

## An *Arabidopsis* promoter microarray and its initial usage in the identification of HY5 binding targets *in vitro*

Ying Gao<sup>1,2</sup>, Jinming Li<sup>3</sup>, Elizabeth Strickland<sup>2</sup>, Sujun Hua<sup>4</sup>, Hongyu Zhao<sup>5</sup>, Zhangliang Chen<sup>1</sup>, Lijia Qu<sup>1</sup> and Xing Wang Deng<sup>1,2,\*</sup>

<sup>1</sup>*Peking-Yale Joint Center of Plant Molecular Genetics and Agrobiotechnology, College of Life Sciences, Peking University, Beijing 100871, PR China;* <sup>2</sup>*Department of Molecular, Cellular, and Developmental Biology, Yale University, New Haven, CT 06520-8104, USA (\*author for correspondence; e-mail xingwang.deng@yale.edu);* <sup>3</sup>*School of Biological Sciences, Nanyang Technological University, Singapore 637616, Singapore;* <sup>4</sup>*Bioinformatics and Computational Biology Track, Biological and Biomedical Sciences Program, Yale University, New Haven, CT 06511, USA;* <sup>5</sup>*Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT 06520, USA*

Received 19 December 2003; accepted in revised form 5 April 2004

**Key words:** *Arabidopsis*, HY5, promoter microarray, transcription factor–promoter interaction

### Abstract

To analyze transcription factor–promoter interactions in *Arabidopsis*, a general strategy for generating a promoter microarray has been established. This includes an integrated platform for promoter sequence extraction and the design of primers for the PCR amplification of the promoter regions of annotated genes in the *Arabidopsis* genome. A web-interfaced primer-retrieval program was used to obtain up to 10 primer pairs with a suitability ranking given to each gene. We selected primer pairs for the promoters of about 3800 genes, and greater than 95% of the promoter fragments from the total genomic DNA were successfully amplified by PCR. These PCR products were purified and used to print an *Arabidopsis* promoter microarray. This initial promoter microarray was used to study the *in vitro* binding of the transcription factor HY5 to its promoter targets. A set of promoter fragments exhibited consistent and strong interaction with the HY5 protein *in vitro*, and computational analysis revealed that they were enriched with the HY5 consensus binding G-box motif. Thus, a promoter microarray can be a useful tool for identifying transcription factor binding sites at the genomic scale in higher plants.

### Introduction

Transcription factor–promoter interactions are fundamentally important for understanding the regulation of genome expression, and, thus, eukaryotic cell growth and development. A series of recent papers revealed critical insights in the genome-wide transcription regulatory network using a global genome-wide analysis of transcription factor binding sites in several model organisms, including yeast (Ren *et al.*, 2000; Iyer *et al.*, 2001; Simon *et al.*, 2001; Wyrick *et al.*, 2001), *Drosophila* (Markstein *et al.*, 2002; Stathopoulos

*et al.*, 2002; Orian *et al.*, 2003), and mammalian cells (Horak *et al.*, 2002; Ren *et al.*, 2002; Weinmann *et al.*, 2002). Although a combination of gene expression analysis and computational prediction strategy has been employed previously to understand genome expression regulation in *Arabidopsis* (Hong *et al.*, 2003; Ramirez-Parra *et al.*, 2003), the analysis of transcription factor–promoter interactions has been largely limited to individual genes (Saha *et al.*, 2001; Egelkrout *et al.*, 2002; Lopez-Molina *et al.*, 2002).

The *Arabidopsis thaliana* genome encodes at least fifteen-hundred transcription factors, which

account for about 6% of the estimated total number of genes in the genome (Riechmann *et al.*, 2000). Compared with yeast, *Drosophila*, and *C. elegans*, for which considerable fractions (85%, >25% and >25%) of their known transcription factors have been characterized (Ruvkun and Hobert, 1998; Costanzo *et al.*, 2000), estimates are that only about 5% (<100) of the transcription factors from *Arabidopsis* have been functionally characterized (Riechmann *et al.*, 2000). Hence, it has become evident that effective and high throughput tools will be required to unravel the function of *Arabidopsis* transcription factors in a systematic manner.

The *Arabidopsis* transcription factor HY5 is one of the best-characterized transcription factors (Oyama *et al.*, 1997). It is a basic leucine-zipper type transcription factor and a positive regulator of photomorphogenesis (Chattopadhyay *et al.*, 1998). Genetic analyses have suggested that HY5 acts downstream of multiple photoreceptor-mediated pathways and functionally interacts with pleiotropic constitutive photomorphogenic/deetiolated/fusca (COP/DET/FUS) genes, which are negative regulators of photomorphogenesis (Ang and Deng, 1994; Wei *et al.*, 1994; Pepper and Chory, 1997; Ang *et al.*, 1998). HY5 is constitutively nuclear localized and directly involved in light regulation of transcriptional activity of the light-responsive genes. A previous study (Chattopadhyay *et al.*, 1998) suggested that HY5 specifically, and directly, bound *in vitro* to the G-box, a well-characterized light-responsive element (LRE) that is commonly found in light-regulated promoters. This binding was proposed to be essential for HY5-mediated light control of promoter activity. Nevertheless, only a couple of G-box containing promoters have been investigated so far (Ang *et al.*, 1998; Chattopadhyay *et al.*, 1998). Therefore, further research is needed to investigate the spectrum of HY5 binding sites at a genome scale.

Commonly used transcription factor–promoter interaction analysis techniques, such as filter membrane-binding assays (Woodbury and Von Hippel, 1983), gel shift assays (Garner and Revzin, 1981), ELISA (Choo and Klug 1993), Southwestern blotting (Bowen *et al.*, 1980), or reporter gene assays (Hanes and Brent, 1991), are generally too laborious to be used directly for genome scale protein–DNA interactions (Bulyk *et al.*, 2001).

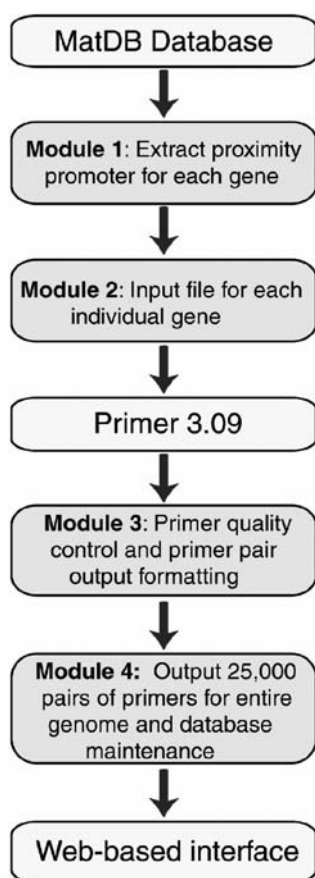
Although *in vitro* binding site selections (Oliphant *et al.*, 1989) and ‘binding-site signatures’ (Choo and Klug, 1994) permitted the sampling of multiple DNA-binding sites for a particular transcription factor, they only provided a partial view of the binding-site specificity because only the highest affinity binding sites were selected. Information from the less than optimal binding sites might be lost in these assays. On the other hand, it is very possible that some lower-affinity DNA binding sites are functionally significant in the transcriptional regulation of gene expression (Bulyk *et al.*, 2001). Therefore, DNA microarray technologies have been used to develop an efficient method for transcription factor–DNA interactions on a large scale.

For global transcription factor binding studies employed in several organisms, an essential tool is the intergenic or promoter microarray. The completely sequenced genome and the well-annotated databases of *Arabidopsis* now provide us with an unique opportunity to develop a genome-wide promoter extraction and a promoter primer design platform that can direct a PCR amplification of the promoter regions of all annotated genes in the *Arabidopsis* genome. Using this platform, we have constructed a 3.8 K *Arabidopsis* promoter microarray, and used it to study *in vitro* HY5–promoter interactions. Our study demonstrates the feasibility of identifying transcription factor binding sites using a promoter microarray.

## Result and discussion

### *Development of an integrated platform for promoter sequence extraction and primer design for all annotated genes in the Arabidopsis genome*

An integrated platform was set up to retrieve the hypothetical promoter region for each individual gene in the *Arabidopsis* genome reported by the MAtDB database (<http://www.mips.biochem.mpg.de/proj/thal/db/>) and to select specific primers for those promoter regions (Figure 1). The integrated platform automated the procedures of sorting genes in terms of their position along the chromosome; locating start and stop codon positions; determining intergenic regions; extracting promoter template sequences from chromosomes; preparing numerous input parameters for Primer3;



*Figure 1.* Flow chart of the promoter primer design and its database. An integrated platform for promoter sequence extraction and the design of primers for PCR amplification of promoter regions of all annotated genes in the *Arabidopsis* genome has been developed. This platform consists of (1) database search and template sequence extraction; (2) preparation of the Primer3 input and picking the primer pairs with Primer3; (3) primer quality control and primer pair output formatting; (4) Construction of the primer pair database and the primer retrieval for a given list of genes. In addition, a web-interfaced primer retrieval program was used to obtain up to 10 primer pairs with a proper ranking for each gene.

