Genome-wide identification and analysis of small RNAs originated from natural antisense transcripts in Oryza sativa

Xuefeng Zhou, Ramanjulu Sunkar, Hailing Jin, et al.

Genome Res. published online October 29, 2008
Access the most recent version at doi:10.1101/gr.084806.108

Genome-wide identification and analysis of small RNAs originated from natural antisense transcripts in *Oryza sativa*

Xuefeng Zhou[1], Ramanjulu Sunkar[3], Hailing Jin[4], Jian-Kang Zhu[5] and Weixiong Zhang[1,2] *
[1]Department of Computer Science and Engineering
[2]Department of Genetics
Washington University in St. Louis
Saint Louis, MO 63130-4899
[3]Department of Biochemistry and Molecular Biology
Oklahoma State University, Stillwater, OK 74078
[4]Departments of Plant Pathology and Microbiology
Center for Plant Cell Biology and Institute for Integrative Genome Biology
[5]Department of Botany and Plant Sciences
University of California, Riverside, CA 92521

Abstract

Natural antisense transcripts (NATs) have been shown to play important roles in post-transcriptional regulation through the RNA interference pathway. We have combined the pyrophosphate-based high-throughput sequencing and computational analysis to identify and analyze, in genome-scale, cis- and trans-NAT small RNAs that are derived under a normal condition and in response to drought and salt stresses in staple plant *Oryza sativa*. Out of a total of 714,202 sequence reads from high-throughput sequencing, we obtained 58,781, 43,003 and 80,990 unique small RNAs matching perfectly to the *O. sativa* genome, from the control, drought and salt libraries, respectively. Computationally, we identified 344 cis-NATs and 7,142 trans-NATs that are formed by protein-coding genes. From the deep sequencing data, we found 108 cis-NATs and 7,141 trans-NATs that gave rise to small RNAs from their overlapping regions. Consistent with early findings, the majority of these 108 cis-NATs seem to be associated with specific conditions or developmental stages. Our analyses also revealed several interesting results. The overlapping regions of the cis- and trans-NATs appear to be more enriched with small RNA loci than non-overlapping regions. The small RNAs generated from cis- and trans-NATs have a length bias of 21-nt, even though their lengths spread over a large range. Furthermore, more than 40% of the small RNAs from cis- and trans-NATs carry an 'A' as their 5' terminal nucleotides. A substantial portion of transcripts are involved in both cis- and trans-NATs, and many trans-NATs can form many-to-many relationships, indicating that NATs may form complex regulatory networks in *O. sativa*. Our work is the first genome-wide investigation of NAT-derived small RNAs in *O. sativa*. It revealed the importance of NATs in biogenesis of small RNAs, and broadened our understanding of the roles of NAT-derived small RNAs in gene regulation, particularly in response to environmental stimuli.

*Corresponding to: Weixiong Zhang, Department of Computer Science and Engineering, Washington University in St. Louis, MO 63130; email: zhang@cse.wustl.edu; phone: (314)935-8788; fax: (314)935-7302.

1 Introduction

Post-transcriptional gene regulation at RNA level has been recently shown to be more widespread and important than previously assumed [6, 9]. While various regulatory RNA molecules have been reported in animals and plants, two prominent types of regulatory RNAs are microRNAs (miRNAs) and endogenous short interfering RNAs (siRNAs). MiRNAs can be encoded in their own genes, which are transcribed by RNA polymerase II, or exist in introns of protein-coding genes [5, 22]. Primary transcripts of miRNAs are processed to give rise to short 20- to 24-nt long mature miRNAs. SiRNA species, on the other hand, seem to be diverse. In plant, up to date, four major groups of endogenous siRNAs - natural antisense siRNA (nat-siRNA) [8, 20, 43] tasi-RNAs [2, 4, 51], heterochromatic siRNA [10, 31, 48] and long siRNAs (lsi-RNA) [24] have been reported. In animals, the Piwi-interacting RNAs [34] and nat-siRNAs [11, 16, 25, 35, 36] have been investigated. Both miRNAs and siRNAs are associated with Argonaute (AGO) proteins to form effector complexes, RNA-induced silencing complexes (RISCs). RISCs are guided by small RNAs in them to their RNA or chromatin targets based on sequence complementarity, and trigger translational repression or cleavage of target mRNAs, or modification of target chromatin [17]. Sorting specific small RNAs into distinct AGOs is very important for post-transcription gene regulation and chromatin modification. In Arabidopsis, 5' terminal nucleotide of small RNAs is the most important factor in sort them into different Argonaute complexes [30, 32].

