolution of plant promoters. The different TFBS motifs in plant promoters are a critical element of spatio-temporal expression of genes. These results are beneficial not only for elucidating the mechanisms of spatio-temporal gene expression and for cloning key plant promoters (or their main motifs), but also for investigating the basic structure of plant promoters and clarifying the evolutionary forces at work in plant promoter diversification.

Acknowledgments

This work was supported financially by National Natural Science Foundation of China (code: 31260335), and Research Fund for the Doctoral Program of Higher Education of China (code: 20123603120002). ASM is supported by an Australian Research Council Discovery Early Career Researcher Award (DE120100668).

References

Abdullah, S. N. A., Omidvar, V., Izadfard, A., Ho, C. L., & Mahmood, M. (2010). The oil palm metallothionein promoter contains a novel AGTTAGG motif conferring its fruit-specific expression and is inducible by abiotic factors. *Planta, 232*(4), 925-936. http://dx.doi.org/10.1007/s00425-010-1220-z

Anish, R., Hossain, M. B., Jacobson, R. H., & Takada, S. (2009). Characterization of transcription from TATA-less promoters: Identification of a new core promoter element XCPE2 and analysis of factor requirements. *PLOS One, 4*(4), e5103. http://dx.doi.org/10.1371/journal.pone.0005103

Bansal, M., & Kanhere, A. (2005). Structural properties of promoters: similarities and differences between prokaryotes and eukaryotes. *Nucleic Acids Research, 33*(10), 3165-3175. http://dx.doi.org/10.1093/nar/gki627

Calistri, E., Livi, R., & Buiatti, M. (2011). Evolutionary trends of GC/AT distribution patterns in promoters. *Molecular Phylogenetics and Evolution, 60*(2), 228-235. http://dx.doi.org/10.1016/j.ympev.2011.04.015

Cameron, M., & Williams, H. E. (2007). Comparing compressed sequences for faster nucleotide BLAST searches. *LEEE-ACM Transactions on Computational Biology and Bioinformatics, 4*(3), 349-364. http://dx.doi.org/10.1109/TCBB.2007.1029

Camp, E., Badhwar, P., Mann, G. J., & Lardelli, M. (2003). Expression analysis of a tyrosinase promoter sequence in zebrafish. *Pigment Cell Research, 16*(2), 117-126. http://dx.doi.org/10.1034/j.1600-0749.2003.00002.x

Carninci, P. (2006). Tagging mammalian transcription complexity. *Trends in Genetics, 22*(9), 501-510. http://dx.doi.org/10.1016/j.tig.2006.07.003

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., … Hayashizaki, Y. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genetics, 38*(6), 626-635. http://dx.doi.org/10.1038/ng1789

Chowdhury, S., Basundra, R., Kumar, A., Amrane, S., Verma, A., & Phan, A. T. (2010). A novel G-quadruplex motif modulates promoter activity of human thymidine kinase 1. *FEBS Journal, 277*(20), 4254-4264. http://dx.doi.org/10.1111/j.1742-4658.2010.07814.x

Conesa, A., Gotz, S., Garcia-Gomez, J. M., Terol, J., Talon, M., & Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics, 21*(18), 3674-3676. http://dx.doi.org/10.1093/bioinformatics/bti610

Cooper, T. G., Georis, I., Tate, J. J., Feller, A., & Dubois, E. (2011). Intranuclear function forprotein phosphatase 2A: Pph21 and Pph22 are required for rapamycin-induced GATA factor binding to the DAL5 promoter in yeast. *Molecular and Cellular Biology, 31*(1), 92-104. http://dx.doi.org/10.1128/MCB.00482-10

Da Maia, L. C., Palmieri, D. A., de Souza, V. Q., Kopp, M. M., de Carvalho, F. I., & Costa de Oliveira, A. (2008). SSR Locator: Tool for simple sequence repeat discovery integrated with primer design and PCR simulation. *International Journal Plant Genomics, 2008*(2008), 412-426. http://dx.doi.org/10.1155/2008/412696