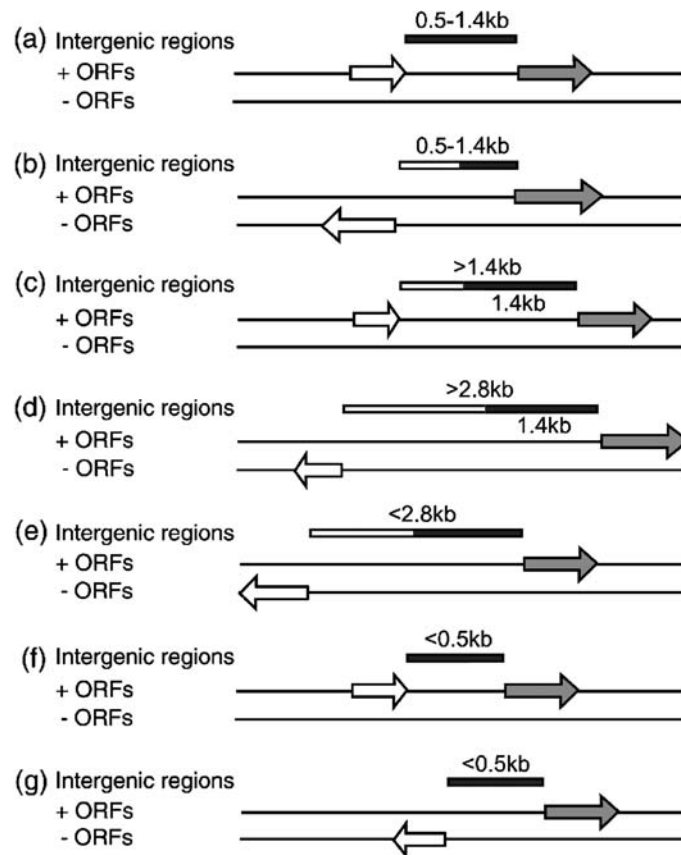
controlling primer quality; preparing for output file; building a primer database for all genes in the genome; and offering a web-based primer retrieval system for any given list of genes with a gene locus ID.

#### *Database search and promoter sequence extraction*

The pre-requisites for developing such a platform for a given organism are (1) the complete genome sequence and (2) well-annotated genome databases.

We retrieved the assembled *Arabidopsis* chromosome sequences and annotation information from MATDB – the MIPS *Arabidopsis thaliana* database (<ftp://ftpmips.gsf.de/cress/>). The annotation information included gene contig names, entry codes, gene structures, and transcription directions. The promoter region of each gene was located according to the annotation information and then was extracted from the chromosome sequences.

Representative promoter deletion analyses have shown that most *Arabidopsis* genes have functional promoters within 1400 bp of their translational start sites (Conley *et al.*, 1994; Tjaden *et al.*, 1995; Honma and Goto, 2000; Haralampidis *et al.*, 2002; Brown *et al.*, 2003). Therefore, we used 1400 bp as an upper limit for our promoter sequence selection of *Arabidopsis* genes. To select promoter fragments for microarray construction, we also considered the need for the uniformity of promoter size, so as to reduce the variation in PCR amplification yield, as well as hybridization efficiency. Therefore, the following principles were followed in selecting promoter fragments for PCR amplification. First, the longest fragment size of the PCR products was 1400 bps. Second, a minimum fragment size of the PCR products was set to 500 bps. Third, the promoter 3' end was always near and no more than 50 bps upstream of the ATG. To apply the above principles, transcription directions of the selected specific gene and the length of the intergenic region between this gene and its upstream neighbor gene were considered. These intergenic regions in the genome were grouped into 14 types, and in each case a distinct formula was used to define the promoter region for PCR amplification (Figure 2). Then the promoter sequences from these defined promoter regions were extracted from the chromosome sequences, stored in the database, and used for primer selection. A database containing annotation and promoter sequences of each gene in *Arabidopsis* with frequent update of new genome annotation is available online (<http://www.yale.edu/denglab/arabpromoter.htm> and Supplementary Data 1 at <http://plantgenomics.biology.yale.edu/>). However, the sequence data used for promoter extraction for the initial promoter microarray was based on the July 5, 2001 version of the MIPS *Arabidopsis thaliana* database (<ftp://ftpmips.gsf.de/cress/>). Due to the lack of full-length cDNA sequences and



**Figure 2.** Principles of locating the putative promoter region for individual genes. The gray arrow represents the target gene, and the bar represents the intergenic region between the target gene and its upstream gene. The gray part of the bar represents the calculated putative promoter region. The figure only shows those cases where the target gene ORFs are on the plus strand. There were seven situations for this case: (a) When the upstream intergenic region of the target gene is between 0.5 and 1.4 Kb and does not contain the promoter region of its upstream gene, the entire intergenic region is assigned as the putative promoter region. (b) When the upstream intergenic region of the target gene was between 0.5 and 1.4 Kb and contained the promoter region of its upstream gene (two divergently transcribed genes shared one intergenic region), half of the intergenic region was assigned as the putative promoter region. (c) When the upstream intergenic region of the target gene was greater than 1.4 Kb, but less than 2.8 Kb, and did not contain the promoter region of its upstream gene, a 1.4 Kb intergenic region was assigned as the putative promoter region. (d) When the upstream intergenic region of the target gene was greater than 2.8 Kb and contained the promoter region of its upstream gene, a 1.4 Kb intergenic region in length was assigned as the putative promoter region. (e) When the upstream intergenic region of the target gene was greater than 1.4 Kb, but less than 2.8 Kb, and contained the promoter region of its upstream gene, half of the intergenic region was assigned as the putative promoter region. (f) When the upstream intergenic region of the target gene was less than 0.5 Kb and did not contain the promoter region of its upstream gene, the entire intergenic region was assigned as the putative promoter region. However, the extracted putative promoter region was extended into the coding region to ensure that the fragment was at least 0.5 Kb. (g) When the upstream intergenic region of the target gene was less than 0.5 Kb and contained the promoter region of its upstream gene, the same putative promoter region was assigned as (f), except that this region was assigned to the two adjacent genes and was only considered once in the primer selection. In the same way, there were also seven possibilities if the target gene was on a minus strand, and the same principles could be used to extract the promoters.

incomplete annotation of the genome at the time, a small fraction of the promoters were somewhat shorter than expected due to long 5' UTR or the presence of an intron within the 5' UTR. This deficiency would need to be corrected in future versions of the promoter microarray.

#### *Preparing the Primer3 input and picking primer pairs with Primer3*

The primer design software Primer3 (Rozen and Skaletsky, 2000) was used to pick primers using the promoter sequences mentioned above as tem-

plate. Primer3 allowed multiple template sequences as the input, as long as each template sequence was accompanied by pre-determined parameters. The parameters for designing forward and reverse primers are described in the experimental procedure section.

#### *Primer quality control and primer pair output formatting*

In general, the output of Primer3 provided multiple primer pairs for a single promoter region. These primer pairs were ranked based on an overall evaluation of a dozen indices by Primer3. The values of these indices for each primer pair were also listed in the output file. We set the requirements to output the 10 best possible pairs of primers for each template. For 94.5% of the entire 25 626 promoters, it could successfully output primers under the standard parameters. For 5.5% of the remaining promoters, there was no primer output under the previous parameter constraints. For this later group of promoters, some of the parameter constraints were loosened until Primer3 could select a PCR primer pair. The resulting output primers, as well as the corresponding brief explanation for each input template, are summarized in a data set (Supplementary Data 2 at <http://plantgenomics.biology.yale.edu/>).

We noted that Primer3 did not select against those primers with a GC-rich region at the 3' end, GC clusters, or repeated nucleotide units. These in our experience cause mis-priming on genomic DNA PCR amplification and result in multiple product bands. Therefore, in order to increase the success rate of PCR, an additional score system was developed for the filtration of the 10 pairs of primers output by Primer3. For these primers with any one of the above problems, a specific number of points were subtracted as the penalty from an initial score of 100 for each perceived problem (Supplementary Table 1). Then the 3 primer pairs with the highest final scores were selected for each gene. If no satisfactory PCR product was obtained by the first pair of primers, another one of the two remaining pairs of primer candidates could be used to try again.

#### *Primer pair database construction and primer retrieval for a given list of genes*

We built a primer database (<http://www.yale.edu/denglab/arabpromoter.htm> and Supplementary

Table 1. Classification of transcription factors.

