Small RNAs in angiosperms: sequence characteristics, distribution and generation

Dijun Chen1,2,†, Yijun Meng1,2,†, Xiaoxia Ma2, Chuanzao Mao2, Youhuang Bai1,3, Junjie Cao1, Haibin Gu1, Ping Wu2,3,∗ and Ming Chen1,2,5,∗

†Department of Bioinformatics, 2State Key Laboratory of Plant Physiology and Biochemistry, College of Life Sciences, Zhejiang University, Hangzhou 310058, P. R. China
2Dijun Chen, 3James D. Watson Institute of Genome Sciences, Zhejiang University, Hangzhou 310058, P. R. China

© The Author 2010. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

Here we report a large-scale analysis of HTS data for plant sRNAs. We have surveyed the GC contents, 5′-terminal compositions and chromosome-wide distribution patterns of sRNAs. De novo natural antisense transcript (NAT) prediction unveiled that the overlapping regions of trans-NATs, but not cis-NATs, were hotspots for sRNA generation. One cis-NAT generates phased natural antisense short interfering RNAs (nat-siRNAs) specifically from flowers in Arabidopsis, while one in rice produces phased nat-siRNAs from grains, suggesting their organ-specific regulatory roles.

2 METHODS
2.1 Data resources
The plant sRNA HTS data and the genomic information were retrieved from public databases. See Supplementary Tables S1 and S2 for details.

2.2 Distribution analysis of sRNAs, transposable elements (TEs) and non-TEs
Scrolling-window analysis was performed as previously described (Kasschau et al., 2007). The 200 000-nt window and 100 000-nt scroll were used. The number of sRNA or gene loci located within each window was counted to produce distribution plots. For sRNA ‘total’ loci, regardless of how repetitive the loci were, the total number of sRNA loci within each window was calculated. For ‘unique’ loci, only the sRNAs possessing unique genomic loci were taken into account. ‘Repeat-normalized’ locus analysis was done by dividing the locus count of a certain sRNA in a specific window by the total locus count of this sRNA in the genome. For non-TE or TE distribution, the number of annotated non-TEs or TE loci (annotations were retrieved from TAIR 8 and TIGR 6 for Arabidopsis and rice, respectively) in each window was calculated to generate distribution plots.

2.3 NAT prediction
Prediction was performed as previously described (Wang et al., 2006; Zhou et al., 2009). Annotated transcription units in each plant were used (see genomic resources in Supplementary Table S2).

For cis-NATs, if a pair of overlapping transcripts was located on opposite strands at adjacent genomic loci, and the overlapping region was longer than 30 nt, then they were considered as a cis-NAT pair.

For trans-NATs, BLASTN (release 2.2.20; downloaded from ftp://ftp.ncbi.nih.gov/blast/executables/release/) (Altschul et al., 1990) with default parameters was used to search for transcript pairs with high sequence complementarity, and the following criteria were used to classify the trans-NATs: based on the BLAST results, if the longest overlapping region of one transcript pair covered more than half the length of either transcript, this trans-NAT was considered to be ‘high-coverage’ (HC). If the two transcripts had a continuous complementary region longer than 100 nt, they were classified as a ‘100 nt’ pair. The predicted NATs with transcripts annotated as transposons or pseudogenes were removed, and

ABSTRACT
High-throughput sequencing (HTS) has opened up a new era for small RNA (sRNA) exploration. Using HTS data for a global survey of sRNAs in 26 angiosperms, elevated GC contents were detected in the monocots, whereas the 5′-terminal compositions were quite uniform among the angiosperms. Chromosome-wide distribution patterns of sRNAs were investigated by using scrolling-window analysis. We performed de novo natural antisense transcript (NAT) prediction, and found that the overlapping regions of trans-NATs, but not cis-NATs, were hotspots for sRNA generation. One cis-NAT generates phased natural antisense short interfering RNAs (nat-siRNAs) specifically from flowers in Arabidopsis, while one in rice produces phased nat-siRNAs from grains, suggesting their organ-specific regulatory roles.

Contact: clspwu@zju.edu.cn; mchen@zju.edu.cn
Supplementary information: Supplementary data are available at Bioinformatics online.

Received on 12 January 2010; revised on 17 March 2010; accepted on 5 April 2010

1 INTRODUCTION
In eukaryotes, small RNAs (sRNAs) are essential regulators in gene expression, heterochromatin formation and virus resistance (Carthew and Sontheimer, 2009). The plant sRNAs are divided into two classes: microRNAs (miRNAs) and siRNAs. The ~21-nt miRNAs exert post-transcriptional regulation on targets mostly through cleavage. However, the 20–24-nt siRNAs are more diverse.

