Genomics 101 (2013) 187–194

Contents lists available at SciVerse ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Identification and analysis of the proximal promoters of microRNA genes in *Arabidopsis*

Xin Zhao [1], Huiyong Zhang [1], Lei Li *

Department of Biology, University of Virginia, Charlottesville, VA 22904, USA

ARTICLE INFO

Article history:
Received 27 November 2012
Accepted 24 December 2012
Available online 4 January 2013

Keywords:
Promoter
RNA polymerase II
Chromatin immunoprecipitation
microRNA
Gene expression
Arabidopsis

ABSTRACT

Endogenous microRNAs (miRNAs) modulate gene expression at the post-transcription level. In plants, a vast majority of *MIR* genes are thought to be transcribed by RNA Polymerase II (Pol II). However, promoter organization is currently unknown for most plant *MIR* genes. This deficiency prevents a comprehensive understanding of miRNA-mediated gene networks. In this study, we performed Pol II chromatin immunoprecipitation (ChIP) analysis in *Arabidopsis* using a genome tiling microarray. Distinct Pol II binding was found at most *MIR* loci, which allowed prediction of the transcription start sites (TSSs) for 167 *MIR* genes in *Arabidopsis* that was validated by average free energy profiling. By employing 99 position weight matrices (PWM), we systematically scanned the regulatory regions upstream of the TSSs. We discovered eleven and ten cis-elements that are statistically over- and under-represented in *MIR* promoters, respectively. Thus, analysis of Pol II binding provides a new perspective for studying miRNAs in plants.

1. Introduction

Following the initial discovery in *Caenorhabditis elegans* [1,2], miRNAs are recognized as a conspicuous class of regulatory small RNA molecules [3,4]. The 20 to 24 nucleotides long mature miRNAs are encoded by endogenous *MIR* genes and processed from much longer primary transcripts known as pri-miRNAs via stem–loop structured intermediates called pre-miRNAs [3,5]. In higher plants, pri-miRNA and pre-miRNA processing is carried out in the nucleus mainly by the endonuclease DICER-LIKE1 [6]. Mature miRNAs are then transported to the cytoplasm and integrated into the RNA-induced silencing complex (RISC) [7,8]. After integration into RISC, miRNAs interact with their cognate target mRNA through base pairing. In plants, such interactions typically lead to repression of gene expression through cleavage of the target transcripts [9,10] and translational attenuation [11]. Recently, down regulation of gene expression by miRNA-directed DNA methylation at the target loci has also been reported [12].

As trans-acting regulators, temporal and spatial control of the abundance of individual miRNAs is immediately relevant to our understanding of any biological process that involves miRNAs. In contrast to the numerous reports that attest to the importance of miRNA-mediated regulation of gene expression, our knowledge on the regulation of *MIR* genes themselves is sparse. Several studies indicate that plant *MIR* genes are transcribed by Pol II, similar to what is found in animals [13,14]. By sequencing the 5′ transcript ends, Xie et al. [15] mapped the TSSs for 52 *MIR* genes. This list was expanded upon through computational prediction of the core promoters [16]. In addition to the TATA box motifs located upstream of the TSSs [15], Megraw et al. [17] identified other transcription factor binding motifs in the promoter of *MIR* genes in *Arabidopsis*. They showed that within the 800 nucleotides region upstream of TSSs, sequences resembling the binding sites for the transcription factors AtMYC2, ARF, SORLREP3, and LFY were overrepresented relative to protein-coding gene promoters and randomly sampled genomic sequences [17].

While the previous studies are instrumental in establishing our working understanding of *MIR* gene transcription in plants, there are several major issues need to be addressed. Binding of Pol II to the identified *MIR* promoter regions has not been demonstrated in plants. In addition, few of the predicted cis-elements have been functionally tested. Further, the number of annotated miRNA genes has since increased and the newly identified genes need to be subject to examination for promoter regions and regulatory sequences. Compared to the previously known miRNAs, most of the new miRNA genes have narrower phylogenetic distribution and exhibit weaker expression level and more prominent tissue-specific expression [18–20]. Based on these observations, it has been argued that continuous gene birth and death allows beneficial miRNAs to be maintained while deleterious ones avoided [18,19,21,22]. Identification of the promoter regions of these *MIR* genes and comparison to those of the conserved are thus highly desirable to fully elucidate miRNA based gene regulation.

MIR genes that encode transcripts which are processed into identical or near identical mature miRNAs are grouped in paralogous families [23]. In contrast to the small but abundant miRNA families in animals, plants have fewer but larger families. For example, in

* Corresponding author. Fax: +1 434 982 5626.
E-mail address: ll4jn@virginia.edu (L. Li).
[1] These authors contributed equally to this work.

0888-7543/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ygeno.2012.12.004

Arabidopsis, the miR169 family contains at least 14 members [20,24]. *MIR* genes of the same family, although encoding identical mature miRNAs, can differ considerably in gene structure and regulatory sequences. Thus, paralogous *MIR* genes may be differentially expressed at different developmental stages or in response to various environmental stimuli. On the other hand, many families contain highly similar members, suggesting recent expansion via tandem gene duplication and segmental duplication events [25]. Therefore, the promoter regions of the paralogous members may contain shared as well as unique motifs. For example, there are six *MIR395* genes (*MIR395a–f*) in *Arabidopsis*. When the promoter of individual family member was used to drive GFP expression, it was found that some family members share the same tissue- and cell-specific patterns of GFP expression while additional GFP expression observed for individual members [26]. Globally cataloging the DNA motifs shared within a paralogous family or unique to individual members thus will help to identify their function and trace their evolution.

The goal of the current study is to comprehensively identify and analyze the proximal promoter regions of *MIR* genes in *Arabidopsis*. Toward this goal, we performed Pol II ChIP followed with a whole genome tiling microarray analysis. Based on the Pol II binding profiles, we designed a computational method to reliably predict the TSSs and hence the proximal promoter regions of 167 *MIR* genes. We show this dataset is useful in identifying cis-regulatory elements that are necessary for fully understanding the regulation of *MIR* genes and elucidating the miRNA networks.

2. Materials and methods

2.1. Plant materials and growth conditions

The plant used in this work was *Arabidopsis thaliana* ecotype Col-0. The seeds were placed on Murashige and Skoog (Sigma-Aldrich) agar plates containing 1% sucrose and incubated at 4 °C for two days after which they were exposed to continuous white light (170 μmol sec^{-1} m^{-2}) at 22 °C for four days. The seedlings were then incubated either under light or in the dark and harvested eight hours thereafter.

2.2. ChIP analyses

Chromatin isolation was performed using four-day-old whole seedlings grown under continuous white light or undergone dark-transition as previously described [27]. The resuspended chromatin pellet was sonicated at 4 °C with a Diagenode Bioruptor set at high intensity for 10 min (30 sec on, 30 sec off intervals). Chromatin was immunoprecipitated with a polyclonal anti-RNA polymerase II antibody (Santa Cruz), washed, reverse cross-linked, amplified, and hybridized to the Affymetrix At35b_MR_v04 genome tiling microarray using the manufacturer supplied protocol (Affymetrix). An aliquot of untreated sonicated chromatin was reverse cross-linked and used as a total input DNA control for microarray hybridization. Four biological replicates were hybridized to the tiling microarrays.