Gene family	Number
MYB	258
AP2/EREBP	118
bHLH	108
NAC	106
C2H2(Zn)	152
HB	83
MADS	91
bZIP	82
WRKY(Zn)	70
GARP A. G2-like	50
GARP B. ARRB class	3
C2C2(Zn) A. DOF	35
C2C2(Zn) B. CO-like	32
C2C2(Zn) C. GATA	22
C2C2(Zn) D. YABBY	5
CCAAT A. HAP2 TYPE	10
CCAAT B. HAP3 TYPE	13
CCAAT C. HAP5 TYPE	13
CCAAT D. DR1	0
GRAS	32
TRIHILIX	9
HSF	16
TCP	21
ARF	23
C3H-TYPE1(Zn)	46
C3H-TYPE2(Zn)	26
SBP	16
NINlike	14
ABI3/VP1	11
TUB	9
E2F/DP	8
CPP(Zn)	8
ALFIN-like	7
EIL	6
LFY	1
AUX/IAA	28
HMG-box	13
ARID	5
JUMONJI	8
PCG;E(z) CLASS	4
Others	18
Total	1580

Data 3 at <http://genomics.biology.yale.edu/>) for the selected primer pairs of all the genes after the additional quality control described above. A primer pair retrieval system was developed based on this database. The retrieval system is web-interfaced, and is accessible from any internet-connected terminal. With this interface, by inputting a gene locus ID or a set of gene locus IDs, users can get 3 candidate primer pairs for each gene promoter in the order of quality rank along with their detailed information (Figure 3).

(a)

**Selecting Primer Pair by Gene Locus ID**

**Arabidopsis Gene Locus ID (Eg., *At1g20020*)**  
Please input one code per line for multiple queries

Primer Select:

Select Primer Pairs

Select Forward Primers

Select Reverse Primers

Last Updated: Aug 11, 2003

Submit Reset

(b)

Summary	
Definition	Transcription factors/ MYB family gene
Code	At3g01140
Annotation	putative Myb-related transcription factor
Explanation	Total OK Forward Primers: 196, Total OK Reverse Primers: 542, Total OK Primer Pairs: 17

Forward Primer 1 20mer 5' TGCCCATGTTTCCCACAS'		
Reverse Primer 1 21mer 5' CCITTCITCTIGCTCCACCA3'		
	Forward Primer	Reverse Primer
Primer TM	63.976	63.776
Primer GC	50	52.381
Primer Location	95	1400
Product Length	1306	

Forward Primer 2 20mer 5' TGCCCATGTTTCCCACAS'		
Reverse Primer 2 21mer 5' CCITTCITCTIGCTCCACCA3'		
	Forward Primer	Reverse Primer
Primer TM	63.976	63.776
Primer GC	50	52.381
Primer Location	96	1400
Product Length	1305	

Forward Primer 3 20mer 5' TGCCCATGTTTCCCACAS'		
Reverse Primer 3 21mer 5' CCITTCITCTIGCTCCACCA3'		
	Forward Primer	Reverse Primer
Primer TM	64.078	63.776
Primer GC	50	52.381
Primer Location	94	1400
Product Length	1307	

Figure 3. Input and output of the web-based interface for the selection of the promoter region primer. (a) A user-friendly website interface was employed to retrieve promoter primers by inputting a gene locus ID. (b) Three candidate primer pairs sorted by qualities with their detailed information can be obtained for each gene in the output.

#### Selection of the initial group of Arabidopsis genes

We selected 3864 known and predicted genes (about 15% of the entire *Arabidopsis* genome) to construct the initial promoter microarray (see Supplementary Data 4 at <http://plantgenomics.biology.yale.edu/>). These genes represent 1580 transcription factors (Table 1), 1341 ubiquitin-proteasome degradation related proteins (Table 2), 619 protein kinases, and 324 genes from our specially selected group, which contained control genes and genes relevant to our ongoing research. A previous study had identified 1533 transcription

Table 2. Classification of Ubiquitin proteasome degradation related proteins.

Gene family	Number
E1 (Ubiquitin-activating enzyme) and E1-like	11
E2 (Ubiquitin-conjugating enzymes) and E2-like	49
E3 HECT-domain (Ubiquitin-transferase)	7
E3 SCF SKP1 Family	20
E3 SCF Cullin-like	10
E3 SCF F-box	690
E3 SCF RBX	2
E3 U-box	36
E3 Ring-finger	417
E3 Anaphase-promoting complex (APC)	14
Ubiquitin-specific protease (UBP)	28
Ubiquitin	16
COP9 Signalosome	10
26S Proteasome	30
<b>Total</b>	<b>1341</b>

factors genes in the *Arabidopsis* genome, or about 6% of the estimated total number of genes in the genome at the time (Riechmann *et al.*, 2000). Using the 33 known families of transcription factor genes (Riechmann *et al.*, 2000), we carried out a similar database search, and identified 1580 genes (Table 1). A total of 1341 ubiquitin-proteasome degradation-related proteins were obtained based on 14 functional/structure groups (Table 2) according to the accepted classification (Gagne *et al.*, 2002; Risseuw *et al.*, 2003). About 619 predicted protein kinase genes were selected from the MIPS database.

Our specially selected group of genes includes 291 light-responsive genes, based on previous whole genome expression profile microarray analysis (Ma *et al.*, 2001). They included those genes with 10-fold or higher regulation by light or specific photosynthesis-related function groups. Both the chalcone synthase (CHS) and ribulose biphosphate carboxylase small subunit 1A (RBCS-1A) gene that represent previously characterized HY5 binding targets were included (Chattopadhyay *et al.*, 1998) and, thus, served as positive controls in the HY5 binding experiments. In addition, 14 commonly used plant transformation marker genes and human genes and a 3X SSC blank control were used as negative controls (Table 3, panel B).

#### Construction of promoter microarray

The top-ranking primer pairs of the 3864 selected gene promoters were synthesized and used for the

Table 3. Ranks and values of positive and negative controls.

Gene code	Gene name	Relative value	Rank	P-value
<i>Panel A: Ranks and values of positive controls</i>				
At5gl3930	CHS	7.06	2	0.01
Atlg67090	RBCS-1A	3.75	32	0.02
<i>Panel B: Ranks and values of negative controls</i>				
Abbreviation				
G10	G10homolog(edg-2)	2.31	168	0.07
BT	B. Thuringiensis cry 1 Ac	2.21	199	0.08
MG	Beta2 microglobulin	1.62	536	0.18
MYL	Myosin light chain 2 (IMAGE: 1592600)	1.48	675	0.22
LUC	Luciferase	0.98	1686	0.51
NPTII	Kanamycin/neomycin phosphotransferase	0.83	2124	0.64
PGK	Phosphoglycerate kinase (pgkl)	0.80	2208	0.66
HSP	HSPC120	0.73	2455	0.72
MYH	Myosin heavy chain(IMAGE: 1593605)	0.70	2545	0.71
GLB	Globin	0.54	3060	0.87
IGF	Insulin-like growth factor(IMAGE: 1576490)	0.51	3118	0.89
GFP	Green fluorescent protein	0.38	3379	0.95
BRP	B-cell receptor protein (IMAGE: 1420858)	0.38	3383	0.95
FLJ	FLJ10917fis	0.24	3496	0.98

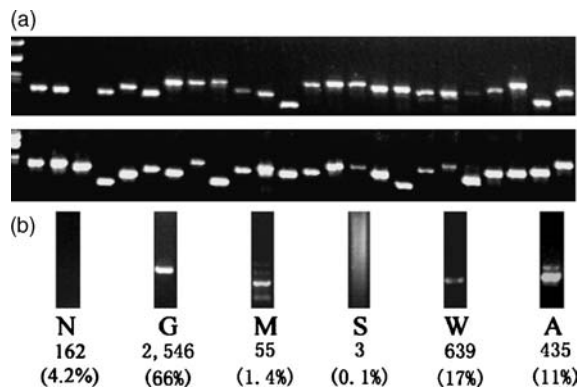
The relative intensities and the *P*-values of each spot were calculated, and spots were sorted by their values. Then the ranks, the relative values, and the *P*-values of the positive and negative controls were retrieved as shown in the table. In panel A, the two well-characterized HY5 binding targets, *CHS* and *RBCS-1A* promoters, have high relative value and rank. Panel B showed the relative intensities of 15 transgenes and human genes and their ranks.

promoter PCR amplification using a genomic DNA template with a 96-well format. The PCR products were gel electrophoresis analyzed and were scored based on six quality classifications G, A, W, M, S and N (Figure 4). The portion of each classification is shown in Figure 4b. The classification of each PCR product is shown in the gene annotation file in Supplementary Data 4. The failed or disqualified PCR amplification reactions were repeated twice with less stringent PCR conditions. If they still failed, then the second best primer pairs were synthesized, and used for PCR amplification, again for three tries. With the two rounds of primer-pair selections, about 95% of the promoters resulted in desirable quantities of quality PCR products. All PCR products were validated by their sizes consistent with the prediction. All of the promoter PCR products, as well as the 16 controls (2 positive controls and 14 negative controls), were printed on poly-L-lysine coated glass slides onto 16 subarrays. Each DNA sample was spotted in duplicate, and the average value of the two duplicates was used in all the subsequent analyses. The 16 positive and negative controls were distributed in six different subarrays. Slides were post-processed by re-mois-

turizing, snap drying, and UV cross-linking. A summary of our microarray information that consistent with the minimum information about a microarray experiment (MIAME) standard is provided in Supplementary Data 6.