Endogenous siRNAs can be derived from long double-stranded RNAs (dsRNAs), which are formed by transcribed repeat sequences, products of RNA-dependent RNA polymerase (RdRP) activities, and overlapped reverse complementary transcripts [39]. Natural antisense transcripts (NATs) form double-stranded RNAs in their overlapping regions, and have been reported in many model species. There are two classes of NATs, cis-NATs [18, 19, 20, 37, 44, 54, 50, 55] and trans-NATs [40, 45, 49]. Cis-NATs are formed by convergent or divergent antisense transcripts at the same genomic loci while sense and antisense transcripts of trans-NATs are derived from different loci. Endogenous siRNAs derived from NATs are referred to as natural antisense siRNAs (nat-siRNAs). Nat-siRNAs direct post-transcriptional gene silencing. In the founding example of post-transcriptional gene silencing directed by nat-siRNA, a pair of cis-NATs, SRO5 and P5CDH, were shown to have antagonistic functions in the regulation of salt tolerance in Arabidopsis [8]. P5CDH is expressed constitutively while SRO5 is induced in response to salt stress. Once SRO5 is expressed, a 24-nt-long siRNA is produced from the overlapped region of these two genes through the action of DCL2, NRPD1a, RDR6, and SGS3. This siRNA directs cleavage of the P5CDH transcripts, and the consequent terminus is used by RdRP to produce dsRNA. Subsequently, the secondary siRNAs are processed from the dsRNAs by DCL1, and these secondary siRNAs can also target P5CDH messages [8]. Later, another endogenous nat-siRNA, *nat-siRNAATGB2*, was found to be specifically induced in Arabidopsis by the bacterial pathogen [43]. Moreover, endogenous siRNAs derived from NATs have recently been reported in *Drosophila melanogaster* [11, 16, 25, 35, 36]. siRNAs typically arise from both genomic strands and are usually enriched in overlapping portions of cis-NATs [11, 35]. These recent discoveries revealed unexpected complexity of the regulatory small RNAs, and may open a window onto understanding the functions of nat-siRNAs in both plant and animal species.

Since the discovery of the founder example of cis-NATs, SRO5 and P5CDH, NATs have been believed to be an important biogenesis mechanism of endogenous siRNAs [8]. Large-scale analysis in several

model species have indicated the widespread existence of cis-NATs [15, 26, 37, 44, 50, 54, 55]. The reported frequencies of cis-NATs in different species vary because of different search criteria used and various sizes of available samples. Generally, 5% to 10% of all neighboring gene pairs in a genome of question seem to be cis-NATs. In the human genome, 4% to 9% of all transcript pairs overlap, while in the murine genome 1.7% to 14% have been reported to be overlapping [26, 44, 54]. Much effort has been devoted to discovering NATs in Arabidopsis. A transcriptome analysis with whole-genome tiling arrays has detected antisense expression of 7,600 transcripts, which are roughly 25% of all annotated genes [53]. An *in silicon* analysis using the Arabidopsis UniGene (Build 45) dataset and genome annotation data (tair5 version) predicted 1,340 potential cis-NAT pairs in Arabidopsis, and further analysis of full-length cDNAs and massively parallel signature sequencing (MPSS) data confirmed sense and antisense transcripts of 957 cis-NAT pairs [50]. Based on some qualitative criteria, these studies concluded that majority cis-NATs showed highly anticorrelated expression [50, 53]. However, there are counter examples or exceptions; some independent studies did not find any evidence of anticorrelated expression, in general, being greater than expected by chance [18, 19]. Henz *et al* found anticorrelation of cis-NATs expressions in only a small number of cis-NAT pairs [18]. Thus, the general biological functions and regulatory mechanisms of cis-NATs remain elusive. In addition to cis-NATs, 1,320 putative trans-NATs in Arabidopsis have been recently reported [49]. Interestingly, a substantial number of transcripts was predicted to form both cis- and trans-NATs, suggesting that antisense transcripts may form a complex regulatory network [49], which, nevertheless, needs to be further investigated.