For ChIP-qPCR analysis, an equal amount of sonicated chromatin was incubated with IgG as a control in parallel to immunoprecipitation by the Pol II antibody. Relative abundance of regions of interest in immunoprecipitated DNA was measured by qPCR using the ABI 7500 system and the Power SYBR Green PCR master mix (Applied Biosystems). Three independent qPCR assays were performed on immunoprecipitated DNA prepared using either IgG or the Pol II specific antibody and compared with the corresponding input DNA.

2.3. Assigning transcription level to protein-coding and MIR genes

The *Arabidopsis* genome and annotation data were extracted from the TAIR 10 release of The *Arabidopsis* Information Resource [28]. Protein-coding genes were selected if they do not overlap with any other genes in both the 2 kb upstream and 2 kb downstream regions. Of the 3953 protein-coding genes meeting this criterion, 2000 were randomly selected for further analysis. RNA-Seq data from 11-day *Arabidopsis* seedlings (GSE30814) [29] were used and processed through the Tophat-Cufflinks pipeline to rank the transcription level for these genes. *MIR* genes (miRBase release 17) [24] with validated TSSs [15] and full-length cDNA support were collected. After excluding those embedded in intron of host genes, 59 *MIR* genes were eventually selected. Their transcription levels were ranked based on qRT-PCR data specifically interrogating the pri-miRNAs as previously reported [30].

2.4. Microarray data analysis

Raw microarray data was processed using Cisgenome with default parameters [31,32]. Log2 transformation was applied during normalization and only probes perfectly matched to the genome were used for intensity computation. Signal intensity from moving average statistics (MA statistic in TileMap) was used for pattern making. For each *MIR* gene, a Pol II binding profile in the −1000 to 1000 bp region relative to the TSS was drawn using a sliding window approach (window size = 100 bp, step = 5 bp). Within each window, the average signal intensity of all probes was calculated and set as the signal intensity for the current position (midpoint of the window). After scanning the entire region, the resultant series of points were lined up, the signal intensity for each point set as the value for each position, and plotted against genome coordinates to make the binding profile. For profiling multiple genes, TSSs were used to align the genes and then the same procedure carried out to average the combined dataset.

2.5. TSS prediction

The graphic characters among Pol II binding profiles of the 59 *MIR* genes with known TSSs were utilized to identify similar Pol II binding pattern for other genes. For each pre-miRNA, the region with extensive declination of Pol II binding was first identified using a sliding window approach (window size = 100 bp, step = 5 bp) to calculate the average Pol II signal intensity for a region. The midpoint of the first window with an average Pol II signal intensity 0.2 lower than the two windows (200 bp) upstream and downstream was designated as the valley. Overall, valleys were observed for 167 *MIR* genes (including 51 of the 59 with known TSSs). To predict TSSs for the 167 *MIR* genes, we developed a three-step procedure. First, the position 500 bp upstream the valley was set as the start point. Second, we searched within the 300 bp flanking sequences of the start point for TATA box like motifs. To this end, Motif Matcher (http://users.soe.ucsc.edu/~kent/improbizer/motifMatcher.html) was used to conduct a search based on PWM for TATA box from 345 experimentally verified plant promoters, which were collected from PlantProm DB [33] and *MIR* genes with identified TATA boxes [15]. Third, the same region was scanned using a PWM based on 236 experimentally verified transcription initiation motifs for dicots collected from PlantProm DB. If a TATA box was found 25 bp upstream to an initiation motif, the 5th nucleotide in the initiation motif was set as the refined TSS. If only a TATA box motif was found within the flanking region, position 25 bp downstream of the TATA box was set as the TSS. Otherwise the original start point was used as approximation for the TSS.

2.6. Average Free Energy profiling

We used dinucleotide parameters in DNA melting based on previously proposed models [34,35] to calculate the Average Free Energy (AFE). The 2000 selected protein-coding genes, 59 *MIR* genes with known TSSs, and 167 *MIR* genes with predicted TSSs were aligned within each group with the TSSs set at the +1 position. The overall

sequences in the −1000 to 1000 regions relative to the TSSs were scanned by calculating the mean value of free energy in DNA melting at each position. A previously described method was employed to reduce noise [36]. In brief, the dinucleotide parameters were averaged over a 15 bp sliding window with one nucleotide step. After that, the mean value assigned to the midpoint of each window was used to generate the AFE profile over all the sequences.

2.7. Analysis of cis-elements

To identify putative cis-regulatory elements, a previously described method was followed [17]. Briefly, PWM for 99 transcription factor binding sites were built based on experimentally validated data derived from the *Arabidopsis thaliana* Promoter Binding Element Database (http://exon.cshl.org/cgi-bin/atprobe/atprobe.pl) and the Arabidopsis Gene Regulatory Information Server dataset [37]. Threshold used for specific matrix was set as the lowest score from using the matrix against all validated binding site variants. For the 2000 selected protein-coding genes, 167 *MIR* genes, and 2000 random genome sequences (1 kb long each), we used the 99 PWM to scan all the −1, 000 to +1 bp regions. For each cis-element, proportion of sequences found to contain at least one copy of the cis-element was calculated for all three datasets (Ppc for protein-coding genes, PmiRNA for *MIR* genes, and Prandom for random sequences). Posterior Probability for all four possibilities (PmiRNA>Ppc, PmiRNA>Prandom, PmiRNA<Ppc, and PmiRNA<Prandom) was calculated using 10,000 times Monte Carlo simulation in Matlab. For specific cis-element, if posterior probability of (PmiRNA>Prandom)>0.85, the cis-element was considered to be enriched in *MIR* promoters. If posterior probability of (PmiRNA<Prandom)>0.85, the cis-element was considered to under-represented in *MIR* promoters.

2.8. Accession number

The original ChIP-chip data have been deposited in the National Institutes of Health Gene Expression Omnibus database under the accession number GSE35608.

3. Results

3.1. Profiling genome-wide Pol II binding sites in Arabidopsis

To obtain *in vivo* Pol II binding sites at the genome scale in *Arabidopsis*, we performed ChIP experiments using a commercial antibody of *Arabidopsis* origin that is specific for the N-terminus of the largest subunit of Pol II. Two independent experiments were performed in young seedlings either grown under continuous white light or undergone light-to-dark transition. Pol II-immunoprecipitated DNA was then hybridized to an Affymetrix genome tiling microarray that interrogates ~97% of the nuclear genome. After data processing as previously described [38], a global profile of Pol II binding signal was generated for both biological samples. For light-grown seedlings, we identified a total of 11,689 high-confidence Pol II-bound regions with a total length of approximately 7.8 Mb, or 6.5% of the sequenced nuclear genome. For seedlings that have undergone dark-transition, a total of 8217 Pol II-bound loci were identified that cover approximately 7.0 Mb or 5.9% of the genome.