Deyneko, I. V., Kalybaeva, Y. M., Kel, A. E., & Blocker, H. (2010). Human-chimpanzee promoter comparisons: Property-conserved evolution? *Genomics, 96*(3), 129-133. http://dx.doi.org/10.1016/j.ygeno.2010.06.003

Freeman, J., Sparks, C. A., West, J., Shewry, P. R., & Jones, H. D. (2011). Temporal and spatial control of transgene expression using a heat-inducible promoter in transgenic wheat. *Plant Biotechnology Journal, 9*(7), 788-796. http://dx.doi.org/10.1111/j.1467-7652.2011.00588.x

Gao, C. H., Xiao, M. L., Jiang, L. Y., Li, J. N., Yin, J. M., Ren, X. D., … Tang, Z. L. (2012). Characterization of transcriptional activation and inserted-into-gene preference of various transposable elements in the Brassica species. *Molecular Biology Reports, 39*(7), 7513-7523. http://dx.doi.org/10.1007/s11033-012-1585-0

Gao, C., Tang, Z., Yin, J., An, Z., Fu, D., & Li, J. (2011). Characterization and comparison of gene-based simple sequence repeats across Brassica species. *Molecular Genetics and Genomics, 286*(2), 161-170. http://dx.doi.org/10.1007/s00438-011-0636-x

Gupta, S., Gallavotti, A., Stryker, G. A., Schmidt, R. J., & Lal, S. K. (2005). A novel class of Helitron-related transposable elements in maize contain portions of multiple pseudogenes. *Plant Molecular Biology, 57*(1), 115-127. http://dx.doi.org/10.1007/s11103-004-6636-z

Halfon, M. S., & Zhu, Q. Q. (2009). Complex organizational structure of the genome revealed by genome-wide analysis of single and alternative promoters in Drosophila melanogaster. *BMC Genomics, 10*(9), 216-228. http://dx.doi.org/10.1186/1471-2164-10-9

Hernandez-Garcia, C. M., Bouchard, R. A., Rushton, P. J., Jones, M. L., Chen, X., Timko, M. P., & Finer, J. J. (2010). High level transgenic expression of soybean (Glycine max) GmERF and Gmubi gene promoters isolated by a novel promoter analysis pipeline. *BMC Plant Biology, 10*(10), 237-252. http://dx.doi.org/10.1186/1471-2229-10-237

Hoskins, R. A., Landolin, J. M., Brown, J. B., Sandler, J. E., Takahashi, H., Lassmann, T., … Celniker, S. E. (2011). Genome-wide analysis of promoter architecture in drosophila melanogaster. *Genome Research, 21*(2), 182-192. http://dx.doi.org/10.1101/gr.112466.110

Hu, F., Wang, D., Zhao, X., Zhang, T., Sun, H., Zhu, L., … Li, Z. (2011). Identification of rhizome-specific genes by genome-wide differential expression analysis in Oryza longistaminata. *BMC Plant Biology, 11*(18),

86-101. http://dx.doi.org/10.1186/1471-2229-11-18

Hwang, B. K., An, S. H., Choi, H. W., & Hong, J. K. (2009). Regulation and function of the pepper pectin methylesterase inhibitor (CaPMEI1) gene promoter in defense and ethylene and methyl jasmonate signaling in plants. *Planta, 230*(6), 1223-1237. http://dx.doi.org/10.1007/s00425-009-1021-4

Janky, R., & van Helden, J. (2008). Evaluation of phylogenetic footprint discovery for predicting bacterial cis-regulatory elements and revealing their evolution. *BMC Bioinformatics, 9*(37), 326-338. http://dx.doi.org/10.1186/1471-2105-9-37

Kumar, S., Tamura, K., Dudley, J., & Nei M. (2007). MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. *Molecular Biology and Evolution, 24*(8), 1596-1599. http://dx.doi.org/10.1093/molbev/msm092

