BIOINFORMATICS DISCOVERY NOTE
Vol. 26 no. 11 2010, pages 1391–1394
doi:10.1093/bioinformatics/btq150

Sequence analysis
Advance Access publication 8 April 2010

Small RNAs in angiosperms: sequence characteristics, distribution and generation

Dijun Chen[1,2,†], Yijun Meng[1,2,†], Xiaoxia Ma[2], Chuanzao Mao[2], Youhuang Bai[1,3], Junjie Cao[1], Haibin Gu[1], Ping Wu[2,3,*] and Ming Chen[1,2,3,*]

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

Associate Editor: Ivo Hofacker

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. Two subspecies, natural antisense short interfering RNAs (nat-siRNAs) (Borsani *et al*., 2005) and *trans*-acting siRNAs (Allen *et al*., 2005; Williams *et al*., 2005), have been characterized. Besides cleavage effect on targets, numerous siRNAs were reported to mediate DNA or histone methylation (Li *et al*., 2008; Lister *et al*., 2008). The development of sensitive high-throughput sequencing (HTS) technology has enabled our efficient sRNAs exploration (Simon *et al*., 2009), and there are a number of huge HTS datasets available (Gustafson *et al*., 2005; Johnson *et al*., 2007). A few analytical frameworks have been proposed for analyzing such data (Fahlgren *et al*., 2009), but it is clear that these datasets contain a wealth of information that has not been fully explored.

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 were hotspots for sRNA generation. We have identified novel *cis*-NATs that produce phase-distributed nat-siRNAs which are likely involved in organ-specific regulation.

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 TEs (annotations were retrieved form 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 pesudogenes were removed, and

[*]To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First authors.

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

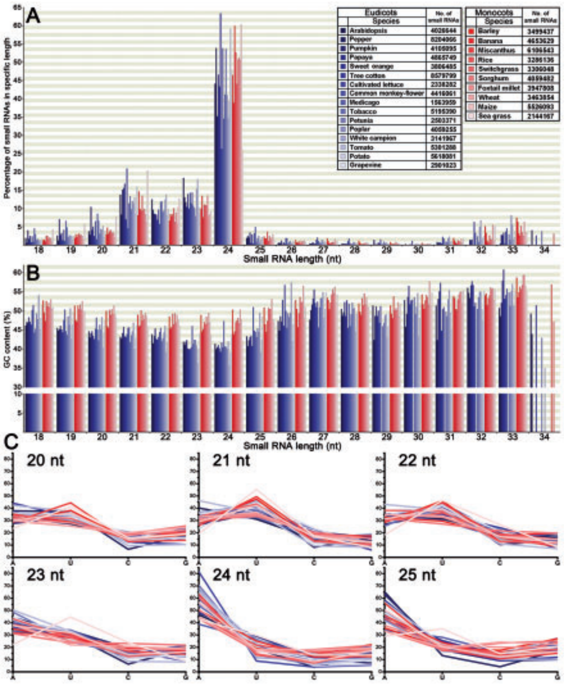
The NATs possessing sRNA loci in the overlapping regions were analyzed. For each NAT: first, we counted the number of sRNA loci in the overlapping region (N_{oi} $(1 \leq i \leq n)$, 'n' represents the number of the *cis*- or *trans*-NATs analyzed) and that in the whole NAT (N_{gi}). Subsequently, the length of the overlapping region ($L_{oi} = 2\times$ the length of the pairing sequence of either transcript) and the total length of the NAT composed by two transcripts (L_{gi}) were measured. Finally, the ratios (no. of sRNA loci per kb), $A_{oi} = N_{oi}/L_{oi}$ and $A_{gi} = N_{gi}/L_{gi}$, were considered as the sRNA locus densities of the overlapping region and the whole NAT, respectively. The average density of the overlapping regions ($A_o = \sum_{i=1}^{n} A_{oi}/n$) and that of the whole NATs ($A_g = \sum_{i=1}^{n} A_{gi}/n$) were calculated, respectively. Besides, the sRNA enrichment score ($S_i = A_{oi}/A_{gi}$) of the overlapping region of each NAT was calculated. Then, the average score ($S = \sum_{i=1}^{n} S_i/n$) 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 siRNAs (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.