If a pair of cis- or trans-NATs are co-expressed in the same cells under the same conditions, they may form double-stranded RNA duplex, which would presumably be necessary for them to spawn endogenous siRNAs. However, there is currently little data to support the co-expression of cis- or trans-NAT pairs in individual cells. The advent of high-throughput sequencing techniques made it efficient to profile, in genome scale, all small RNAs species including nat-siRNAs. Rajagopalan *et al* recently applied the high-throughput pyrosequencing to analyze small RNAs in Arabidopsis [39]. More than 340,000 unique small RNA sequences were obtained from libraries from whole seedlings, rosette leaves, whole flowers, and siliques [39]. They also noticed that some protein-coding genes in Arabidopsis had a particularly high propensity for spawning small RNAs. Eleven of the top 20 hot-spot genes of siRNAs in Arabidopsis were found to be convergently transcribed with neighboring genes. These include an antisense gene pair, At2g16580/At2g16575. Both genes have ORFs with unknown functions, with one ORF falling largely within an intron of the other [39]. With the same high-throughput sequencing technique, Kasschau *et al* profiled and analyzed small RNAs for silencing pathway mutants with defects in three RNA-dependent RNA polymerase (RDR) and four Dicer-like (DCL) genes [23]. Our recent study of cis-NATs in Arabidopsis identified 1,008 cis-NAT pairs of protein-coding genes [20]. A further analysis of small RNA datasets obtained by high-throughput sequencing techniques found that at least one gene in 646 pairs out of these 1,008 cis-NATs generates small RNAs under certain conditions or in certain development stages [20]. Moreover, high-throughput sequencing techniques have also been applied to study endogenous siRNAs in *Drosophila melanogaster* [11, 16, 25, 35, 36]. These sequencing based profiling techniques have successfully discovered miRNAs that have escaped earlier detection because they are not well conserved in related genomes or they are expressed with low abundance [7, 14, 27, 41, 42, 46].

In this study, we take advantage of two complementary approaches, computational analysis and high-

throughput sequencing, to study cis-NATs and trans-NATs in model plant species *O. sativa*. First, we perform a genome-wide computational analysis to identify cis-NATs and trans-NATs in *O. sativa*. Then, we deep-clone small RNAs in seedlings of *O. sativa* grown under normal and stress conditions, and apply a high-throughput pyrosequencing technique, 454 [28], to sequence and profile small RNA species. Finally, we identify and analyze the small RNAs derived from the identified NATs. Different from recent studies of nat-siRNAs in Arabidopsis, which only focus on cis-NATs, we consider endogenous siRNAs derived from trans-NATs as well as cis-NATs in *O. sativa*.

2 Results and Discussions

2.1 Antisense transcript pairs in *Oryza sativa*

In order to analyze small RNAs spored from NATs, we first identified cis-NATs, which are located at the same or adjacent loci. The *O. sativa* genome contains 67,450 transcription units based on the Version 5 annotation from TIGR (`www.tigr.org`). To identify potential NAT pairs, we compared the genomic loci of all annotated transcription unites to search for gene pairs that overlap in an antiparallel manner. As shown in Table 1, 530 pairs of neighboring transcription units are reverse complementarily overlapped. According to the directions of the involved transcription units, cis-NATs can be categorized in three groups, convergent (with 3'-ends overlapped), divergent (with 5'-ends overlapped) and enclosed (with one transcription unit being entirely overlapped by the other) cis-NATs. Among the 530 cis-NAT pairs, 514 (97%) are arranged in the convergent orientation, 16 (3%) are "enclosed" cis-NATs, but none of them is divergent (see Table 1). Plant genomes contain many transposable elements. In *O. sativa*, more than 24,000 transcription units annotated by TIGR are transposons. In some of cis-NATs (44, 8%), both transcription units are transposons, and in some other cis-NATs (142, 27%) one transcript is a transposon. As we expected, majority cis-NATs (344, 65%) are from protein-coding genes that are not found in any transposons. Since many transposons also code for protein-coding genes, in the rest of this paper, protein-coding genes refer to the subset of protein-coding genes not found in any transposons.

The number of cis-NATs that we identified in *O. sativa* is lower than what was previously reported for Arabidopsis. A previous study of adjacent genes in *S. cerevisiae* suggested that there may be evolutionary pressure to select against convergent genes [13]. Since *O. sativa* has a much larger genome than Arabidopsis and both genomes encode a similar amount of transcription units, we expect rice to have fewer cis-NATs. Moreover, we identified 514 convergent cis-NATs but no diverged cis-NATs. A plausible explanation for this is that transcriptional exclusion mechanisms are preferentially inhibitory to transcriptional initiation. We also note that our analysis may underestimate the number of cis-NATs in *O. sativa*, because the current rice genome annotation may still lack the extreme 5' and 3' ends for many transcripts, and even miss some transcription units.

Besides cis-NATs, which are derived from the same genomic loci, there are many trans-NATs, which are produced by transcription units from distinct genomic regions. As a major group of antisense transcripts, trans-NATs also widely exist and seem to have important functions. We performed a genome-wide screen of trans-encoded NATs in *O. sativa*. Specifically, we searched for transcript pairs sharing sequence complementarities using NCBI BLAST. Two transcripts were considered as a trans-NAT pair if their over-

Table 1: Statistics of candidate cis-NATs in *O. sativa*. (a): Chromosome. (b): Transcription units. (c): Convergent cis-NAT (with 3'-ends overlapped). (d): One transcription unit being completely reverse complementarily overlapped by the other. (e): cis-NAT pairs of transposons and protein-coding genes. (f): cis-NAT pairs of transposons. (g): cis-NAT pairs of protein-coding genes.