#### *Quality assessment of the promoter microarray*

The DNA content in an individual spot on the array was semi-quantified by hybridization with labeled genomic DNA fragments (Figure 5a). After scanning of the hybridization image, the spot intensity approximately reflects the amount of DNA in each spot. A diagram of the spots' intensity distribution is shown in Figure 6a, and the numbers of spots with intensity values below several selected threshold levels are summarized in Supplementary Table 2. To examine the inter-slide and intra-slide variability, we calculated Pearson correlation coefficients between intensities of duplicate spots for individual arrays, and correlation coefficients between arrays. The two types of correlation coefficients were quite consistent each time in multiple independent experiments. The correlation coefficient between duplicate spots in



**Figure 4.** Gel electrophoresis and quality control of the PCR products. All of the promoter PCR amplification products were purified, analyzed by gel electrophoresis, recorded, and scored. (a) Examples of gel electrophoresis for a batch of PCR products. (b) The PCR products were scored by six classifications based on the gel electrophoresis results: N – no PCR product visible; G – single band, right size, and high intensity; M – multiple bands were roughly same intensity; S – smear; W – single band, right size, but low intensity; A – acceptable, multiple bands, but was a single dominant band with correct size visible. The PCR products with score G, W, and A were considered successful, and were kept; The PCR products with score N, M, and S were considered failures and were repeated. The number of promoter PCR products in each class was indicated under each demonstration gel picture, with the percentage of the total in the bracket.

the same slide was in the range of 0.95 or higher, and the correlation coefficient between slides was around 0.84–0.86, which suggested a reasonable consistency. The correlation coefficient for intra-slide duplicated spots are higher than inter-slides spots, suggesting printing and slide handling is the major source of the variability.

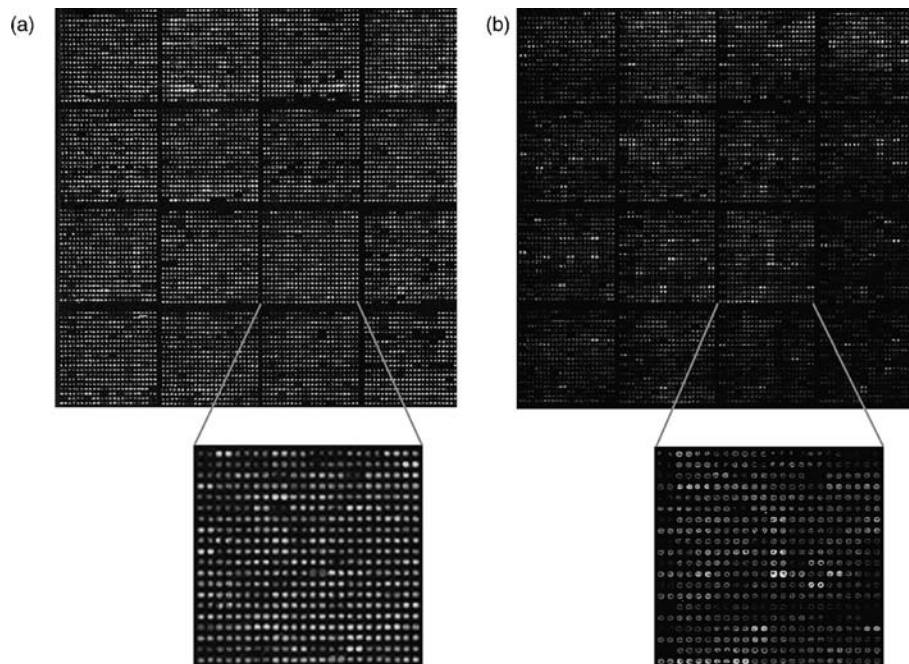
In order to further check the inter-slide variability we calculated the distribution of CVs (coefficient of variance = standard deviation/mean) of the signal intensities for all the features on the slides. A typical result is shown in Supplementary Table 3. This analysis indicates that the CVs for about 50% of the features are less than 0.10 (or 10% of mean), and the average CV (standard deviation) is 0.16 (0.14). We also analyzed the intra-slide variability by calculating the distribution of CVs of the intensities for two duplicated spots on the same slide, and a representative result is shown in Supplementary Table 4. This analysis indicated that about 70% of features in all our tests have CVs below 0.10. The average CV is consistently around 0.09–0.10, which suggested relatively small intra-slide variability in general.

#### *Analysis of HY5-promoter binding in vitro*

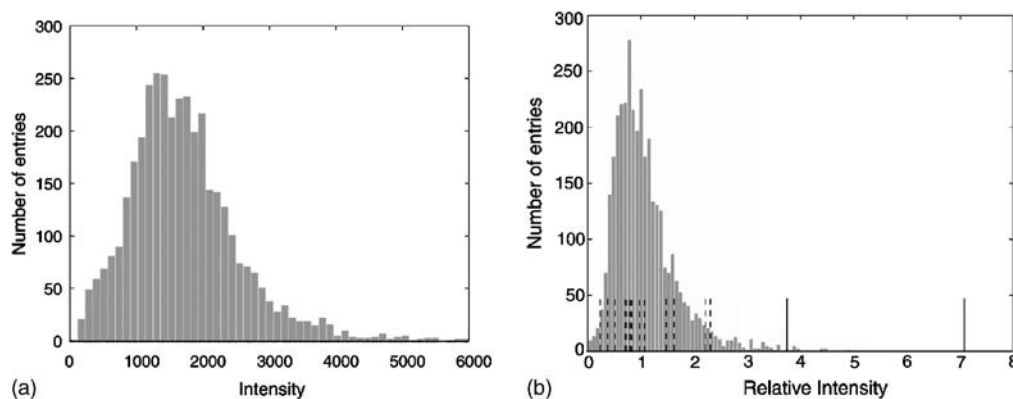
In order to identify the HY5 binding promoters *in vitro*, we did two sets of experiments. One set is genomic DNA binding experiment as described above, the other set is the *in vitro* HY5-promoter binding experiment by incorporating the labeled HY5 protein probe with the promoter microarray. Based on a pilot test (data not shown) with a range of concentration of representative promoters and control DNA fragments, it was evident that the HY5 binding intensity normalized against the promoter DNA amounts better reflects the affinity between the giving promoter and HY5 in our binding assay. Thus the ratio of the GST-HY5 binding intensity and the DNA spot intensity after hybridization with genomic DNA probes was calculated, which we call 'relative promoter binding value'. All the elements on the microarray were ranked by this ratio and the HY5 binding candidates were determined by their rank.

The glutathione S transferase (GST)-HY5 fusion protein (GST-HY5) was expressed in *Escherichia coli*, and was affinity purified. The purified protein was labeled by Cy3 dye. After the unincorporated Cy3 dye had been removed, the labeled GST-HY5 was used to probe the promoter microarray slides. Slides were scanned after free GSTHY5 had been removed, and the resulting images were analyzed (Figure 5b). The correlation coefficients were calculated after four independent GST-HY5 protein labeling and *in vitro* binding experiments. In each experiment two replicate slides were used. The correlation coefficient between duplicate spots in each slide was around 0.93, and the correlation coefficient between the two replicated slides was about 0.85.

To further check the inter-slide variability we calculated the distribution of CVs of the signal intensities for all the features on the two slides of one representative experiment (see Supplementary Table 3). In this typical example, the CVs for about 50% of the features are less than 0.10, and the average CV is 0.11 with a standard deviation of 0.09. This analysis supports a reasonably good consistency between the replicated slides. We also checked the inter-slide variability by examining the distribution of CVs of the intensities for two duplicated spots on the same slide, and a representative result is shown in Supplementary Table 4. In this typical example, about 70% of



*Figure 5.* Microarray image after hybridization with genomic DNA and probed with labeled GST-HY5 protein. (a) Individual spots on the array were quantified by the hybridization with a randomly sheared and Cy3-labeled genomic DNA, and 95% of the spots were found to produce effective signals in this array. This matched well with the fact that approximately 5% of the 7680 spots printed on the array were defective due to PCR failure. (b) Microarray image after being probed with the Cy3-labeled GST-HY5 protein. The differential binding of specific spots in comparison with the image in panel (a) was evident.