2 METHODS
2.1 Data resources
The plant sRNA HTS data and the genomic information were retrieved from public databases. See Supplementary Tables S1 and S2 for details.

2.2 Distribution analysis of sRNAs, transposable elements (TEs) and non-TEs
Scrolling-window analysis was performed as previously described (Kasschau et al., 2007). The 200 000-nt window and 100 000-nt scroll were used. The number of sRNA or gene loci located within each window was counted to produce distribution plots. For sRNA ‘total’ loci, regardless of how repetitive the loci were, the total number of sRNA loci within each window was calculated. For ‘unique’ loci, only the sRNAs possessing unique genomic loci were taken into account. ‘Repeat-normalized’ locus analysis was done by dividing the locus count of a certain sRNA in a specific window by the total locus count of this sRNA in the genome. For non-TE or TE distribution, the number of annotated non-TEs or TE loci (annotations were retrieved from TAIR 8 and TIGR 6 for Arabidopsis and rice, respectively) in each window was calculated to generate distribution plots.

2.3 NAT prediction
Prediction was performed as previously described (Wang et al., 2006; Zhou et al., 2009). Annotated transcription units in each plant were used (see genomic resources in Supplementary Table S2).

For cis-NATs, if a pair of overlapping transcripts was located on opposite strands at adjacent genomic loci, and the overlapping region was longer than 30 nt, then they were considered as a cis-NAT pair.

For trans-NATs, BLASTN (release 2.2.20; downloaded from ftp://ftp.ncbi.nih.gov/blast/executables/release/) (Altschul et al., 1990) with default parameters was used to search for transcript pairs with high sequence complementarity, and the following criteria were used to classify the trans-NATs: based on the BLAST results, if the longest overlapping region of one transcript pair covered more than half the length of either transcript, this trans-NAT was considered to be ‘high-coverage’ (HC). If the two transcripts had a continuous complementary region longer than 100 nt, they were classified as a ‘100 nt’ pair. The predicted NATs with transcripts annotated as transposons or pseudogenes were removed, and

© The Author 2010. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

those containing transcripts with alternatively spliced forms were considered as distinct ones. Furthermore, DINAmeLT (Markham and Zuker, 2005) was used to verify whether the overlapping regions could melt into RNA–RNA duplexes in silico. All the BLAST-based trans-NATs were subjected to DINAmeLT hybridization validation. The trans-NAT was reserved for further analyses if it satisfied: (i) at least 80% similarity existed between DINAmeLT-based longest pairing region and BLAST-based one and (ii) the largest bubble predicted by DINAmeLT in the longest pairing region should be no longer than 10% of the region. Owing to the heavy computational work, the BLAST-based trans-NATs containing transcripts longer than 5 kb were not subjected to DINAmeLT validation. Instead, it was considered as verified trans-NAT if BLAST-based longest pairing region was longer than 10% of its longer transcript.

2.4 Enrichment of sRNA loci in NAT overlapping regions

The NATs possessing sRNA loci in the overlapping regions were analyzed. For each NAT, first, we counted the number of sRNA loci in the overlapping region (N_{oi} [1 ≤ i ≤ n]). ‘n’ represents the number of the cis- or trans-NATs analyzed) and that in the whole NAT (N_{oi}). Subsequently, the length of the overlapping region (L_{oi} = 2 × the length of the pairing sequence of either transcript) and the total length of the NAT composed by two transcripts (L_{ot}) were measured. Finally, the ratios (no. of sRNA loci per kb), A_{oi} = N_{oi}/L_{oi} and A_{ot} = N_{ot}/L_{ot}, were considered as the sRNA locus densities of the overlapping region and the whole NAT, respectively. The average density of the overlapping regions (A_{oi} = ∑_{i=1}^{n} A_{oi}/n) and that of the whole NATs (A_{ot} = ∑_{i=1}^{n} A_{ot}/n) were calculated, respectively. Besides, the sRNA enrichment score (S_{oi} = A_{oi}/A_{ot}) of the overlapping region of each NAT was calculated. Then, the average score (S = ∑_{i=1}^{n} S_{oi}/n) for all the cis- or trans-NATs was obtained.

3 RESULTS

3.1 Sequence characteristics

The HTS data of 16 eudicots and 10 monocots were retrieved from public databases (Supplementary Table S1). All the redundant sequences and those containing ambiguous nucleotide ‘N’ were removed. Considering the length range of sequences in the resulting datasets, the 18–34 nt sRNAs were analyzed. We found that the plant sRNAs were enriched in 21–24 nt, particularly in 24 nt (Fig. 1A), which was the typical length of endogenous sRNAs (Voinnet, 2009).