chrom$^{(a)}$	TU$^{(b)}$	cis-NATs	3'-3'$^{(c)}$	enclosed$^{(d)}$	transp-PC$^{(e)}$	transp-transp$^{(f)}$	PC-PC$^{(g)}$
1	8203	86	85	1	22	5	59
2	6787	65	64	1	22	6	37
3	7139	84	83	1	17	7	60
4	6292	58	56	2	11	5	42
5	5618	50	49	1	13	5	32
6	5573	32	31	1	9	1	22
7	5322	30	30	0	9	1	20
8	4938	36	32	4	14	3	19
9	4025	25	24	1	7	3	15
10	4070	26	23	3	8	2	16
11	4800	17	17	0	5	4	8
12	4683	21	20	1	5	2	14
total	67450	530	514	16	142	44	344

lapping region was longer than 100nt and could form RNA-RNA duplexes. We classified the resulting trans-NATs into three categories based on their origins. The first category contains 7,142 trans-NATs originated all from protein-coding genes, the second has 25,677 trans-NATs from the combinations of transcripts of protein-coding genes and transposons, and the third category includes 504,935 trans-NATs composed purely of transposons.

2.2 Small RNAs derived from NATs

The founder example of cis-NATs that involved in post-transcriptional gene regulation was discovered to be related to salt stress [8]. We expected more cis-NATs to be underlying various abiotic stress, and also anticipated to identify more NAT-associated siRNAs in *O. sativa* that are related to drought and salt stress. We thus performed a pyrophosphate-based high-throughput sequencing experiment to profile small RNAs in *O. sativa* under drought and salt stresses and normal condition, and obtained a total of 714,202 raw sequence reads. Specifically, 58,781, 43,003 and 80,990 unique small RNAs from control, drought and salt libraries, respectively, match perfectly to the rice genome (see Section 3.3 for details).

We searched in our three libraries for small RNAs that 100% match to the putative cis-NATs and trans-NATs discussed in Section 2.1. We found 49 cis-NATs pairs and 4,769 trans-NATs whose overlapping regions spawned small RNAs in at least one of our three libraries. In addition to our three small RNA libraries, we also searched two public datasets, MPSS [33] and CSRDB (Cereal small RNA Database) [21], for small RNAs with 100% match to *O. sativa* cis-NATs and trans-NATs. Specifically, we identified 84 cis-NATs and 5,190 trans-NATs whose overlapping regions match small RNAs in MPSS dataset, and 2

Table 2: Numbers of cis-NATs in *O. sativa* that spawn small RNAs under both conditions/stages listed in the second row and the first column. (a): Seedlings treated with ABA. (b): Nipponbare immature panicles-90 days old plants. (c): Germinating seedlings infected with Magnaporthe grisea. (d): Germinating seedlings. (e): Stem. (f): Seedling control for ABA treatment. The small RNAs were downloaded from MPSS small RNA database [33].

condition	aba$^{(a)}$	flr$^{(b)}$	snm$^{(c)}$	snu$^{(d)}$	stm$^{(e)}$	unt$^{(f)}$
aba	15	10	6	8	7	8
flr		46	15	8	8	12
snm			26	6	5	9
snu				15	5	7
stm					17	7
unt						25

cis-NATs and 2,338 trans-NATs that match small RNAs in CSRDB dataset. In total, for 108 cis-NATs and 7,141 trans-NATS, we found small RNAs in at least one of the three datasets 100% match to their overlapping regions. Supplemental Table S1 shows the numbers of unique small RNAs derived from cis-NATs and trans-NATs in different stages or under different conditions in all three datasets. In summary, 2,592 unique small RNAs match to 323 pairs of cis-NAT genes, and 56,790 unique small RNAs, match to 7,141 pairs of trans-NAT genes.

Among these 49 cis-NATs matching small RNAs in our libraries, 40 (81.6%) are specific to one of the three conditions, i.e., 11, 13 and 16 of these 49 cis-NATs exclusively match to small RNAs in the control, drought and salt libraries, respectively. In addition, 3 of the 49 cis-NATs appear in control and drought libraries, 4 match to small RNAs in control and salt libraries, 2 match to those in drought and salt libraries, but none of them appears in all three libraries. Supplemental Table S2 lists these 49 cis-NATs, the conditions under which they spawn siRNAs, and the annotation of the genes involved. Similar observations of cis-NATs were also obtained in the MPSS dataset. As shown in Table 2, most cis-NATs produce small RNAs in specific developmental stages/tissues. All these observations suggest that small RNAs derived from cis-NAT are associated with specific conditions or developmental stages. However, as shown in Supplemental Figure 1 and Supplemental Table S3, the trans-NATs related to specific conditions are relatively fewer, compared with the trans-NATs that generate small RNAs under all conditions.