*Figure 6.* Distribution of the genomic DNA binding intensities and the HY5 relative intensities. (a) The distribution of spot intensities in the genomic DNA hybridization experiment. The spots whose intensities were not 200-arbitrary units higher than background intensity were removed before the data analysis. Note that most of the spots (81%) with intensities fell within the range of 500–2500 units. (b) HY5 binding was quantified by the relative intensity of each spot. The greater the relative intensity, the more significant the binding. The dotted bars represent 14 negative controls that were transgenes and human genes; and the black bars represent 2 positive controls that were the *CHS* gene promoter and the *RBCS-1A* gene promoter, which ranked 2nd ( $P$ -value 0.01) and 32nd ( $P$ -value 0.02), respectively.

duplicated features have CVs below 0.10. The average CV is 0.07, with a standard deviation of 0.06–0.07, which suggested a very small intra-slide variability in general.

In order to test whether the GST portion of the GST-HY5 fusion protein had a significant effect on HY5 binding to promoters on the slides, the duplicated slides were probed with

Table 4. List of the 42 genes whose promoters exhibited high HY5 binding affinity.

Gene code	Relative binding value	P-value	Gene group	Gene family/name
At4g27090	7.39	.008	Light responsive gene	Ribosomal protein L14 -like protein
At5g13930	7.06	.008	Light responsive gene	Chalcone synthase
At3g16320	7.06	.008	Ubiquitin proteasome degradation related	E3 Anaphase-promoting complex (APC)
At3g05670	6.81	.008	Ubiquitin proteasome degradation related	E3 Ring-finger
At3g58930	5.86	.010	Ubiquitin proteasome degradation related	E3 SCF (F-box)
At2g20580	5.66	.010	Ubiquitin proteasome degradation related	26s Proteasome
At5g42460	5.12	.012	Ubiquitin proteasome degradation related	E3 SCF (F-box)
Atlg23540	5.01	.012	Protein kinase	Putative serine/threonine protein kinase
At4g27410	4.94	.0013	Transcription factor	NAC
Atlg01150	4.92	.013	Transcription factor	MYB Superfamily
At3g45580	4.68	.014	Ubiquitin proteasome degradation related	E3 Ring-finger
Atlg19000	4.54	.014	Transcription factor	MYB Superfamily
At5g47850	4.50	.016	Protein kinase	Receptor kinase-like protein
At5g41440	4.49	.015	Ubiquitin proteasome degradation related	E3 Ring-finger
At2g07180	4.46	.015	Protein kinase	Putative protein kinase
At4g33190	4.41	.015	Transcription factor	MADS
Atlg68400	4.40	.015	Protein kinase	Putative receptor kinase
Atlg53820	4.39	.015	Ubiquitin proteasome degradation related	E3 Ring-finger
At5g22670	4.30	.017	Ubiquitin proteasome degradation related	E3 SCF (F-box)
Atlg70320	4.14	.017	Ubiquitin proteasome degradation related	E3 HECT -domain (Ubiquitin-transferase)
At4g33210	4.02	.018	Ubiquitin proteasome degradation related	E3 SCF (F-box)
At2g01950	3.97	.019	Protein kinase	Putative receptor protein kinase
Atlg52490	3.95	.019	Ubiquitin proteasome degradation related	E3 SCF (F-box)
Atlg20800	3.93	.019	Ubiquitin proteasome degradation related	E3 SCF (F-box)
At5g38410	3.92	.019	Light responsive gene	RuBisCO small subunit 3b
Atlg69220	3.90	.020	Protein kinase	Putative serine/threonine kinase
At2g26860	3.88	.020	Ubiquitin proteasome degradation related	E3 SCF (F-box)
Atlg0210	3.88	.020	Protein kinase	Putative mitogen-activated protein kinase
At5g49720	3.86	.020	Light responsive gene	Cellulase homolog OR16pep precursor
Atlg19680	3.83	.020	Ubiquitin proteasome degradation related	E3 Ring-finger
Atlg48220	3.83	.020	Protein kinase	Pto kinase interactor 1
Atlg67090	3.75	.022	Light responsive gene	RuBisCO small subunit la
At5g03000	3.73	.021	Ubiquitin proteasome degradation related	E3 SCF (F-box)
Atlg05080	3.66	.022	Ubiquitin proteasome degradation related	E3 SCF (F-box)
Atlg57800	3.61	.023	Ubiquitin proteasome degradation related	E3 Ring-finger
Atlg23980	3.60	.023	Ubiquitin proteasome degradation related	E3 Ring-finger
Atlg11340	3.60	.023	Protein kinase	Putative receptor kinase
At2g37180	3.59	.024	Light responsive gene	Aquaporin water channel protein
Atlg76180	3.58	.024	Cold regulatory protein	ERD14
Atlg31420	3.58	.024	Protein kinase	Hypothetical protein
At3g10510	3.54	.024	Ubiquitin proteasome degradation related	E3 SCF (F-box)
At2g17830	3.54	.024	Ubiquitin proteasome degradation related	E3 SCF (F-box)

labeled GST-HY5 alone, or with labeled GST-HY5 plus a two molar excess of GST protein as a competitor. A correlation coefficient of 0.80 was obtained between the two experiments. This result indicated that the GST tag in the GST-HY5 fusion protein might have some effects on the HY5 binding, but that they are relatively minor. This result was somewhat consistent with previous gel shift assay experiments that showed that GST had no detectable effect on GST-HY5 binding with the ribulose biphosphate carboxylase small subunit (RBCS)-1A minimal light-responsive promoter (Chattopadhyay *et al.*, 1998) and the CHS promoter (Holm, 2002). We suspected that GST might largely affect the non-specific association between GST-HY5 and DNA in general. However, the GST addition does not seem to affect specific binding to HY5 targets in the microarray experiment. For example, the two hybridization sets (with or without GST competitive addition) share 37 common promoters out of the 40 top binding targets (92.5%).

#### *Identification of the HY5 binding candidate promoters*

After the two sets of experiments, we calculated the 'relative promoter binding value', which is the ratio of the GST-HY5 binding intensity and the DNA spot intensity after hybridization with genomic DNA probes. The HY5 binding assay and the genomic DNA hybridization experiment could not be performed on the same microarray slide due to their mutually exclusive experimental procedures. This brings some additional variations to the relative promoter binding value, because the amount of promoter DNA printed on the duplicated slides might not be exactly identical. To minimize this variation, we always used the adjacent slides from the same printing batch for each experiment set, one for DNA quantification and one for HY5 binding assay. To make the GST-HY5 binding strength of distinct promoters comparable, we first normalized our HY5 binding intensities and the genomic DNA intensities using the median centering method (normalization factor = median of genomic DNA intensities/median of GST-HY5 binding intensities). After this normalization, the relative promoter binding value was calculated (see Supplementary Data 5 at

<http://plantgenomics.biology.yale.edu/>). We adopted the single-array error model (Hughes *et al.*, 2000) to assign a *P*-value to the results of microarray data.

Microarray data quality control is essential for reducing the false positive rate of binding site identification. We set a lower limit of 200 as the threshold for genomic DNA level; all those spots that have a genomic DNA intensity less than 200 were removed from further analysis. For GST-HY5 binding, some spots with extremely high binding signals (outliers) are another major source of false positive rate, as we noted that these very high signals normally represent false positives and are not reproducible. This problem may result from the fact that extremely high signals may have a non-linear relationship with DNA contents. In this case our definition of relative binding value may not be valid. To limit this problem, we removed the spots whose signals are greater than mean of intensities + 2x standard deviation of intensities. This corresponds to the range that is greater than the upper limit of 95% confidence interval of the GST-HY5 binding intensities. In this way spots with a binding signal higher than 2500 or lower than 10 were removed from the further consideration. List of all promoters that were flagged out or remained in the final analysis are provided in Supplementary data 7.

In Table 3, the ranks and relative values of the positive and negative controls are summarized. While most of the negative controls were ranked lower, both positive controls were ranked in the top 1% in relative GST-HY5 binding strength (see also Figure 6b). For example, the CHS minimal promoter ranked second, with a relative value of 7.06, as well as a *P*-value of 0.01. The RBCS-1A minimal light-responsive promoter ranked 32nd, with a relative value of 3.75, as well as a *P*-value of 0.02. We set a *P*-value < 0.025, which corresponds a relative binding value of 3.54, as a threshold for HY5-binding candidate promoters. Using this threshold, 42 out of 3543 promoters that passed the initial data quality control were selected (Table 4). These 42 candidate promoters included promoters of 4 transcription factors, 21 ubiquitin proteasome degradation-related genes, 10 protein kinases, 6 light-responsive genes, and 1 cold-responsive factor gene.