Martin, D. E. K. (2006). The exact joint distribution of the sum of heads and apparent size statistics of a "tandem repeats finder" algorithm. *Bulletin of Mathematical Biology, 68*(8), 2353-2364. http://dx.doi.org/10.1007/s11538-006-9146-0

Mastroeni, P., Janis, C., Grant, A. J., McKinley, T. J., Morgan, F. J. E., John, V. F., … Dougan, G. (2011). In vivo regulation of the vi antigen in salmonella and induction of immune responses with an in vivo-inducible promoter. *Infection and Immunity, 79*(6), 2481-2488. http://dx.doi.org/10.1128/IAI.01265-10

Molina, C., & Grotewold, E. (2005). Genome wide analysis of Arabidopsis core promoters. *BMC Genomics, 6*(25), 147-159. http://dx.doi.org/10.1186/1471-2164-6-25

Morgante, M., Hanafey, M., & Powell, W. (2002). Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nature Genetics, 30*(2), 194-200. http://dx.doi.org/10.1038/ng822

Nozaki, T., Yachie, N., Ogawa, R., Kratz, A., Saito, R., & Tomita, M. (2011). Tight associations between transcription promoter type and epigenetic variation in histone positioning and modification. *BMC Genomics, 12*(1), 416-429. http://dx.doi.org/10.1186/1471-2164-12-416

Obara, N., Suzuki, N., Ki-Bom, K., Imagawa, S., Nagasawa, T., & Yamamoto, M. (2005). GATA motif on the erythropoietin gene promoter is essential for repression of ectopic constitutive erythropoietin production. *Blood, 106*(11), 878A-878A.

Pandey, S. P., & Somssich, I. E. (2009). The role of WRKY transcription factors in plant immunity. *Plant Physiology, 150*(4), 1648-1655. http://dx.doi.org/10.1104/pp.109.138990

Parida, S. K., Dalal, V., Singh, A. K., Singh, N. K., & Mohapatra, T. (2009). Genic non-coding microsatellites in the rice genome: characterization, marker design and use in assessing genetic and evolutionary relationships among domesticated groups. *BMC Genomics, 10*(6), 140-152. http://dx.doi.org/10.1186/1471-2164-10-140

Romania, M., Brigati, C., Banelli, B., Casciano, I., Di Vinci, A., Matis, S., … Allemanni, G. (2011). Epigenetic mechanisms regulate Delta NP73 promoter function in human tonsil B cells. *Molecular Immunology, 48*(4), 408-414. http://dx.doi.org/10.1016/j.molimm.2010.09.001

Rombauts, S., Florquin, K., Lescot, M., Marchal, K., Rouze, P., & van de Peer, Y. (2003). Computational approaches to identify promoters and cis-regulatory elements in plant genomes. *Plant Physiology, 132*(3), 1162-1176. http://dx.doi.org/10.1104/pp.102.017715

Rushton, P. J., Somssich, I. E., Ringler, P., & Shen, Q. X. J. (2010). WRKY transcription factors. *Trends in Plant Science, 15*(5), 247-258. http://dx.doi.org/10.1016/j.tplants.2010.02.006

Seliverstov, A. V., Lysenko, E. A., & Lyubetsky, V. A. (2009). Rapid evolution of promoters for the plastome gene ndhF in flowering plants. *Russian Journal of Plant Physiology, 56*(6), 838-845. http://dx.doi.org/10.1134/S1021443709060144

Shahmuradov, I. A., Gammerman, A. J., Hancock, J. M., Bramley, P. M., & Solovyev, V. V. (2003). PlantProm: a database of plant promoter sequences. *Nucleic Acids Research, 31*(1), 114-117. http://dx.doi.org/10.1093/nar/gkg112

Singh, D. P., Bhargavan, B., Chhunchha, B., Kubo, E., Kumar, A., & Fatma, N. (2012). Transcriptional protein Sp1 regulates LEDGF transcription by directly interacting with its cis-elements in GC-rich region of TATA-less gene promoter. *PLOS One, 7*(5), 3701-3712. http://dx.doi.org/10.1371/journal.pone.0037012