To further explore whether NATs exhibit a higher likelihood to give rise to small RNAs than other protein-coding genes, we computed the density of small RNA loci in the overlapping regions, along the whole regions of the genes involved in NATs and along the whole regions of all protein-coding genes in *O. sativa* genome, respectively, in all three datasets. Here, the density is the number of small RNA loci within 1kb. As shown in Table 3, small RNAs do not seem to be enriched in cis-NATs genes. The densities of small RNAs loci in the whole regions of cis-NAT genes and in whole regions of all protein-coding genes are similar. Nevertheless, small RNAs appear to be enriched in the overlapping regions of cis-NATs. The density of small RNA loci in overlapping regions is at least three times greater than that in the whole regions of cis-NATs. This is consistent with earlier reports about derivation of small RNAs from cis-NATs [18]. We further assessed the statistical significance of the enrichment of small RNA loci

Table 3: Density of small RNA loci in different genome regions in *O. sativa*. The density is the number of small RNA loci per kb. (a): whole regions of all protein-coding genes. (b): overlapping regions that generate small RNAs. (c): whole regions of all genes involved, respectively, in cis-NATs and trans-NATs whose overlapping regions generate small RNAs. (d): enrichment score, the ratio of (b) and (c) (see Section 3.4). (e): p-value for the enrichment score (see Section 3.4). Development stages and treatment conditions of MPSS dataset are the same as in Table 2.

			cis-NAT				trans-NAT			
	condition	all PC$^{(a)}$	overlaps$^{(b)}$	total$^{(c)}$	score$^{(d)}$	p-value$^{(e)}$	overlaps$^{(b)}$	total$^{(c)}$	score$^{(d)}$	p-value$^{(e)}$
our	control	0.335	1.141	0.366	3.12	0.0004	6.989	0.782	8.94	<0.0001
libraries	drought	0.285	1.327	0.351	3.78	0.0003	4.581	0.560	8.17	<0.0001
	salt	0.294	1.231	0.333	3.7	0.0003	4.221	0.545	7.75	<0.0001
MPSS	aba	0.297	1.244	0.345	3.61	0.0006	2.162	0.426	5.07	<0.0001
	flr	0.611	2.315	0.592	3.91	0.0002	4.741	0.927	5.11	<0.0001
	snm	0.306	0.906	0.288	3.15	0.0003	2.119	0.437	4.85	<0.0001
	snu	0.309	1.039	0.282	3.69	0.0005	2.385	0.439	5.43	<0.0001
	stm	0.385	1.463	0.416	3.52	0.0003	3.034	0.582	5.21	<0.0001
	unt	0.378	1.201	0.409	2.93	0.0021	2.851	0.568	5.02	<0.0001
CSRDB		0.308	2.089	0.284	7.36	0.0084	2.562	0.416	6.16	<0.0001

in the cis-NAT overlapping regions with p values obtained by a randomization procedure (see Section 3.4). The small p values in Table 3 suggest that the enrichment of small RNA loci in the overlapping regions is statistically significant. The case for trans-NATs is more complicate. In all datasets, small RNAs are enriched in overlapping regions of trans-NATs. The density of small RNA loci in the overlapping regions is five- and six-fold greater than that in the whole regions of trans-NATs and all protein-coding genes, respectively. The enrichment of small RNA loci in trans-NAT overlapping regions is statistically significant with their p values being < 0.0001 (Table 3). However, the density of small RNA loci spans a broad range across different datasets. For example, in the small RNA library that we profiled under the normal condition as control, the density in the whole regions of trans-NAT genes is about two-fold of that in the whole regions of all protein-coding genes. But, in the other two libraries that we sequenced, the MPSS dataset and the CSRDB dataset, there is little difference in density between the whole regions of trans-NAT genes and all protein-coding genes.

2.3 Characteristics of small RNAs derived from NATs

In *D. melanogaster*, two genes on the opposite strands of cis-NATs contribute approximately the same amount of small RNAs [11, 16, 25, 35, 36]. However, in our study, we found that majority cis-NATs and trans-NATs spawn small RNAs only from one gene in a NAT pair. Specifically, among the 58 pairs of cis-NATs that produce small RNAs in our three libraries, 52 generate small RNAs only from one transcript of the pair, two cis-NATs generate at least two-fold more small RNA reads from one transcript than the other in the pair, while only 4 spawn small RNA reads roughly equally from both strands. Here, these numbers of cis-NATs are different from the numbers described in Section 2.1 since several cis-NATs give rise to small RNAs in more than one library. As shown in Table 4, in MPSS dataset, cis-NATs also produce small RNAs with a strand bias. Moreover, similar observation on trans-NATs was also obtained. In all datasets,