Eudicots		No. of small RNAs
	Species	
■	Arabidopsis	4026644
■	Pepper	6204966
■	Pumpkin	4105895
■	Papaya	4665749
■	Sweet orange	3666485
■	Tree cotton	8579799
■	Cultivated lettuce	2338282
■	Common monkey-flower	4416061
■	Medicago	1563959
■	Tobacco	5195390
■	Petunia	2503371
■	Poplar	4059255
■	White campion	3141967
■	Tomato	5381288
■	Potato	5618081
□	Grapevine	2981823

Monocots		No. of small RNAs
	Species	
■	Barley	3499437
■	Banana	4653629
■	Miscanthus	6106543
■	Rice	3286136
■	Switchgrass	3309048
■	Sorghum	4059482
■	Foxtail millet	3947908
■	Wheat	3463854
■	Maize	5526093
□	Sea grass	2144967

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

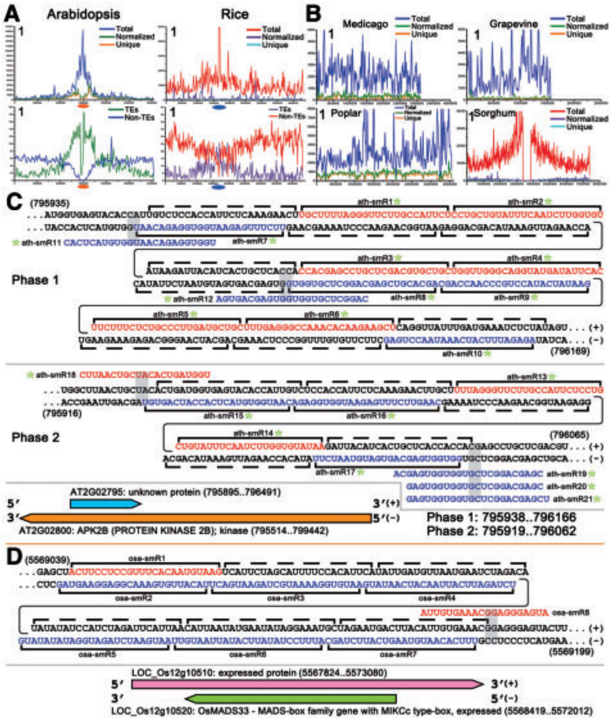


Fig. 2. sRNA distribution and generation. (**A**) Chromosome-wide distribution patterns of sRNAs, TEs and non-TEs in Arabidopsis and rice. Ovals represent the centromeres. (**B**) Chromosome-wide distribution patterns of sRNAs in Medicago, poplar, grapevine and sorghum. For both (**A** and **B**), the 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 all 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.

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 ($P = 0.0458$). Only three and 84 *cis*-NATs 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_Os12g10510* 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.

Table 1. sRNA locus density in NATs

Species	*Cis/Trans*-NATs (transcripts)[a]	*Cis/Trans* -NATs with sRNA loci in their overlapping regions (transcripts)[b] (total/unique)[c]	Overlap[d] (total/unique)[c]	All[e] (total/unique)[c]	Average score[f] (total/unique)[c]	*P*-value[g] (total/unique)[c]
Cis-NATs						
Arabidopsis	2056 (3252)	687 (1097)/623 (1000)	38.89/7.11	10.62/5.63	3.10/1.95	< 0.0001/0.0448
Poplar	10 (20)	3 (6)/2 (4)	8.42/11.19	5.42/2.68	2.61/5.26	0.4525/0.1548
Grapevine	–	–	–	–	–	–
Papaya	118 (220)	80 (151)/71 (134)	7.05/3.85	4.66/2.33	1.99/1.97	0.0094/0.0011
Medicago	–	–	–	–	–	–
Rice	1343 (1931)	634 (895)/578 (799)	3.28/1.13	4.62/0.58	1.62/2.31	0.0011/<0.0001
Maize	2437 (3599)	1455 (2215)/449 (678)	13.33/1.73	11.68/1.19	1.32/2.24	0.0458/<0.0001
Sorghum	148 (266)	84 (157)/69 (128)	8.13/3.64	8.11/2.54	1.69/2.17	0.9836/0.0727
Trans-NATs						
Arabidopsis	574 (716)	392 (475)/318 (394)	169.65/60.06	48.62/19.00	3.74/3.51	< 0.0001/ < 0.0001
Poplar	6932 (2527)	5880 (1896)/2113 (1021)	159.94/9.19	23.80/2.63	8.63/5.48	< 0.0001/ < 0.0001
Grapevine	72 572 (11 452)	69 239 (9841)/16 446 (5725)	35.25/0.74	17.87/0.47	2.39/1.95	< 0.0001/ < 0.0001
Papaya	1567 (1282)	1267 (970)/888 (698)	26.84/7.52	20.14/7.13	1.56/1.42	< 0.0001/0.2838
Medicago	92 431 (11 578)	86 153 (9771)/14 989 (5771)	61.37/5.00	28.49/1.74	3.17/4.53	< 0.0001/ < 0.0001
Rice	256 452 (18 545)	243 974 (15 631)/74 968 (12 537)	210.30/6.23	17.33/2.65	14.06/7.03	< 0.0001/ < 0.0001
Maize	152 334 (27 850)	134 127 (22 451)/20 320 (9181)	116.44/6.97	18.97/1.61	7.13/6.15	< 0.0001/ < 0.0001
Sorghum	26 268 (3851)	25 693 (3492)/5164 (2308)	344.77/5.17	64.09/2.39	10.22/3.37	< 0.0001/ < 0.0001

[a] The number of NATs and the total number of the transcripts forming the NATs in each plant.