To validate the identified Pol II occupancy, we performed two sets of experiments. First, we examined the distribution of Pol II binding activity across the five chromosomes in both samples using a sliding window approach. We found that the global Pol II binding pattern correlates in general with the gene density (Fig. S1), suggesting that the detected Pol II binding reflects the transcriptional activity. As an example, analysis of chromosome 1 is illustrated in Fig. 1A. Second, we randomly selected 13 Pol II-occupied regions identified from either the light or dark-transition samples and performed quantitative PCR analysis following the ChIP assay (ChIP-qPCR). As shown in Fig. 1B, specific Pol II binding was confirmed for all 13 examined loci, attesting to the reliability of the microarray analysis.

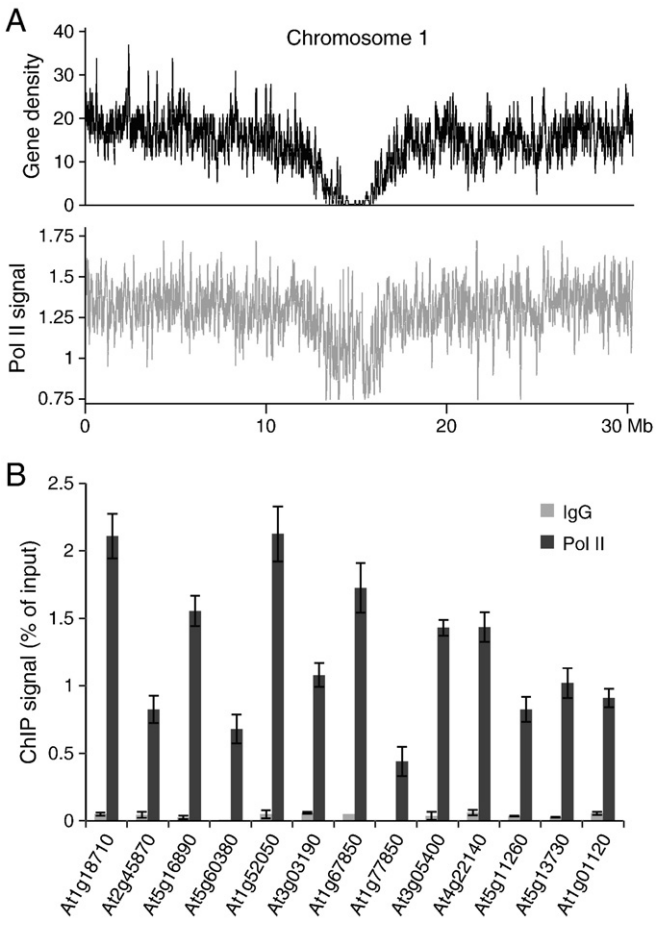

Fig. 1. Identification and confirmation of Pol II binding along the *Arabidopsis* genome. (A) Distribution of gene density (top track) and Pol II signal (bottom track) from the light sample along chromosome 1. The tracks were generated using 50 Kb sliding windows with 1 Kb step. Within each window, number of annotated genes and average Pol II binding signal were calculated and aligned to the chromosomal coordinate. (B) ChIP-qPCR confirmation of Pol II binding on selected loci. From microarray data, 13 Pol II-occupied regions were randomly selected. ChIP-qPCR was performed using either IgG or the Pol II specific antibody and normalized against the input genomic DNA. Error bars indicate standard deviation derived from three independent qPCR experiments.

3.2. Characteristic Pol II binding around the TSSs of MIR genes

To characterize Pol II binding at the gene level, we first selected 2000 protein-coding genes that contain reliable information on the TSSs. We grouped these genes into three equal-sized subsets based on their ranked transcription level (see Materials and methods). We then determined Pol II binding profiles in the light sample, which is consistent with conditions under which the expression data was obtained, by plotting the probe intensity from tiling microarray against the distance from the TSSs. We found that protein-coding genes with different transcription levels possess distinct Pol II binding profiles around the TSSs (Fig. 2A). Genes ranked in the top one-third in terms of transcription level collectively show a strong Pol II binding peak at the TSSs. The profile then gradually increases in the gene bodies (Fig. 2A). By contrast, for genes ranked in the middle and bottom one-third, the overall Pol II binding profiles are distinct from the high expression genes. For these genes, no strong Pol II binding peak at the TSSs was observed while even weaker binding was found in the gene body (Fig. 2A). Further, for all three subsets, the average levels of Pol

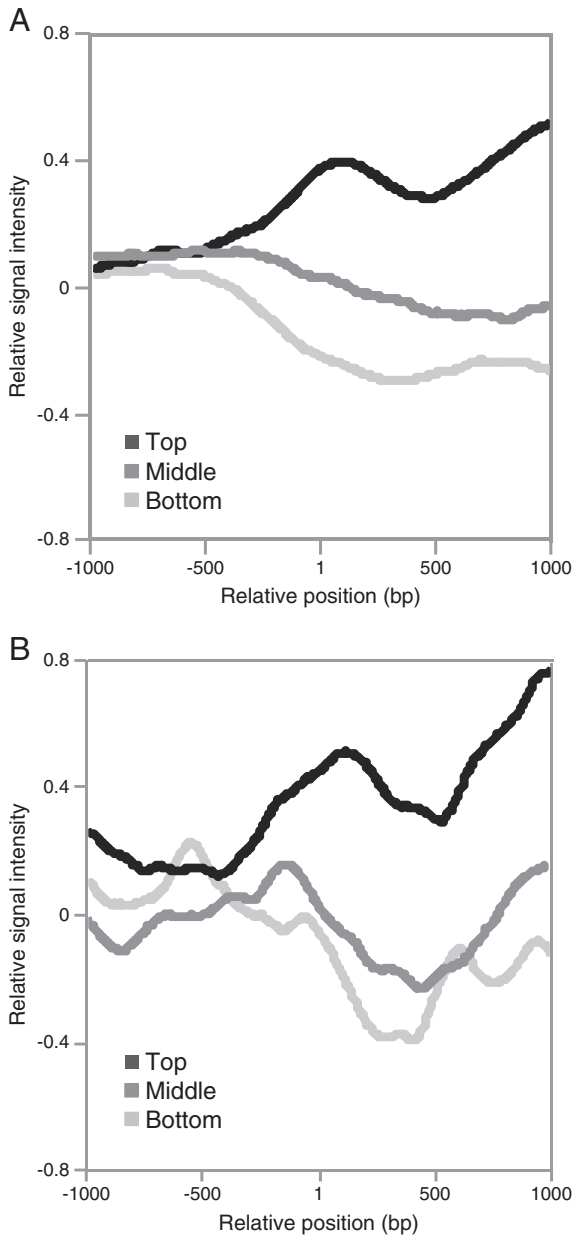

Fig. 2. Pol II binding profiles in the vicinity of TSSs of protein-coding and *MIR* genes in *Arabidopsis*. (A) A total of 2000 protein-coding genes with known TSSs were randomly selected and divided into top one-third (Top), middle one-third (Middle), and bottom one-third (Bottom) based on their ranked transcription levels. Genes in each rank were then aligned at the TSS, which is designated the +1 position. Within each rank, the average log2-transformed Pol II ChIP signal from the light sample was calculated for the −1000 to the 1000 regions and plotted against the relative position from the TSS. (B) A total of 59 *MIR* genes with known TSSs were selected and similarly analyzed.