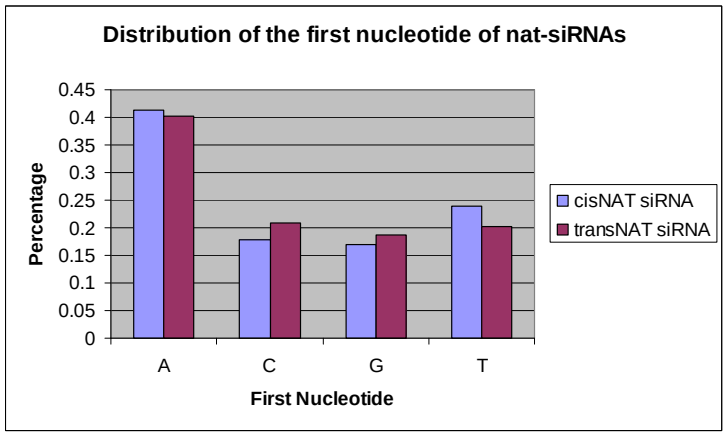


Figure 1: Distribution of the first nucleotide of small RNAs generated from ci-NATs and trans-NATs.

majority trans-NATs prefer one strand in the pair to spawn small RNAs.

It has been reported that endogenous small RNAs from cis-NATs in *D. melanogaster* have been shown to be included in AGO2 complex [25]. However, no direct evidence has been collected for which Argonaute complexes are associated with plant nat-siRNAs. In Arabidopsis sorting of small RNAs in Argonaute complexes is likely to be directed by the 5' terminal nucleotides [30, 32]. AGO2 and AGO4 preferentially recruit small RNAs with a 5' terminal adenosine [30]. However, no similar study on *O. sativa* has been reported so far. To gain further insight into endogenous small RNAs in *O. sativa*, we analyzed the lengths and 5' terminal nucleotides of small RNAs derived from cis-NATs and trans-NATs in the three libraries that we sequenced. As shown in Figure 1, more than 40% of the small RNAs from NATs have an 'A' with their 5' terminals. We further computed the natural logarithm odds ratios [1] of cis-NAT and trans-NAT small RNAs to all small RNAs that are not mapped to any NATs. Specifically, the log odds ratios of cis-NAT and trans-NAT small RNAs are 0.48 and 0.44, respectively. We also analyzed the first nucleotide frequencies in the total reads of NAT-derived small RNAs, and obtained very similar results. These NAT-derived small RNAs also have a detectable 21-nt length bias, even though their lengths spread over a wide range from 17- to 31-nt (Figure 2). It has been reported that different biogenesis mechanisms of small RNAs in Arabidopsis will generate small RNAs with different sizes [38, 52]. However, no such study on *O. sativa* has been reported.

2.4 Networks formed by NATs

It has been reported that several Arabidopsis genes are involved in two cis-NATs. One is convergent cis-NAT with overlapped 3'-ends while the other is divergent cis-NAT with overlapped 5'-ends [20, 50]. In general, one transcript in a cis-NAT pair has only one antisense partner, and we did not find any gene with more than one cis-NAT in *O. sativa*. However, 162 of 688 genes involved in the 344 cis-NATs also form trans-NATs with other genes. Moreover, one transcript may have more than one antisense transcription partner at a different location to form more than one trans-NAT. As discussed in Section 2.1, we identified 7,142 trans-NATs with overlapping regions longer than 100nt in *O. sativa*. 7,141 of them spawn small RNAs in at least one of the datasets that we studied. These trans-NATs consist of 4,753 genes, among

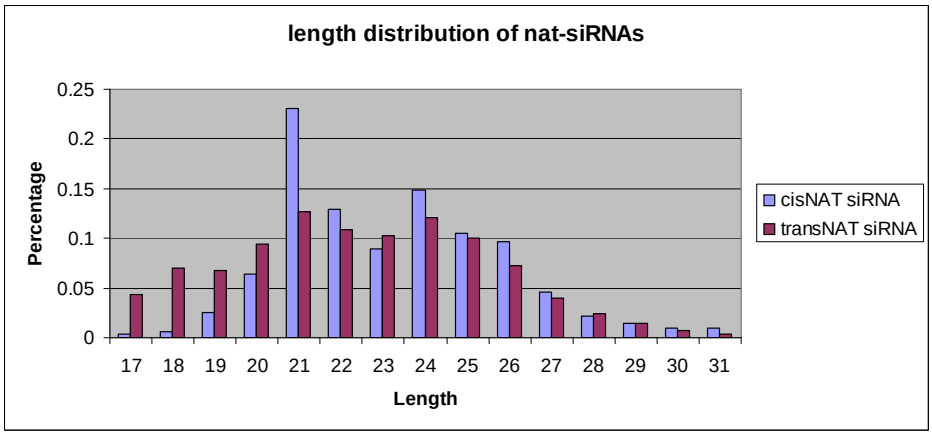


Figure 2: Distribution of the lengths of small RNAs generated from cis-NATs and trans-NATs.