Smith, A. J., Chudnovsky, L., Simoes-Barbosa, A., Delgadillo-Correa, M. G., Jonsson, Z. O., Wohlschlegel, J. A., & Johnson, P. J. (2011a). Novel core promoter elements and a cognate transcription factor in the divergent

unicellular eukaryote Trichomonas vaginalis. *Molecular and Cellular Biology, 31*(7), 1444-1458. http://dx.doi.org/10.1128/MCB.00745-10

Smith, C. L., Lee, S. C., & Magklara, A. (2011b). HDAC activity is required for efficient core promoter function at the mouse mammary tumor virus promoter. *Journal of Biomedicine and Biotechnology, 2011*(2011), 169-185. http://dx.doi.org/10.1155/2011/416905

Solovyev, V. V., Shahmuradov, I. A., Gammerman, A. J., Hancock, J. M., & Bramley, P. M. (2003). PlantProm: a database of plant promoter sequences. *Nucleic Acids Research, 31*(1), 114-117. http://dx.doi.org/10.1093/nar/gkg041

Suhre, K., Audic, S., & Claverie, J. M. (2005). Mimivirus gene promoters exhibit an unprecedented conservation among all eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America, 102*(41), 14689-14693. http://dx.doi.org/10.1073/pnas.0506465102

Van Deursen, D., Botma, G. J., Jansen, H., & Verhoeven, A. J. (2007). Comparative genomics and experimental promoter analysis reveal functional liver-specific elements in mammalian hepatic lipase genes. *BMC Genomics, 8*(8), 99-112. http://dx.doi.org/10.1186/1471-2164-8-99

Van Oers, M. M., Nalcacioglu, R., Ince, I. A., Vlak, J. M., & Demirbag, Z. (2007). The Chilo iridescent virus DNA polymerase promoter contains an essential AAAAT motif. *Journal of General Virology, 2007*(88), 2488-2494. http://dx.doi.org/10.1099/vir.0.82947-0

Wang, Y. J., Xu, W. R., Yu, Y. H., Zhou, Q., Ding, J. H., Dai, L. M., … Zhang, C. H. (2011). Expression pattern, genomic structure, and promoter analysis of the gene encoding stilbene synthase from Chinese wild vitis pseudoreticulata. *Journal of Experimental Botany, 62*(8), 2745-2761. http://dx.doi.org/10.1093/jxb/erq447

Wolf, E., Aigner, B., & Klymiuk, N. (2010). Transgenic pigs for xenotransplantation: selection of promoter sequences for reliable transgene expression. *Current Opinion in Organ Transplantation, 15*(2), 201-206. http://dx.doi.org/10.1097/MOT.0b013e328336ba4a

Wu, Q. Y., & Huang, W. (2004). The ManR specifically binds to the promoter of a Nramp transporter gene in Anabaena sp PCC 7120: a novel regulatory DNA motif in cyanobacteria. *Biochemical and Biophysical Research Communications, 317*(2), 578-585. http://dx.doi.org/10.1016/j.bbrc.2004.03.089

Yun, K. Y., Park, M. R., Mohanty, B., Herath, V., Xu, F., Mauleon, R., … de Los Reyes, B. G. (2010). Transcriptional regulatory network triggered by oxidative signals configures the early response mechanisms of japonica rice to chilling stress. *BMC Plant Biology, 10*(16), 163-182. http://dx.doi.org/10.1186/1471-2229-10-16

Zhang, P., Li, W. L., Fellers, J., Friebe, B., & Gill, B. S. (2004). BAC-FISH in wheat identifies chromosome landmarks consisting of different types of transposable elements. *Chromosoma, 112*(6), 288-299. http://dx.doi.org/10.1007/s00412-004-0273-9

Zhang, Y. J., & Wang, L. J. (2005). The WRKY transcription factor superfamily: its origin in eukaryotes and expansion in plants. *BMC Evolutionary Biology, 5*(1), 1236-1248. http://dx.doi.org/10.1186/1471-2148-5-1