II binding in the gene body are highly consistent with the transcription level (Fig. 2A). Such pronounced patterns suggest that the detected binding activity is primarily from Pol II in the pre-initiating and the elongating states [39].

Our next goal is to quantitatively examine the distribution of Pol II binding in the miRNA loci. To this end, we compiled 59 *MIR* genes with TSSs either validated experimentally [15] or supported by full-length cDNA collections in *Arabidopsis*. These *MIR* genes were also divided into three subsets based on ranked expression level. Plotting the Pol II binding profiles revealed similar pattern for highly expressed *MIR* genes as protein-coding genes with Pol II binding peaking at the TSSs followed by a gradually increasing profile (Fig. 2B). For the middle- and bottom-ranked *MIR* genes, despite that the Pol II pattern fluctuates more due to the small dataset, the basic features of Pol II profiles remain the same as the protein-coding genes. These include relatively higher Pol II signal in upstream regions, decreased Pol II binding in the gene body, and a correlation of expression level with Pol II binding in the gene body (Fig. 2B). These results demonstrate that the identified Pol II binding is relevant to the expression of *MIR* genes and useful to study their transcriptional regulation.

It can be observed that a "valley" locates around the 500 bp position downstream of TSSs in the Pol II binding profile for highly expressed protein-coding genes (Fig. 2A). Intriguingly, this valley is more profound for *MIR* genes regardless of the expression level (Fig. 2B). Indeed, when the 59 *MIR* genes with known TSSs were individually examined in the two biological samples, we found that 51 (86%) genes possess a Pol II binding valley around 500 bp downstream the TSSs in at least one of the samples. To test whether this pattern is robust for individual genes, we compared the Pol II binding profiles under the light and dark-transition conditions. Some representative examples are illustrated in Fig. 3. This analysis revealed that overall Pol II binding profiles for individual genes, even those in the same family, are different in terms of the peak position and shape (Fig. 3). However, the characters of Pol II binding near the TSSs observed under different growth conditions were in general conserved for the same *MIR* gene (Fig. 3). These results indicate that extensive declination of Pol II binding downstream of TSS is characteristic for *MIR* genes with known TSSs in *Arabidopsis*.

The above observation prompted us to examine all the 232 annotated *MIR* genes in *Arabidopsis*. Of these, 21 are embedded within the intron of another gene. As they are likely co-transcribed with their host gene and controlled by the host gene promoter [40], they were excluded from further analysis. Additionally, there are 20 *MIR* genes that are either too close to or overlap with other genes and 7 *MIR* genes that have poor Pol II signal. These genes were also excluded as their Pol II binding is indistinguishable from the background. For the remaining 191 *MIR* genes, we were able to identify a total of 167 (87%) displaying the characteristic Pol II binding pattern (Table S1), a proportion identical to the *MIR* genes with known TSSs. Together these results indicate that strong and unique Pol II binding is associated with a majority of the *MIR* genes transcribed as independent units.

3.3. Predicting TSSs for MIR genes based on Pol II binding pattern

To harvest further information in the Pol II binding profile, we sought to identify the promoter regions for the 167 *MIR* genes with discernable Pol II binding pattern. To this end, we developed a method to predict the TSSs for these *MIR* genes, which is motivated by the observation that 51 of the 59 *MIR* genes exhibit spatial correlation between the Pol II binding valley and the known TSSs. To utilize the Pol II binding profile for predicting TSS of individual *MIR* genes, a three-step procedure was followed. First, we used the base of the valley as the start point and set the position 500 bp upstream of the valley as an approximation for the TSS (Fig. 4A). Then, we searched within the local sequence context for TATA box like motifs based on the previous observation that TATA box is present in the core promoter of most *MIR* genes [15]. Finally, we searched approximately 25 bp downstream of the identified TATA boxes for sequences similar to the weak consensus motif around known TSSs. After each step, the prediction was refined in case these motifs were found to arrive at predicted TSSs for all 167 *MIR* genes (Table S1).

We performed two sets of analyses to evaluate the predicting power of Pol II binding profile for mapping TSSs of *MIR* genes. In the first set of analyses, we utilized the 51 *MIR* genes with known TSSs as the benchmark to determine the accuracy of the predicted TSSs. For these genes, we found that the absolute distance between the predicted TSSs and the actual TSSs is 32 bp on average. Further, we applied false discovery rate (FDR) control to the null hypothesis