### G-box motif distribution among the HY5 binding candidates

It had been shown that HY5 could specifically and directly bind to the G-box motif, one type of light-responsive element with a palindromic hexanucleotide core CACGTG (Chattopadhyay *et al.*, 1998). To test if this G-box motif is over-represented in the promoter regions of putative HY5 target promoters, computational analysis on their PCR amplified promoter regions was performed. First, an alignment matrix, or a position-weight matrix (PWM) (Stormo, 2000), that represents the G-box motif was constructed from the sequence alignment of 33 known plant G-box motifs (Thijs *et al.*, 2002) in the PlantCARE database (Lescot *et al.*, 2002). The sequence logo, which depicts the G-box motif with position-weight matrix (Schneider and Stephens, 1990), is shown in Figure 7. Second, MotifScanner (Coessens *et al.*, 2003) was used to detect possible G-box motifs in a given promoter with the G-box position-weight matrix. This analysis revealed 14 G-box motif carrying promoters among the 42 putative HY5 targets based on the above mentioned threshold criteria. We also used Motif Scanner to detect exact matches for given consensus sequences for the other 13 hexamer motifs described in the PlantCARE database (<http://oberon.fvms.ugent.be:8080/PlantCARE/index.html>) and PLACE database (<http://www.dna.affrc.go.jp/htdocs/PLACE/>). The analysis was done for promoter regions of all 3485 gene promoters with lower HY5 binding affinity based on our analysis and for 42 gene promoters in our selected list. For each of the hexamer motifs we listed in Table 5, the number and the proportion of



Figure 7. G-box motif matrix diagram. The sequence logo of the G-box motif. The logo was constructed from the alignment of 33 known plant G-box motifs in the PlantCARE database (Lescot *et al.*, 2002). The sequence conservation, which was measured in bits of information, was depicted by the height of the stack of letters for each position in the G-box motif. The relative heights of the letters within a stack were proportional to their frequencies (Schneider and Stephens, 1990).

the detected potential motif-carrying promoter regions in the whole genome as well as in our 42 top binding promoters are summarized. The result shows that the G-box motif is the only one exhibited significant enrichment in the 42 selected genes, with a *P*-value of 0.05. This result suggests that the promoter microarray could be a useful tool for the large-scale identification of transcription factor binding targets *in vitro*. Together with the newly developed chromosome immunoprecipitation technique (Horak and Snyder, 2002), the promoter microarray could be a valuable tool to identify *in vivo* binding sites of a given transcription factor at a genome scale. However, the application of microarray technology is not without drawbacks. For example, although the distribution of a specific sequence motif accounts for much of the binding specificity, the context and *in vivo* properties of the chromatin surrounding the motif may provide additional specificity by defining the promoter conformations or cooperativity of transcription factors.

### Experimental procedures

#### Parameter settings for the promoter PCR primer design

There are two major classes of input parameters for Primer3 – ‘global’ input parameters and ‘sequence’ input parameters. The ‘global’ input parameters are the general parameters for Primer3, and the values of these parameters persisted among the input records of all genes. For these parameters, we set the minimum primer size at 17 bases; the maximum size at 27 bases; the optimal primer size at 20 bases; the optimal primer  $T_m$  at 64.0 °C; the maximum primer  $T_m$  at 70.0 °C; the maximum  $T_m$  difference between the forward and reverse primers at 3.0 °C; the minimum primer GC content at 30%; the maximum primer GC content at 70%; the optimal primer GC content at 50%; the maximum number of *N* (unknown bases) in the primer sequence at 1; the product size range at 270–1600 base pairs; the number of the output primers at 10; the primer GC clamp at 0 (by default); the maximum alignment score for both the primer self-complementation and complementation between forward and reverse primers at 4.00; the maximum length of a mononucleotide repeat

Table 5. Enrichment of known sequence motifs in selected hy5 target promoters.

Name	Consensus	Annotation	Frequency in the lower ranking promoters (%)	Frequency in target promoters (%)	Enrichment <i>P</i> -value
G-Box	CACGTG	Light responsive element	22.7	33.3	0.05
HexamerAt H4	CCGTCG	Hexamer motif of A. t. histone H4 promoter	17	13.9	0.70
MYCAter DI	CATGTG	MYC recognition sequence necessary for expression of erdl (early responsive to dehydration) in dehydrated A.t.	42.2	30.2	0.94
TBoxAtGA PB	ACTTTG	“Tbox” found in the A.t. GAPB gene promoter	63.7	44.1	0.99
GCC-Core	GCCGCC	Core of GCC-box found in many pathogen-responsive genes	13.9	11.6	0.67
MY-CAtRD22	CACATG	Binding site for MYC (rd22BPI) in A.t. dehydration-responsive gene, rd22	42.2	30.2	0.94
CAT-box	GCCACT	Cis-acting regulatory element related to meristem expression	20.5	6.9	0.99
CCGTCC-box	CCGTCC	Cis-acting regulatory element related to meristem specific activation	11.7	16.2	0.18
GT1-motif MBS	GGTTAA TAACTG	Light responsive element MYB binding site involved in drought-inducibility	50.4 35.9	32.5 23.2	0.99 0.96
TCT-motif Wbox	TCTTAC TTGACC	Part of a light responsive element Wounding and pathogen response	53.1 41.4	48.8 41.8	0.71 0.48
CBox	TGACGT	Light responsive element	26.4	23.2	0.68
I-box	GATAA[T/G]	Part of a light responsive element	82.1	76.7	0.82
MYBIAt	[A/T]AA CCA	MYB recognition site found in the promoters of the dehydration-responsive gene rd22 and many other genes in A.t.	89.6	72	1.00

at 4; the first base index in the input sequence at 1; and the maximum stability for the five 3' bases of primer at 9. We also set all the other global parameters by default.

The second class of input parameters was the 'sequence' input parameters. These describe a particular input sequence to Primer3, and are reset after every boulder record. We had three sequence input parameters for Primer3: Primer sequence ID, Sequence, and Target. Target defines the primer location. For one specified target, a legal primer pair must flank it.

#### *PCR amplification and PCR products post-processing*

The pairs of forward and reverse primers for the PCR amplification were diluted into 96-well plates, and the PCR amplification reactions were performed in 96-well thermal cycling plates. For each 96-well plate PCR amplification, 96  $\mu$ l of a master

mix containing the following ingredients was made: 10  $\mu$ l 10X Taq DNA polymerase buffer, 2  $\mu$ l *Arabidopsis* genomic DNA (100 ng/ $\mu$ l), 8  $\mu$ l dNTP (10 mM/ $\mu$ l), 2.5  $\mu$ l Taq DNA polymerase (5 units/ $\mu$ l), and 73.5  $\mu$ l water. Then 2  $\mu$ l of each forward and reverse primer (10  $\mu$ M) was added. The standard PCR condition was 94 °C for 5 min, 35 cycles of 92 °C for 60 s, 58 °C for 60 s, and 72 °C for 90 s, and a final extension step at 72 °C for 10 min. One-tenth of the PCR products were subsequently analyzed by gel electrophoresis on a 1% agarose gel, and the gel images were recorded. For each PCR amplicon, the product size was checked to see if it matches the predicted size. Only the amplicon that had the correct size was accepted in order to minimize the error rate. The failed PCR amplification reactions were repeated twice with less stringent PCR conditions. If suitable PCR products were still not obtained, then the second set of primers for these genes were selected, synthesized, and used for PCR amplification again.

The PCR products that passed these quality control criteria were precipitated with 2 volumes of 100% ethanol. The DNA concentrations ranged from 50 to 500 ng/ $\mu$ l. The purified PCR products were re-suspended in appropriate amounts of water according to following rules. For the PCR products scored G and A (Figure 4b), the pellets were re-suspended in 80  $\mu$ l water; and for the PCR products scored W, the pellets were re-suspended in 40  $\mu$ l water. From all these samples, a 20  $\mu$ l solution was transferred from the 96-well plates to the 384-well plates as a working stock. They were dried, and then re-suspended again in 20  $\mu$ l 3X SSC immediately before being printed onto slides.

#### *Microarray printing and processing*

Glass microscope slides (Gold Seal # 3010) were purchased from VWR and coated with poly-L-lysine (Sigma). All of the 3800 PCR products were printed from the 384-well plate working stock onto the poly-L-lysine coated glass slides by using an arrayer with 16 microspotting pins, resulting in 16 sub-arrays on the final slides. For spotting, the temperature and humidity were maintained at 25 °C and 50%, respectively. Each sample was spotted in duplicate next to each other. Printed slides were post-processed by re-moisturizing, snap drying, and UV cross-linking. The cross-linking was done using a Stratallinker UV-crosslinker at 200 mJ.