[b] The number of the NATs with sRNA loci in the overlapping regions and the total number of the transcripts forming these NATs. Note: only these NATs were further analyzed.

[c] 'Total' and 'unique' loci (see Section 2 for their definitions) were calculated, and the results were separated by '/'.

[d] Average sRNA locus density (no. of sRNA loci per kb) in the overlapping regions ('A_o', see Section 2). [e] Average sRNA locus density (no. of sRNA loci per kb) in the whole NATs ('A_g', see Section 2).

[f] Average sRNA enrichment score ('S', see Section 2).

Zhu's group reported that nat-siRNAs generated from the overlapping region of one *cis*-NAT in Arabidopsis were involved in salt tolerance, and suggested that the *cis*-NAT defined a mode of siRNA biogenesis and function, which might be applied to other *cis*-NATs in eukaryotes (Borsani *et al*., 2005). Here, additional *cis*-NATs producing phased nat-siRNAs in their overlapping regions have been identified, and they are potentially involved in organ-specific regulation. Similar to the previous work (Borsani *et al*., 2005), all the phased nat-siRNAs and the sRNAs mediating first cleavage to initiate the phases are not miRNAs. Hence, the *cis*-NAT-defined mode of nat-siRNA generation and function can extend to the angiosperms.

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, 30971743); Program for New Century Excellent Talents in University of China (NCET-07-0740).

Conflict of Interest: none declared.

REFERENCES

Allen,E. *et al*. (2005) MicroRNA-directed phasing during trans-acting siRNA biogenesis in plants. *Cell*, **121**, 207–221.

Altschul,S.F. *et al*. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Borsani,O. *et al*. (2005) Endogenous siRNAs derived from a pair of natural cis-antisense transcripts regulate salt tolerance in Arabidopsis. *Cell*, **123**, 1279–1291.

Carthew,R.W. and Sontheimer,E.J. (2009) Origins and mechanisms of miRNAs and siRNAs. *Cell*, **136**, 642–655.

Fahlgren,N. *et al*. (2009) Computational and analytical framework for small RNA profiling by high-throughput sequencing. *RNA*, **15**, 992–1002.

Gustafson,A.M. *et al*. (2005) ASRP: the Arabidopsis small rna project database. *Nucleic Acids Res.*, **33**, D637–D640.

Johnson,C. *et al*. (2007) CSRDB: a small RNA integrated database and browser resource for cereals. *Nucleic Acids Res.*, **35**, D829–D833.

Kasschau,K.D. *et al*. (2007) Genome-wide profiling and analysis of Arabidopsis siRNAs. *PLoS Biol.*, **5**, e57.

Li,X. *et al*. (2008) High-resolution mapping of epigenetic modifications of the rice genome uncovers interplay between DNA methylation, histone methylation, and gene expression. *Plant Cell*, **20**, 259–276.

Lister,R. *et al*. (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, **133**, 523–536.

Llave,C. *et al*. (2002) Endogenous and silencing-associated small RNAs in plants. *Plant Cell*, **14**, 1605–1619.

Lu,C. *et al*. (2005) Elucidation of the small RNA component of the transcriptome. *Science*, **309**, 1567–1569.

Markham,N.R. and Zuker,M. (2005) DINAMelt web server for nucleic acid melting prediction. *Nucleic Acids Res.*, **33**, W577–W581.

Mette,M.F. *et al*. (2002) Short RNAs can identify new candidate transposable element families in Arabidopsis. *Plant Physiol.*, **130**, 6–9.

Nobuta,K. *et al*. (2007) An expression atlas of rice mRNAs and small RNAs. *Nat. Biotechnol.*, **25**, 473–477.

Simon,S.A. *et al*. (2009) Short-read sequencing technologies for transcriptional analyses. *Annu. Rev. Plant Biol.*, **60**, 305–333.

Voinnet,O. (2009) Origin, biogenesis, and activity of plant microRNAs. *Cell*, **136**, 669–687.

Wang,H. *et al*. (2006) Prediction of trans-antisense transcripts in Arabidopsis thaliana. *Genome Biol.*, **7**, R92.

Williams,L. *et al*. (2005) A database analysis method identifies an endogenous trans-acting short-interfering RNA that targets the Arabidopsis ARF2, ARF3, and ARF4 genes. *Proc. Natl Acad. Sci. USA*, **102**, 9703–9708.

Yu,J. *et al*. (2002) A draft sequence of the rice genome (Oryza sativa L. ssp. indica). *Science*, **296**, 79–92.

Zhou,X. *et al*. (2009) Genome-wide identification and analysis of small RNAs originated from natural antisense transcripts in Oryza sativa. *Genome Res.*, **19**, 70–78.