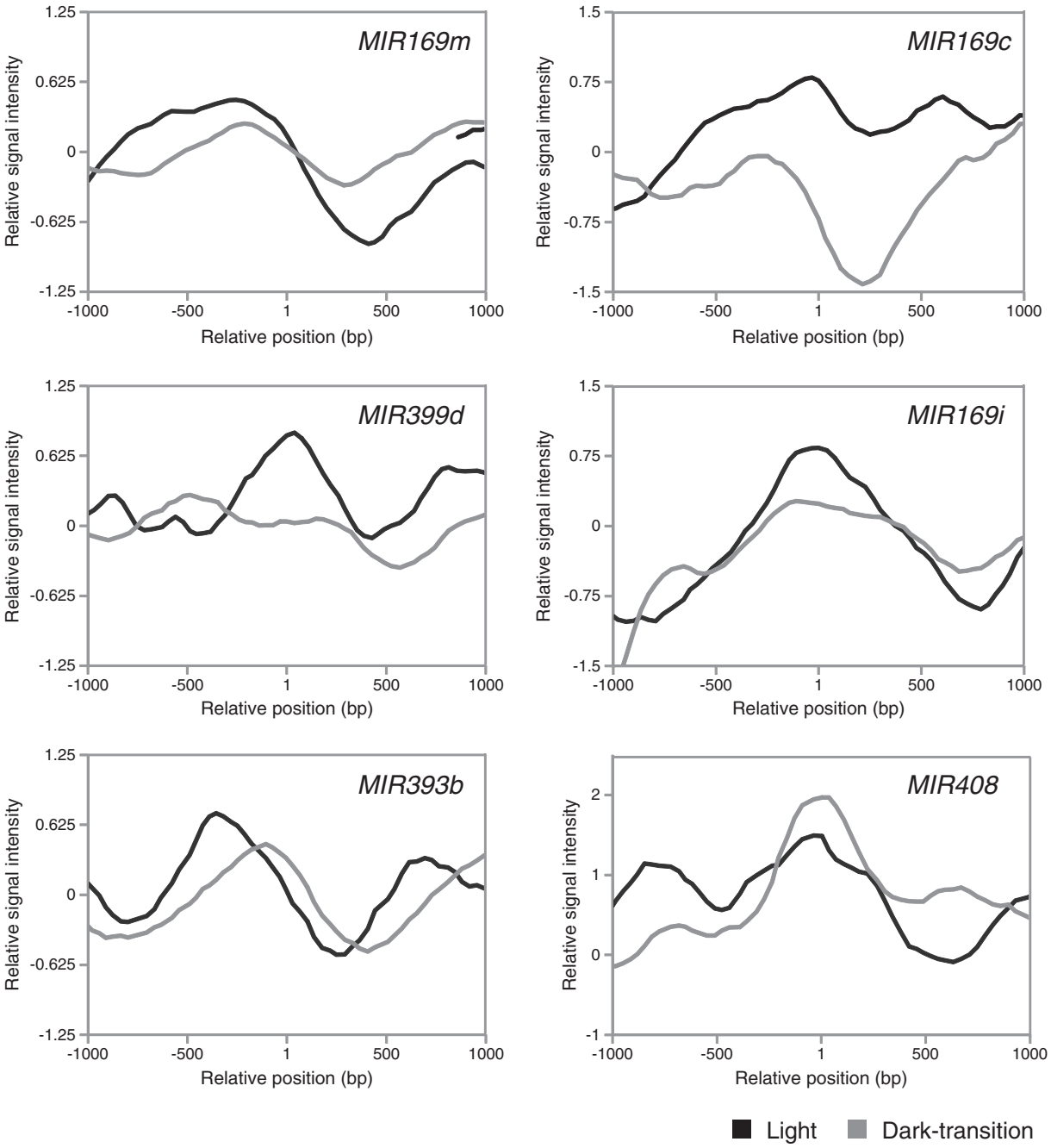

Fig. 3. Comparison of the Pol II binding profile of individual *MIR* genes in different growth conditions. Pol II ChIP signal from the light and dark-transition samples was calculated and plotted separately in the −1000 to the 1000 bp region for individual *MIR* genes with known TSSs. Shown are six representative *MIR* genes: *MIR169c*, *MIR169i*, *MIR169m*, *MIR393b*, *MIR399d*, and *MIR408*.

that the predicted TSSs are more than 200 bp away from the known TSSs and found the FDR only to be 2.0%. Next, we calculated the distance from the TSSs to the first nucleotide of the pre-miRNAs for the 59 *MIR* genes with known TSSs and genes with predicted TSSs. We found that the distribution of this measurement is essentially identical between the two groups (Fig. 4B). Thus, this set of experiments proved that the predicted TSSs are physically close to the true TSSs.

The second set of experiments aimed at examining whether the DNA structural features of the predicted TSSs are the same as the known TSSs. To this end, we generated AFE profiles on the basis of free energy change in DNA melting [36] in the vicinity of TSS for protein-coding genes and *MIR* genes. Similar to the previous report [36], we found that protein-coding genes with known TSSs show an AFE profiles with a significant difference between upstream and downstream regions and a sharp spike immediately upstream of the TSSs (Fig. 5A). Such an AFE profile is consistent with the general regulatory landscape in which the upstream promoter region is less stable while the downstream region relatively more stable. As previous reported [36], the spike found ubiquitously at approximately the −35 bp region was found to coincide with several AT-rich tetramers.

Similar to the protein-coding genes, we found a spike upstream of TSS in the AFE profiles for the 59 *MIR* genes with known TSSs (Fig. 5B). Interestingly, the decrease in the AFE profile downstream of the TSS is less profound for *MIR* genes (Fig. 5B), which is consistent with the absence of open reading frames in the *MIR* genes. For the 167

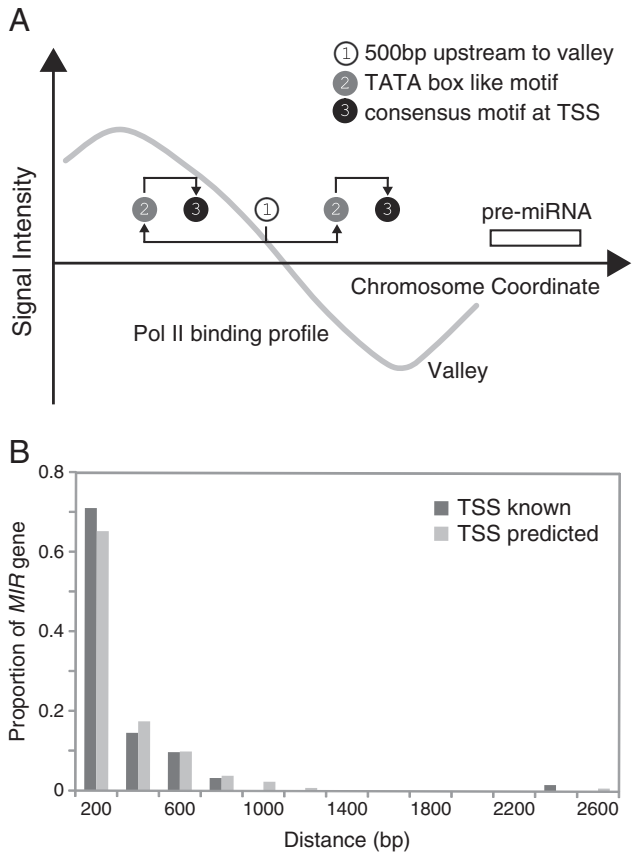

Fig. 4. Prediction of TSSs for *MIR* genes based on Pol II binding pattern. (A) TSSs were predicted from the Pol II binding profile in a three-step-procedure. In step 1, positions 500 bp upstream of the Pol II signal valley were used as approximations for the TSSs. In step 2, local sequences flanking the putative TSSs were scanned for the TATA box motifs. In step 3, sequence approximately 25 bp downstream of the identified TATA boxes were searched for the weak consensus motif found at known TSSs. After steps 2 and 3, the predicted TSSs were refined. (B) Distance measured in nucleotides between the TSSs and the first nucleotide of the pre-miRNAs was calculated. For the 59 known and 167 predicted TSSs, the proportion of *MIR* genes having a given distance were respectively calculated and plotted in 200 bp intervals.

MIR genes with predicted TSSs, we found that all features are observed including the sharp AFE spike immediately preceding the predicted TSSs (Fig. 5C). Taken together, these results attest to the effectiveness of identifying TSSs from the Pol II binding profiles and generate accurate and reliable TSSs for 167 *MIR* genes in *Arabidopsis*.