#### *Genomic DNA hybridization*

The *Arabidopsis* genomic DNA was extracted from 7-day-old wild type *Arabidopsis* seedlings (Columbia ecotype). Two microgram genomic DNA was diluted to 500  $\mu$ l and randomly sheared into 300–400 bp fragments by sonication. Then it was concentrated to 100 ng/ $\mu$ l using a YM-30 concentrator (Micron), and labeled with cy3-dUTP (Amersham) using the BioPrime DNA Labeling System (Invitrogen) according to the manufacturer's protocol. Labeled DNA was cleaned up by YM-30 and heat denatured at 90 °C for 2 min. Microarray slides were pre-hybridized with pre-hybridization buffer (25% formamide, 5X SSC, 0.1% SDS, 1% BSA) and successively washed with water, 70% ethanol, and 100% ethanol. Then the labeled DNA was applied to the slide in hybridization buffer, (50% formamide, 10X SSC, 0.2% SDS). After hybridization at 42 °C in a HybChamber<sup>TM</sup> (Genemachines) for one

overnight, the slide was washed twice successively at room temperature with 2X SSC/0.1% SDS, 0.2X SSC/0.1% SDS, 0.2X SSC, and 0.02X SSC. The hybridization signal was then scanned using a Genepix 4000B scanner (Axon Instruments).

#### *GST-HY5 expression and purification*

Glutathione S-transferase (GST)-HY5 fusion protein production in *Escherichia coli* was described previously (Holm *et al.*, 2002). GST-HY5 was affinity purified according to the manufacturer's standard procedure (New England Biolabs). The purified GST-HY5 was concentrated to 1mg/ml using a YM-50 concentrator (Micron).

#### *In vitro HY5-DNA binding experiment*

About 1 mg GST-HY5 was labeled by the Cy3 monofunction protein labeling kit (Amersham Pharmacia) according to the manufacturer's standard procedure. To optimize the hybridization, the labeled GST-HY5 was diluted to 10X, 100X, 1000X and 10 000X using gel shift assay binding buffer (15 mM HEPES, pH7.5, 35 mM KCl, 1 mM EDTA, 6% glycerol, 1 mM DTT, 2  $\mu$ g poly(dI-dC), 1% Top-block (Sigma)). We found that 10X and 100X dilution produced saturated signals while 10 000X dilution produced weak signals with the main peak of the intensity below 1000. Therefore we chose 1000X dilution which produced the signals with the main peak of the intensity around 3000. The microarray slide was blocked with 1% Top-block for one hour and rinsed with 1X PBS/0.1% Tween-20. Then the 200  $\mu$ l probe was applied to the microarray and incubated at room temperature in darkness for 1 h. The slide was washed with 1X PBS/0.1% Tween-20 three times at room temperature, spun down to dry, and then scanned.

#### *Calculation of P-value for HY5 binding*

We have adopted the whole chip error model (Hughes *et al.*, 2000) to calculate the confidence P-value of HY5 binding for each promoter region on the microarray.

The statistic used to define significance is

$$X = (a_2 - a_1) / [s_1^2 + s_2^2 + f^2(a_1^2 + a_2^2)]^{1/2} \quad (1)$$

where  $a_1$  is the GST-HY5 binding intensity;  $a_2$  is the genomic DNA intensity measured for each spot;  $s_1$  and  $s_2$  are the uncertainties due to back-

ground subtraction for HY5 binding signal and genomic DNA signal; and  $f$  is a fractional multiplicative error such as from hybridization non-uniformities, fluctuations in the dye incorporation efficiency, scanner gain fluctuations, etc. The Statistic  $X$  was found approximately normal (Hughes *et al.*, 2000). The parameter  $f$  was chosen so that  $X$  has unit variance. The significance of an enrichment of magnitude  $X$  is then calculated as described (Lee *et al.*, 2002, Supplementary Online Material)

$$P\text{-value} = 1 - \text{Erf}(X) \quad (2)$$

We use 0.025 as the  $P$ -value threshold for significant bindings, and the corresponding relative binding value is about 2.0.

#### Calculation of $P$ -value for hexamer motifs

The enrichment  $P$ -value was calculated based on a normal sample proportion distribution  $N(\pi, \sigma^2)$ , where  $\pi$  is the frequency of the motif-carrying promoter regions in the whole genome, and  $\sigma^2 = \pi(1-\pi)/n$ , where  $n = 42$ .

#### Acknowledgments

We are grateful to Dr Kenneth Nelson for his help in microarray printing and Montrel D. Seay and S.P. Dinesh-Kumar for sharing the *Arabidopsis* protein kinase gene list with us before publication. This research project was supported by a strategic international cooperation project grant (#302-21120261) and a regular grant (#30170092 to LJQ) from the National Natural Science Foundation of China and a grant from the National Institutes of Health (GM-47850 to XWD). Y.G. was a Peking-Yale Center Monsanto fellow.

#### References

- Ang, L.H., Chattopadhyay, S., Wei, N., Oyama, T., Okada, K., Batschauer, A. and Deng, X.W. 1998. Molecular interaction between COP1 and HY5 defines a regulatory switch for light control of *Arabidopsis* development. *Mol. Cell* 1: 213–222.
- Ang, L.H. and Deng, X.W. 1994. Regulatory hierarchy of photomorphogenic loci: allele-specific and light-dependent interaction between the HY5 and COP1 loci. *Plant Cell* 6: 613–628.
- Bowen, B., Steinberg, J., Laemmli, U.K. and Weintraub, H. 1980. The detection of DNA-binding proteins by protein blotting. *Nucl. Acids Res.* 8: 1–20.
- Brown, R.L., Kazan, K., McGrath, K.C., Maclean, D.J. and Manners, J.M. 2003. A role for the GCC-box in jasmonate-mediated activation of the PDF1.2 gene of *Arabidopsis*. *Plant Physiol.* 132: 1020–1032.
- Bulyk, M.L., Huang, X., Choo, Y. and Church, G.M. 2001. Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. *Proc. Natl. Acad. Sci. USA* 98: 7158–7163.
- Chattopadhyay, S., Ang, L.H., Puente, P., Deng, X.W. and Wei, N. 1998. *Arabidopsis* bZIP protein HY5 directly interacts with light-responsive promoters in mediating light control of gene expression. *Plant Cell* 10: 673–683.
- Choo, Y. and Klug, A. 1993. A role in DNA binding for the linker sequences of the first three zinc fingers of TFIIIA. *Nucl. Acids Res.* 21: 3341–3346.
- Choo, Y. and Klug, A. 1994. Selection of DNA binding sites for zinc fingers using rationally randomized DNA reveals coded interactions. *Proc. Natl. Acad. Sci. USA* 91: 11168–11172.
- Coessens, B., Thijs, G., Aerts, S., Marchal, K., De Smet, F., Engelen, K., Glenisson, P., Moreau, Y., Mathys, J. and De Moor, B. 2003. INCLUSive: a web portal and service registry for microarray and regulatory sequence analysis. *Nucl. Acids Res.* 31: 3468–3470.
- Conley, T.R., Park, S.C., Kwon, H.B., Peng, H.P. and Shih, M.C. 1994. Characterization of *cis*-acting elements in light regulation of the nuclear gene encoding the A subunit of chloroplast isozymes of glyceraldehyde-3-phosphate dehydrogenase from *Arabidopsis thaliana*. *Mol. Cell Biol.* 14: 2525–2533.
- Costanzo, M.C., Hogan, J.D., Cusick, M.E., Davis, B.P., Fancher, A.M., Hodges, P.E., Kondu, P., Lengieza, C., Lew-Smith, J.E., Lingner, C., Roberg-Perez, K.J., Tillberg, M., Brooks, J.E. and Garrels, J.I. 2000. The yeast proteome database (YPD) and *Caenorhabditis elegans* proteome database (WormPD): comprehensive resources for the organization and comparison of model organism protein information. *Nucl. Acids Res.* 28: 73–76.
- Egelkrou, E.M., Mariconti, L., Settlege, S.B., Cella, R., Robertson, D. and Hanley-Bowdoin, L. 2002. Two E2F elements regulate the proliferating cell nuclear antigen promoter differently during leaf development. *Plant Cell* 14: 3225–3236.
- Gagne, J.M., Downes, B.P., Shiu, S.H., Durski, A.M. and Vierstra, R.D. 2002. The F-box subunit of the SCF E3 complex is encoded by a diverse superfamily of genes in *Arabidopsis*. *Proc. Natl. Acad. Sci. USA* 99: 11519–11524.
- Garner, M.M. and Revzin, A. 1981. A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the *Escherichia coli* lactose operon regulatory system. *Nucl. Acids Res.* 9: 3047–3060.
- Hanes, S.D. and Brent, R. 1991. A genetic model for interaction of the homeodomain recognition helix with DNA. *Science* 251: 426–430.
- Haralampidis, K., Milioni, D., Rigas, S. and Hatzopoulos, P. 2002. Combinatorial interaction of *cis* elements specifies the expression of the *Arabidopsis* AtHsp90-1 gene. *Plant Physiol.* 129: 1138–1149.
- Holm, M., Ma, L.G., Qu, L.J. and Deng, X.W. 2002. Two interacting bZIP proteins are direct targets of