Table 4: Strand bias of cis- and trans-NATs that give rise to small RNAs in *O. sativa*. (a): Only one strand of the NATs gives rise to small RNAs. (b) one strand of the NATs spawns at least two-fold more small RNAs than the other. (c): Both strands of the NATs contribute roughly equal number of small RNAs. Develop stages and treatment conditions of MPSS dataset are the same as Table 2.

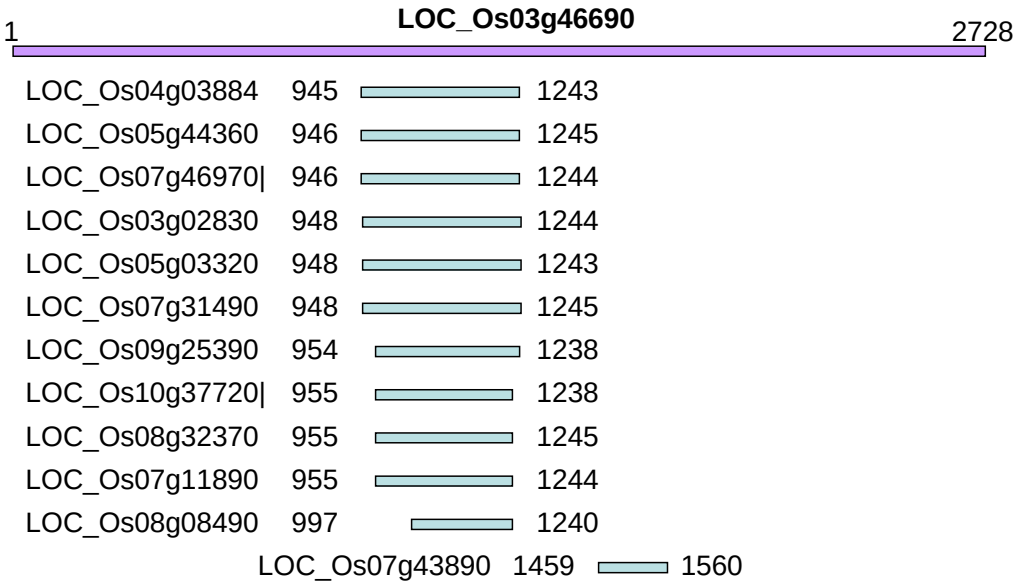
		cis-NAT			trans-NAT		
dataset	condition	one strand$^{(a)}$	bias$^{(b)}$	equal$^{(c)}$	one strand$^{(a)}$	bias$^{(b)}$	equal$^{(c)}$
our libraries	control	15	0	3	1293	2146	495
	drought	17	0	1	1710	1286	392
	salt	20	2	0	1477	1375	367
	total	52	2	4	4480	4807	1254
mpss	aba	13	1	1	1646	1018	365
	flr	35	5	6	1205	1424	175
	snm	25	0	1	1686	747	284
	snu	13	0	2	1435	835	238
	stm	15	0	2	1616	1674	405
	unt	23	1	1	1879	1625	293
	total	124	7	13	9467	7323	1760
CSRDB		1	0	1	1365	646	327

which 271 (11%) genes have at least 10 antisense transcription partners, 2,028 (37%) have more than one but fewer than 10 antisense partners, and the remaining 2,454 (51.6%) have single antisense partners. These observations indicate that antisense transcripts might form complex regulatory networks in *O. sativa*.

Theoretically, multiple paralogous genes may form RNA duplexes with the same antisense transcripts. We further analyzed the structures of possible networks formed by these 7,141 trans-NATs. The first is a star structure. Figure 3 shows an example. *LOC_Os03g46690*, encoding an F-box domain containing protein, forms trans-NATs with 12 other transcripts. The GO annotations of these 12 genes are listed in Figure 3(B). Eleven of these 12 antisense transcripts are reverse complementary to the same region of *LOC_Os03g46690* while the last one base-pairs with a different region of this gene. This example shows that a transcript may pair with more than one antisense transcript at different locations of its sequence. Another type of structure can be represented by a bipartite graph as shown in Figure 4. Trans-NATs involved in this kind of networks have many-to-many relationships. The two networks shown in Figure 4 may have function in regulation of the photosynthesis process. We did not find any clique structure formed by three genes in the trans-NAT network.