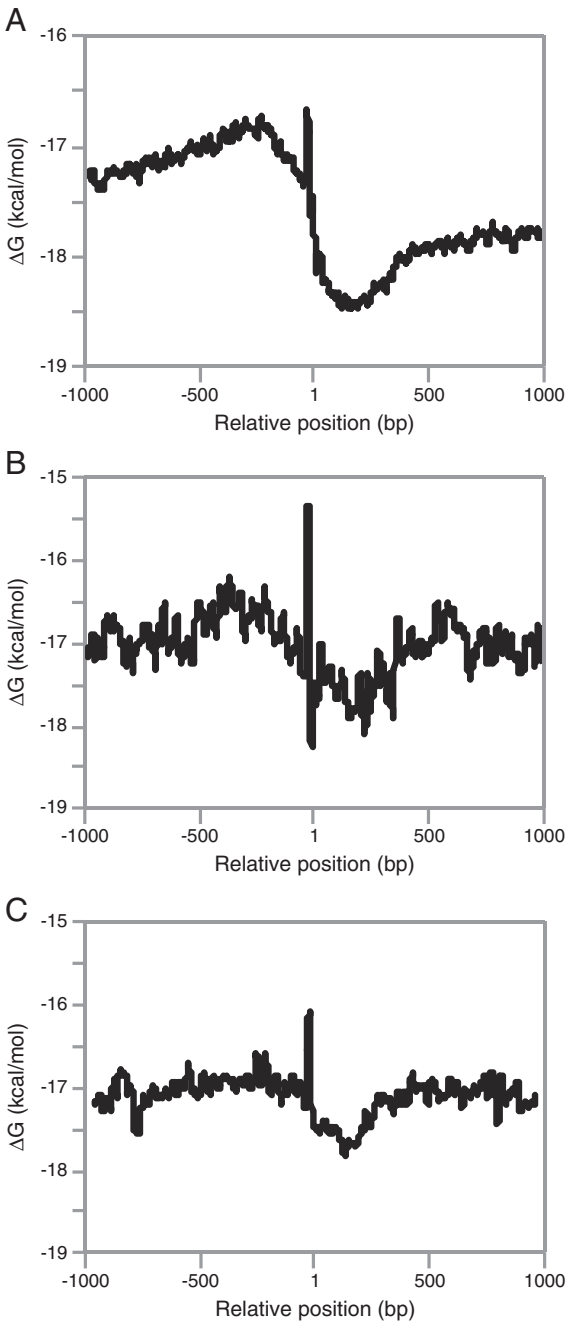

Fig. 5. AFE profiles in the vicinity of known and predicted TSSs. (A) AFE profiles for 2000 protein-coding *Arabidopsis* genes with known TSSs. For calculating the AFE, sequences in the −1000 to the 1000 bp region were aligned with the TSS set as the +1 position. To obtain an average profile, the mean value of free energy change in DNA melting based on dinucleotide parameters was calculated at each position and smoothed using a previously reported sliding window approach [36]. (B) AFE profiles for 59 *MIR* genes with experimentally validated TSSs. (C) AFE profiles for 167 *MIR* genes with predicted TSSs.

3.4. Analyzing the cis-regulatory motifs in the MIR promoters

The reliably predicted TSSs enabled us to precisely pinpoint the proximal promoter region for each of the 167 *MIR* genes. As it was shown for *MIR* genes that 90% of predicted cis-elements fall within 800 bp from the TSSs [17], we used DNA fragment corresponding to the 1 kb upstream region from the TSSs as approximation for miRNA promoters to comprehensively identify putative cis-regulatory elements. Previously, 99 position weight matrices (PWM) derived from known transcription factor binding sites were used to search 52 *MIR* promoters in *Arabidopsis* [17]. Based on posterior probability against random genomic sequences, it was reported that four cis-elements, TATA box, AtMYC2, ARF, and SORLREP3 were most enriched in *MIR* promoters [17]. Following the same PWM procedure, we analyzed all 167 promoters (Table S2). We found from the expanded dataset that three of the four motifs (except the ARF motif) were indeed over-represented in *MIR* promoters. Additionally, we found eight more cis-elements (G-box, SORLIP1, RY-repeat, LTRE, EveningElement, TELO-box, DRE-like, and AtMYB2) that also show significant enrichment in *MIR* promoters based on high posterior probability ($P(\text{PmiRNA} > \text{Prandom}) > 0.85$; Fig. 6A).

Further, we were able to identify ten under-represented cis-elements (GATA box, LFY motif, T-box, GCC-box, RAV1-B, Bellringer BS3, CArG, HSEs, Ibox and CCA1) in the *MIR* promoters that were not previously reported ($P(\text{PmiRNA} < \text{Prandom}) > 0.85$; Fig. 6B). Since the exact method was used in the current and the previous studies [17], these results demonstrate the importance of comprehensive and accurate promoter information in interpreting the cis-regulatory motifs of *MIR* genes. Interestingly, compared to protein-coding genes, approximately half

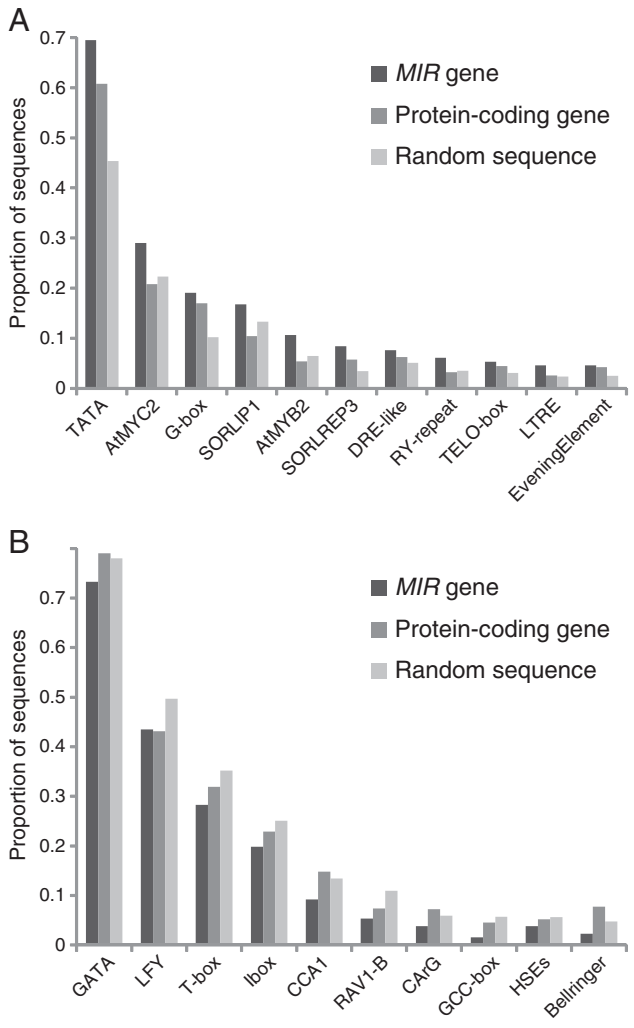

Fig. 6. Over- and under-represented cis-elements found in *MIR* promoters. (A) Upstream 1 Kb regions from the predicted TSSs were scanned for cis-elements using 99 PWM. Compared to random *Arabidopsis* genome sequences, eleven motifs with posterior probability (PmiRNA > Prandom) greater than 0.85 were considered to be over-represented in *MIR* promoters. Proportion of *MIR* promoters, protein-coding gene promoters as well as random genome loci containing these cis-elements is shown. (B) Under-represented cis-elements in *MIR* promoters. Ten motifs with posterior probability (PmiRNA < Prandom) greater than 0.85 were considered to be under-represented. Proportion of *MIR* promoters, protein-coding gene promoters as well as random genome loci containing these cis-elements is shown. See Table S2 for details.

of the 21 motifs also show significant difference in their frequency (P (Ppc > Prandom) > 0.85 or P (Ppc < Prandom) > 0.85; Table S2), suggesting that *MIR* genes may preferentially use certain cis-elements to control their expression.