3.1.1 GC contents Compared to the eudicots, GC contents of the 21–24 nt sRNAs, especially the 23–24 nt ones, were elevated in the monocots (Fig. 1B). The raw sRNA sequencing data were filtered through genome-wide mapping. Thus, 16 eudicots and 9 monocots with available genomic information (Supplementary Table S2) were further analyzed. The size of filtered dataset was greatly reduced largely owing to the incomplete genomic information of most plants investigated. However, after filtering, the 21–24 nt sRNAs were still the dominant contribution, and the elevated GC contents could be observed in the 21–24 nt sRNAs in the monocots (Supplementary Fig. S1). Previous studies proposed that monocots probably possessed higher genomic GC contents than eudicots (Yu et al., 2002). Here, we show that the GC contents of the dominant sRNA portion ranging from 21 to 24 nt are higher in the monocots. Hence, we suggest that the GC contents of these sRNAs were discrepant after the eudicot–monocot divergence, and are still under selective constraint in both plant classes.

Fig. 1. Sequence characteristics. (A) The 21–24 nt sRNAs are enriched. For each plant species, the x-axis indicates the length range, and the y-axis indicates the percentage of the sRNAs in specific length. (B) GC contents are higher in the monocots. Seventeen sRNA populations (the x-axis) were analyzed. The y-axis indicates the average GC content. (C) 5′-terminal compositions. The 20–25 nt sRNAs were analyzed. For each plant species, the y-axis indicates the percentage of the sRNAs beginning with a specific nucleotide (the x-axis). The total number of sRNAs analyzed in each plant is listed on the upper right of (A).

3.1.2 5′-terminal compositions The 20–25 nt sRNAs were analyzed. Interestingly, the 20–22 nt sRNAs, and especially the 21 nt ones, begin with U predominantly, whereas the 23–25 nt sRNAs, especially the 24 nt ones, predominantly begin with A. This phenomenon is quite uniform among the angiosperms (Fig. 1C). After filtering as mentioned above, similar result was obtained (Supplementary Fig. S2). Thus, the terminal composition turning point forms between the 20–22 nt sRNAs beginning with 5′U and the 23–25 nt ones beginning with 5′A, which is quite uniform among the angiosperms. It has been known that the length and the 5′-terminals of plant sRNAs are critical factors involved in loading sRNAs into ARGONAUTE-associated RNA-induced silencing complexes (Voinnet, 2009). Based on our result, we propose that the sRNA loading mechanism is quite conserved among the angiosperms, and this mechanism was born before the eudicot–monocot split. Selection pressure on the 5′-terminals of the sRNAs of different lengths should be maintained to ensure this evolutionary conservation.

3.2 Chromosome-wide distribution

"Total" loci (see Section 2 for the definition) were highly enriched in the centromeres/pericentromeres in Arabidopsis, but much less conspicuous in rice (Fig. 2A and Supplementary Fig. S3). Consistent with previous report (Kasschau et al., 2007), the densities of
An in-depth analysis of sRNA distribution was performed in Arabidopsis and rice (Supplementary Table S3). As previously reported (Lu et al., 2005), a dominant portion of sRNA loci (∼80%) were assigned to intergenic regions. More loci were assigned to the exons and less to the introns in Arabidopsis compared to rice. However, this difference may be caused by, at least partially, the distinct gene structures, as the average length ratio between introns and exons was lower in Arabidopsis than rice.

Taken together, our results support the previous notion that a sizable portion of plant sRNAs can be mapped to TEs and intergenic regions (Llave et al., 2002; Mette et al., 2002). However, the ‘total’ sRNA loci were reported to distribute evenly along the rice chromosomes (Nobuta et al., 2007). Here, 20 times more sRNAs than previously used were included, and our more representative result showed that the ‘total’ loci were enriched in the centromeres/pericentromeres in rice, but not as conspicuous as Arabidopsis. The distribution patterns of ‘total’, ‘repeat-normalized’ and ‘unique’ sRNA loci appear to be species specific.

3.3 SRNA generation in NATs

NATs formed by complementary transcripts are divided into two classes: cis- and trans-NATs. Although several groups have predicted NATs in Arabidopsis and rice (Wang et al., 2006; Zhou et al., 2009), we carried out de novo prediction using uniform criteria (see Section 2). All the predicted NATs could serve as a repository for further analysis (Table 1 and Supplementary Material S1).