- COP1-mediated control of light-dependent gene expression in *Arabidopsis*. *Genes Dev.* 16: 1247–1259.
- Hong, R.L., Hamaguchi, L., Busch, M.A. and Weigel, D. 2003. Regulatory elements of the floral homeotic gene *AGAMOUS* identified by phylogenetic footprinting and shadowing. *Plant Cell* 15: 1296–1309.
- Honma, T. and Goto, K. 2000. The *Arabidopsis* floral homeotic gene *PISTILLATA* is regulated by discrete *cis*-elements responsive to induction and maintenance signals. *Development* 127: 2021–2030.
- Horak, C.E., Mahajan, M.C., Luscombe, N.M., Gerstein, M., Weissman, S.M. and Snyder, M. 2002. GATA-1 binding sites mapped in the beta-globin locus by using mammalian ChIP-chip analysis. *Proc. Natl. Acad. Sci. USA* 99: 2924–2929.
- Horak, C.E. and Snyder, M. 2002. ChIP-chip: a genomic approach for identifying transcription factor binding sites. *Meth. Enzymol.* 350: 469–483.
- Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D., Kidd, M.J., King, A.M., Meyer, M.R., Slade, D., Lum, P.Y., Stepaniants, S.B., Shoemaker, D.D., Gachotte, D., Chakraburty, K., Simon, J., Bard, M. and Friend, S.H. 2000. Functional discovery via a compendium of expression profiles. *Cell* 102: 109–126.
- Iyer, V.R., Horak, C.E., Scafe, C.S., Botstein, D., Snyder, M. and Brown, P.O. 2001. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* 409: 533–538.
- Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., Zeitlinger, J., Jennings, E.G., Murray, H.L., Gordon, D.B., Ren, B., Wyrick, J.J., Tagne, J.B., Volkert, T.L., Fraenkel, E., Gifford, D.K. and Young, R.A. 2002. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* 298: 799–804.
- Lescot, M., Dehais, P., Thijs, G., Marchal, K., Moreau, Y., Van de Peer, Y., Rouze, P. and Rombauts, S. 2002. PlantCARE, a database of plant *cis*-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. *Nucl. Acids Res.* 30: 325–327.
- Lopez-Molina, L., Mongrand, S., McLachlin, D.T., Chait, B.T. and Chua, N.H. 2002. ABI5 acts downstream of ABI3 to execute an ABA-dependent growth arrest during germination. *Plant J.* 32: 317–328.
- Ma, L., Li, J., Qu, L., Hager, J., Chen, Z., Zhao, H. and Deng, X.W. 2001. Light control of *Arabidopsis* development entails coordinated regulation of genome expression and cellular pathways. *Plant Cell* 13: 2589–2607.
- Markstein, M., Markstein, P., Markstein, V. and Levine, M.S. 2002. Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the *Drosophila* embryo. *Proc. Natl. Acad. Sci. USA* 99: 763–768.
- Oliphant, A.R., Brandl, C.J. and Struhl, K. 1989. Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein. *Mol. Cell Biol.* 9: 2944–2949.
- Orian, A., van Steensel, B., Delrow, J., Bussemaker, H.J., Li, L., Sawado, T., Williams, E., Loo, L.W., Cowley, S.M., Yost, C., Pierce, S., Edgar, B.A., Parkhurst, S.M. and Eisenman, R.N. 2003. Genomic binding by the *Drosophila* Myc, Max, Mad/Mnt transcription factor network. *Genes Dev.* 17: 1101–1114.
- Oyama, T., Shimura, Y. and Okada, K. 1997. The *Arabidopsis* HY5 gene encodes a bZIP protein that regulates stimulus-induced development of root and hypocotyl. *Genes Dev.* 11: 2983–2995.
- Pepper, A.E. and Chory, J. 1997. Extragenic suppressors of the *Arabidopsis* *det1* mutant identify elements of flowering-time and light-response regulatory pathways. *Genetics* 145: 1125–1137.
- Ramirez-Parra, E., Frundt, C. and Gutierrez, C. 2003. A genome-wide identification of E2F-regulated genes in *Arabidopsis*. *Plant J.* 33: 801–811.
- Ren, B., Cam, H., Takahashi, Y., Volkert, T., Terragni, J., Young, R.A. and Dynlacht, B.D. 2002. E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints. *Genes Dev.* 16: 245–256.
- Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T.L., Wilson, C.J., Bell, S.P. and Young, R.A. 2000. Genome-wide location and function of DNA binding proteins. *Science* 290: 2306–2309.
- Riechmann, J.L., Heard, J., Martin, G., Reuber, L., Jiang, C., Keddie, J., Adam, L., Pineda, O., Ratcliffe, O.J., Samaha, R.R., Creelman, R., Pilgrim, M., Broun, P., Zhang, J.Z., Ghandehari, D., Sherman, B.K. and Yu, G. 2000. *Arabidopsis* transcription factors: genome-wide comparative analysis among eukaryotes. *Science* 290: 2105–2110.
- Risseuw, E.P., Daskalchuk, T.E., Banks, T.W., Liu, E., Cotelesage, J., Hellmann, H., Estelle, M., Somers, D.E. and Crosby, W.L. 2003. Protein interaction analysis of SCF ubiquitin E3 ligase subunits from *Arabidopsis*. *Plant J.* 34: 753–767.
- Rozen, S. and Skaletsky, H.J. 2000. Primer3 on the WWW for general users and for biologist programmers. *Meth. Mol. Biol.* 132: 365–386 (Code available at [http://www.genome.wi.mit.edu/genome\\_software/other/primer3.html](http://www.genome.wi.mit.edu/genome_software/other/primer3.html)).
- Ruvkun, G. and Hobert, O. 1998. The taxonomy of developmental control in *Caenorhabditis elegans*. *Science* 282: 2033–2041.
- Saha, S., Nicholson, A. and Kapler, G.M. 2001. Cloning and biochemical analysis of the tetrahymena origin binding protein TIF1: competitive DNA binding *in vitro* and *in vivo* to critical rDNA replication determinants. *J. Biol. Chem.* 276: 45417–45426.
- Schneider, T.D. and Stephens, R.M. 1990. Sequence logos: a new way to display consensus sequences. *Nucl. Acids Res.* 18: 6097–6100.
- Simon, I., Barnett, J., Hannett, N., Harbison, C.T., Rinaldi, N.J., Volkert, T.L., Wyrick, J.J., Zeitlinger, J., Gifford, D.K., Jaakkola, T.S. and Young, R.A. 2001. Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell* 106: 697–708.
- Stathopoulos, A., van Drenth, M., Erives, A., Markstein, M. and Levine, M. 2002. Whole-genome analysis of dorsal-ventral patterning in the *Drosophila* embryo. *Cell* 111: 687–701.
- Stormo, G.D. 2000. DNA binding sites: representation and discovery. *Bioinformatics* 16: 16–23.
- Tjaden, G., Edwards, J.W. and Coruzzi, G.M. 1995. *Cis* elements and trans-acting factors affecting regulation of a

- nonphotosynthetic light-regulated gene for chloroplast glutamine synthetase. *Plant Physiol.* 108: 1109–1117.
- Thijs, G., Marchal, K., Lescot, M., Rombauts, S., de Moor, B., Rouzé, P. and Moreau, Y. 2002. A Gibbs Sampling method to detect over-represented motifs in upstream regions of coexpressed genes. *J. Comp. Biol.* 9: 447–464.
- Wei, N., Kwok, S.F., von Arnim, A.G., Lee, A., McNellis, T.W., Piekos, B. and Deng, X.W. 1994. Arabidopsis COP8, COP10, and COP11 genes are involved in repression of photomorphogenic development in darkness. *Plant Cell* 6: 629–643.
- Weinmann, A.S., Yan, P.S., Oberley, M.J., Huang, T.H. and Farnham, P.J. 2002. Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis. *Genes Dev.* 16: 235–244.
- Woodbury Jr., C.P. and Von Hippel, P.H. 1983. On the determination of deoxyribonucleic acid–protein interaction parameters using the nitrocellulose filter-binding assay. *Biochemistry* 22: 4730–4737.
- Wyrick, J.J., Aparicio, J.G., Chen, T., Barnett, J.D., Jennings, E.G., Young, R.A., Bell, S.P. and Aparicio, O.M. 2001. Genome-wide location analysis of ORC and MCM proteins: high-resolution mapping of replication origins in *S. cerevisiae*. *Science* 294: 2357–2360.