4. Discussion

MIR genes are mainly transcribed by RNA Pol II [14]. The resulting primary transcript is capped at the 5′ end and polyadenylated at the 3′ end [13], similar to mRNAs. Because the abundance of pri-miRNAs ultimately determines the level of mature miRNAs present in the cell, temporal and spatial control of the transcription of individual *MIR* genes is thus critical to miRNA-based gene regulation. Mapping the genomic regions upstream of the stem−loop-structured pre-miRNAs through nucleosome positioning and Pol II ChIP analysis has been carried out for human cells [41–43]. These studies indicate that many characteristics of *MIR* promoters, including the relative frequencies of CpG islands, TATA box, TFIIB recognition, initiator elements and other chromatin signatures, are similar to those of protein-coding genes [41–44].

In our current study performed in young seedlings of *Arabidopsis*, we found that global Pol II binding pattern for *MIR* genes generally agrees with that of protein-coding genes (Figs. 1–3). However, we noticed three features in the global Pol II binding profiles that differentiate *MIR* genes from the protein-coding genes under our experimental conditions. First, though a Pol II binding peak at the TSS for highly transcribed protein-coding genes and *MIR* genes was observed, the declination of Pol II signals downstream of the TSS is more profound for *MIR* genes at all transcription levels (Fig. 2). A global Pol II signal valley was found at a position approximately 500 bp downstream of TSSs. We speculate that such a distinct pattern is generated due to the different structure of *MIR* genes compared to protein-coding genes although further experiments are required to fully explain this phenomenon. Practically speaking, the highly similar but unique Pol II binding profile for *MIR* genes allowed us to reliably predict the TSSs and hence the promoter region for 167 *MIR* genes in *Arabidopsis* (Fig. 4; Table S1).

Second, we found that the structural features of DNA in the vicinity of TSS are different for protein-coding and *MIR* genes. As shown in the AFE profiles, protein-coding genes exhibit a free energy change of about 1.5 kcal/mol when the immediate upstream and downstream regions of the TSS are compared (Fig. 5A). Such an AFE profile indicates that the promoter region is thermodynamically less stable than the 5' untranslated and coding regions [36]. However, for *MIR* genes the AFE difference upstream and downstream of the TSS is much milder (Figs. 5B, C). This finding is consistent with the fact that translation is omitted and transcription thus the primary mechanism in controlling the expression of *MIR* genes. In support of this notion, we found that *MIR* genes exhibit generally higher Pol II binding than protein-coding genes (Fig. 2). Recent studies in plants revealed that new *MIR* genes are continuously appearing in evolution [18–20]. It was argued that genetic changes resulting in beneficial miRNAs are maintained while deleterious or nonproductive changes are purged or allowed to drift [21,22]. Therefore, DNA structural features could be an important determinant in the evolution of young *MIR* genes in plants that have not yet been adequately investigated.

Third, we found that *MIR* promoters have distinctive cis-element composition. Employing 99 PWM derived from known cis-regulatory elements [17], we systematically scanned the 167 putative *MIR* promoters. We found eleven and ten cis-elements are over- and under-represented, tested against randomly sampled genomic sequences (Fig. 6). Compared to protein-coding genes, about half of the 21 motifs also show significant difference in their frequency (P (PmiRNA > P pc) > 0.85 or P (PmiRNA < P pc) > 0.85; Table S2). For example, TATA box is the most abundant motifs in *MIR* promoters and shows a high posterior probability of enrichment relative to both protein-coding and random sequences (Fig. 6A; Table S2). This is probably one of the reasons that lead to our accurate prediction of TSSs. Another conspicuous motif in *MIR* promoters is the G-box, which is implicated in environment sensing and responses, although it has a somewhat lower posterior probability relative to protein-coding sequences (Table S2).

In addition to providing testable candidates for functional studies, our analysis of the *MIR* promoters represents a new step toward reconstituting the miRNA networks. Our results demonstrate that global Pol II binding profile is a useful tool in the dissection of *MIR* promoters in *Arabidopsis*. As Pol II is highly conserved, our method should be easily applicable to other plant species. Identification and analysis of cis-regulatory elements of *MIR* genes provides important temporal and spatial measurements regarding transcription initiation, and therefore are useful to illustrate the regulatory networks in a broad range of plant species. Given the crucial roles of miRNAs in plant development and responses to environmental challenges, a comparative approach [45] should prove fruitful in identifying adaptable miRNA gene batteries and tracing their evolution to help us understand the physiological diversity and successful adaptation across plant species.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ygeno.2012.12.004.

Acknowledgments

This work was supported by a grant (DBI-0922526) from the National Science Foundation.