The overlapping regions of both cis- and trans-NATs were reported to be hotspots for sRNA generation in rice (Zhou et al., 2009). We extended our survey to eight plants (Table 1), and our results showed that sRNA loci were enriched in the overlapping regions of trans-NATs (paired t-test, $P < 0.0001$), except for papaya (=$P = 0.2838$) when ‘unique’ loci were calculated. However, we found a less clear-cut description of sRNA loci in cis-NATs. The sRNA loci were enriched in the overlapping regions of the cis-NATs in papaya and rice. In Arabidopsis, the ‘total’ loci were enriched (=$P < 0.0001$) but not the ‘unique’ loci (=$P = 0.0448$). Contrarily, the enrichment of the ‘unique’ loci was significant in maize ($P < 0.0001$), but not the ‘total’ loci (=$P = 0.0459$). Only three and 84 cis-NAT with sRNA loci in the overlapping regions were identified in poplar and sorghum, respectively, but no significant enrichment was detected.

Two cis-NATs with phase-distributed sRNA loci in their overlapping regions were identified in Arabidopsis, and one such cis-NAT in rice. Two distinct 24-nt phases exist in the overlapping region formed by AT2G02795 and AT2G02800 (Fig. 2C). In this cis-NAT, both the phased sRNAs called nat-siRNAs and the ones mediating first cleavage to initiate the phases have unique genomic loci. More interestingly, all these sRNAs were exclusively cloned from floral organs in Arabidopsis (Supplementary Table S4), indicating their organ-specific origin and potential involvement in plant reproduction. The rice cis-NAT was formed by LOC_Os11g10510 and LOC_Os12g10520 (Fig. 2D). Except for osa-smR1, six phased nat-siRNAs and the ones initiating the phase were exclusively cloned from rice grains (Supplementary Table S5), indicating their organ-specific origin and potential role also in reproduction. Additionally, four 24-nt phases and three 21-nt ones were found in the overlapping region of another cis-NAT in Arabidopsis (Supplementary Fig. S5). Most nat-siRNAs in this cis-NAT were also from reproductive organs, such as inflorescences, flower buds and siliques.

Fig. 2. sRNA distribution and generation. (A) Chromosome-wide distribution patterns of sRNAs, TEs and non-TEs in Arabidopsis and rice. Ovals represent the centromeres. (B) Chromosome-wide distribution patterns of sRNAs in Medicago, poplar, grapevine and sorghum. For both (A and B), the x-axis measures the chromosome length (chromosome 1 of each plant is shown), and the y-axis indicates the number of the loci in each scrolling window. Note that the maximum values plotted were capped according to the maximum values on the y-axes. ‘Total’: total sRNA loci. ‘Normalized’: normalized sRNA loci. ‘Unique’: sRNAs with unique loci. See Section 2 for their detailed definitions. (C) Phased sRNA loci in the overlapping region of the cis-NAT in Arabidopsis. Two distinct 24-nt phases are shown. (D) Phased sRNA loci in the overlapping region of the cis-NAT in rice. For both (C and D), the phased sRNAs are indicated by square brackets (sRNAs supported by HTS data are indicated by solid brackets and others by dashed ones). The sRNAs supported by HTS data were numbered, and those with unique loci were marked with asterisks. The gray shadows indicate potential cleavage sites for nat-siRNA initiation. The physical positions of the partial overlapping regions are indicated. The information of the NAT genes is shown at the bottom.

‘unique’ (see Section 2) and ‘repeat-normalized’ (see Section 2) loci were greatly reduced, and peaked in the pericentromeric regions of Arabidopsis. The reduced locus densities were also observed in rice, whereas no extensive enrichment was detected surrounding the centromeres. Instead, many isolated peaks were scattered along the rice chromosomes. In both Arabidopsis and rice, the ‘total’ locus distribution was similar to that of the TEs, but complementary to the non-TEs, indicating that sRNAs play an essential role in TE transposition control (Fig. 2A and Supplementary Fig. S3). The same analysis of sRNAs was performed in Medicago, poplar, grapevine and sorghum. Extensive sRNA enrichment was detected on the sorghum chromosomes, indicating the positions of its centromeres approximately (Fig. 2B and Supplementary Fig. S4). No conspicuous density peak of ‘repeat-normalized’ or ‘unique’ loci was observed in these four plants.
We are grateful to Dr Andrew Harrison for comments on an earlier draft of our manuscript.

Acknowledgements

We are grateful to Dr Andrew Harrison for comments on an earlier draft of our manuscript.

Funding: National Key Basic Research and Development Program of China (2005CB120901), National High Technology Research and Development Program of China (2008AA10Z125); National Natural Sciences Foundation of China (30771326, 30971437); Program for New Century Excellent Talents in University of China (NCET-07-0740).

Conflict of Interest: none declared.

References