References

[1] R.C. Lee, R.L. Feinbaum, V. Ambros, The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14, Cell 75 (1993) 843–854.
[2] B. Wightman, I. Ha, G. Ruvkun, Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in *C. elegans*, Cell 75 (1993) 855–862.
[3] D.P. Bartel, MicroRNAs: genomics, biogenesis, mechanism, and function, Cell 116 (2004) 281–297.
[4] O. Voinnet, Origin, biogenesis, and activity of plant microRNAs, Cell 136 (2009) 669–687.
[5] X. Yang, H. Zhang, L. Li, Alternative mRNA processing increases the complexity of microRNA-based gene regulation in *Arabidopsis*, Plant J. 70 (2012) 421–431.
[6] I. Papp, M.F. Mette, W. Aufsatz, L. Daxinger, S.E. Schauer, A. Ray, et al., Evidence for nuclear processing of plant micro RNA and short interfering RNA precursors, Plant Physiol. 132 (2003) 1382–1390.
[7] A. Khvorova, A. Reynolds, S.D. Jayasena, Functional siRNAs and miRNAs exhibit strand bias, Cell 115 (2003) 209–216.
[8] D.S. Schwarz, G. Hutvagner, T. Du, Z. Xu, N. Aronin, P.D. Zamore, Asymmetry in the assembly of the RNAi enzyme complex, Cell 115 (2003) 199–208.
[9] C. Llave, Z. Xie, K.D. Kasschau, J.C. Carrington, Cleavage of Scarecrow-like mRNA targets directed by a class of *Arabidopsis* miRNA, Science 297 (2002) 2053–2056.
[10] B.J. Reinhart, E.G. Weinstein, M.W. Rhoades, B. Bartel, D.P. Bartel, MicroRNAs in plants, Genes Dev. 16 (2002) 1616–1626.
[11] P. Brodersen, L. Sakvarelidze-Achard, M. Bruun-Rasmussen, P. Dunoyer, Y.Y. Yamamoto, L. Sieburth, et al., Widespread translational inhibition by plant miRNAs and siRNAs, Science 320 (2008) 1185–1190.
[12] L. Wu, H. Zhou, Q. Zhang, J. Zhang, F. Ni, C. Liu, et al., DNA methylation mediated by a microRNA pathway, Mol. Cell 38 (2010) 465–475.
[13] X. Cai, C.H. Hagedorn, B.R. Cullen, Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs, RNA 10 (2004) 1957–1966.
[14] Y. Lee, M. Kim, J. Han, K.H. Yeom, S. Lee, S.H. Baek, et al., MicroRNA genes are transcribed by RNA polymerase II, EMBO J. 23 (2004) 4051–4060.
[15] Z. Xie, E. Allen, N. Fahlgren, A. Calamar, S.A. Givan, J.C. Carrington, Expression of Arabidopsis MIRNA genes, Plant Physiol. 138 (2005) 2145–2154.
[16] X. Zhou, J. Ruan, G. Wang, W. Zhang, Characterization and identification of microRNA core promoters in four model species, PLoS Comput. Biol. 3 (2007) e37.
[17] M. Megraw, V. Baev, V. Rusinov, S.T. Jensen, K. Kalantidis, A.G. Hatzigeorgiou, MicroRNA promoter element discovery in *Arabidopsis*, RNA 12 (2006) 1612–1619.
[18] R. Rajagopalan, H. Vaucheret, J. Trejo, D.P. Bartel, A diverse and evolutionarily fluid set of microRNAs in *Arabidopsis thaliana*, Genes Dev. 20 (2006) 3407–3425.
[19] N. Fahlgren, M.D. Howell, K.D. Kasschau, E.J. Chapman, C.M. Sullivan, J.S. Cumbie, et al., High-throughput sequencing of *Arabidopsis* microRNAs: evidence for frequent birth and death of MIRNA genes, PLoS One 2 (2007) e219.
[20] X. Yang, H. Zhang, L. Li, Global analysis of gene-level microRNA expression in *Arabidopsis* using deep sequencing data, Genomics 98 (2011) 40–46.
[21] K. Chen, N. Rajewsky, The evolution of gene regulation by transcription factors and microRNAs, Nat. Rev. Genet. 8 (2007) 93–103.
[22] M.J. Axtell, J.L. Bowman, Evolution of plant microRNAs and their targets, Trends Plant Sci. 13 (2008) 343–349.
[23] B.C. Meyers, F.F. Souret, C. Lu, P.J. Green, Sweating the small stuff: microRNA discovery in plants, Curr. Opin. Biotechnol. 17 (2006) 139–146.
[24] A. Kozomara, S. Griffiths-Jones, miRBase: integrating microRNA annotation and deep-sequencing data, Nucleic Acids Res. 39 (2011) D152–D157.
[25] A. Li, L. Mao, Evolution of plant microRNA gene families, Cell Res. 17 (2007) 212–218.
[26] C.G. Kawashima, N. Yoshimoto, A. Maruyama-Nakashita, Y.N. Tsuchiya, K. Saito, H. Takahashi, et al., Sulphur starvation induces the expression of microRNA-395 and one of its target genes but in different cell types, Plant J. 57 (2009) 313–321.
[27] C. Bowler, G. Benvenuto, P. Laflamme, D. Molino, A.V. Probst, M. Tariq, et al., Chromatin techniques for plant cells, Plant J. 39 (2004) 776–789.
[28] P. Lamesch, T.Z. Berardini, D. Li, D. Swarbreck, C. Wilks, R. Sasidharan, et al., The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools, Nucleic Acids Res. 40 (2012) D1202–D1210.
[29] X. Gan, O. Stegle, J. Behr, J.G. Steffen, P. Drewe, K.L. Hildebrand, et al., Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*, Nature 477 (2011) 419–423.
[30] D. Bielewicz, J. Dolata, A. Zielezinski, S. Alaba, B. Szarzynska, M.W. Szczesniak, et al., mirEX: a platform for comparative exploration of plant pri-miRNA expression data, Nucleic Acids Res. 40 (2012) D191–D197.
[31] H. Ji, W.H. Wong, TileMap: create chromosomal map of tiling array hybridizations, Bioinformatics 21 (2005) 3629–3636.
[32] H. Ji, H. Jiang, W. Ma, D.S. Johnson, R.M. Myers, W.H. Wong, An integrated software system for analyzing ChIP-chip and ChIP-seq data, Nat. Biotechnol. 26 (2008) 1293–1300.
[33] I.A. Shahmuradov, A.J. Gammerman, J.M. Hancock, P.M. Bramley, V.V. Solovyev, PlantProm: a database of plant promoter sequences, Nucleic Acids Res. 31 (2003) 114–117.
[34] H.T. Allawi, J. SantaLucia Jr., Thermodynamics and NMR of internal G.T mismatches in DNA, Biochemistry 36 (1997) 10581–10594.
[35] J. SantaLucia Jr., A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics, Proc. Natl. Acad. Sci. U. S. A. 95 (1998) 1460–1465.
[36] C. Morey, S. Mookherjee, G. Rajasekaran, M. Bansal, DNA free energy-based promoter prediction and comparative analysis of *Arabidopsis* and rice genomes, Plant Physiol. 156 (2011) 1300–1315.
[37] R.V. Davuluri, H. Sun, S.K. Palaniswamy, N. Matthews, C. Molina, M. Kurtz, et al., AGRIS: *Arabidopsis* gene regulatory information server, an information resource of *Arabidopsis* cis-regulatory elements and transcription factors, BMC Bioinforma. 4 (2003) 25.
[38] H. Zhang, H. He, X. Wang, X. Yang, L. Li, X.W. Deng, Genome-wide mapping of the HY5-mediated gene networks in *Arabidopsis* that involve both transcriptional and post-transcriptional regulation, Plant J. 65 (2011) 346–358.
[39] H.P. Phatnani, A.L. Greenleaf, Phosphorylation and functions of the RNA polymerase II CTD, Genes Dev. 20 (2006) 2922–2936.
[40] S. Baskerville, D.P. Bartel, Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes, RNA 11 (2005) 241–247.
[41] F. Ozsolak, L.L. Poling, Z. Wang, H. Liu, X.S. Liu, R.G. Roeder, et al., Chromatin structure analyses identify miRNA promoters, Genes Dev. 22 (2008) 3172–3183.
[42] D.L. Corcoran, K.V. Pandit, B. Gordon, A. Bhattacharjee, N. Kaminski, P.V. Benos, Features of mammalian microRNA promoters emerge from polymerase II chromatin immunoprecipitation data, PLoS One 4 (2009) e5279.
[43] G. Wang, Y. Wang, C. Shen, Y.W. Huang, K. Huang, T.H. Huang, et al., RNA polymerase II binding patterns reveal genomic regions involved in microRNA gene regulation, PLoS One 5 (2010) e13798.
[44] B.N. Davis-Dusenbery, A. Hata, Mechanisms of control of microRNA biogenesis, J. Biochem. 148 (2010) 381–392.
[45] N. Warthmann, H. Chen, S. Ossowski, D. Weigel, P. Herve, Highly specific gene silencing by artificial miRNAs in rice, PLoS One 3 (2008) e1829.