Networks formed by NATs reflect the complexity of their post-transcriptional regulation. However, under certain conditions, some genes may not be transcribed. Thus, many NAT pairs may not appear under specific conditions. As discussed in Section 2.2, we applied 454 high-throughput sequencing technique to identify small RNAs induced by salt and drought stresses. 802 and 866 trans-NAT pairs specifically spawn small RNAs under drought and salt stress conditions, respectively. We used an enrichment analysis of Gene Ontology (GO) terms [47] to functionally characterize these two sets of condition-specific trans-NATs. The 802 drought-specific trans-NAT pairs are formed by 1,056 genes. Only 429 of these 1,056 genes have function annotation. The 866 salt-specific trans-NAT pairs consist of 987 genes. Similarly, only 442 of these 987 genes have been annotated. As shown in Supplemental Figure 2, among GO terms related to biological processes, cellular process, metabolic process and response to stimulus are the three most enriched terms in both sets of genes that form drought and salt-specific trans-NATs. Binding, catalytic, transcription regulator and molecule transducer are the four most enriched GO terms that are related to molecular functions. All these enriched GO terms have p-values small than 0.001, which were obtained by a hypergeometric distribution test. If we applied the Bonferroni correction, the enrichment of these GO terms are not statistically significant. The main reason is that function annotation of *O. sativa* genes is very preliminary, the GO terms that we reported are very general, and most analyzed genes have not been annotated yet.

UDP-glucosyl transferase family proteins are important enzymes for catalyzing transportation of sugars. In Arabidopsis, it has been reported that 44 of 115 UDP-glucosyl transferase family members have one or more pairing trans-NATs, and 5 of these 44 genes also have putative cis-NATs [49]. Another 13 UDP glucosyl transferase genes have pairing cis-NATs only [49]. In *O. sativa*, 28 genes have been annotated as UDP-glucosyl transferase proteins. Our analysis show that 10 of these 28 genes form trans-NATs with their antisense partners, and 2 of them have putative cis-NATs. This implies that regulations of some biological processes through the mechanism of NATs may be conserved to certain degree.

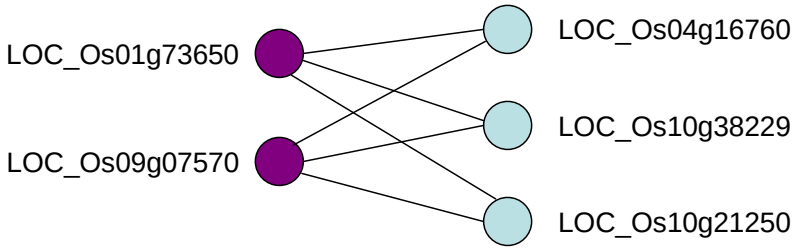


(A)

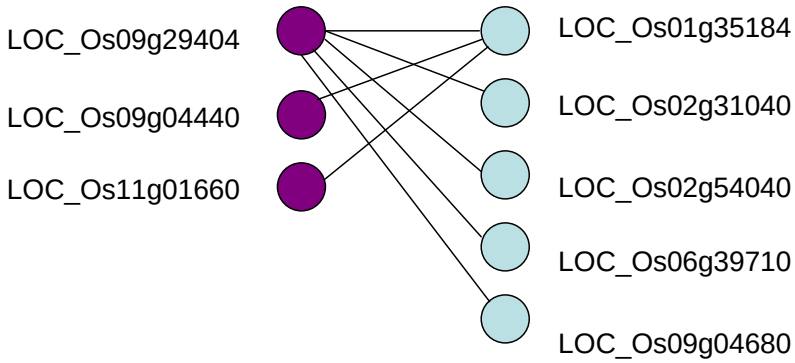
LOC_Os03g02830	cell division control protein 50, putative
LOC_Os04g03884	expressed protein
LOC_Os05g03320	expressed protein
LOC_Os05g44360	oligosaccharyl transferase STT3 subunit, putative
LOC_Os07g11890	cytochrome P450 71D7, putative
LOC_Os07g31490	expressed protein
LOC_Os07g43980	ATP-dependent RNA helicase DDX49, putative
LOC_Os07g46970	sex determination protein tasselseed-2, putative
LOC_Os08g08490	hypothetical protein
LOC_Os08g32370	mucin-associated surface protein, putative
LOC_Os09g25390	phospholipase D alpha 1, putative
LOC_Os10g37720	catalytic/ hydrolase, putative

(B)

Figure 3: One transcript may form trans-NATs with multiple antisense transcripts. The network formed by these NATs shows a star structure.



LOC_Os01g73650	hypothetical protein
LOC_Os09g07570	ferredoxin-thioredoxin reductase catalytic chain chloroplast precursor
LOC_Os04g16760	photosystem I P700 chlorophyll a apoprotein A1
LOC_Os10g21250	photosystem I P700 chlorophyll a apoprotein A1
LOC_Os10g38229	photosystem I P700 chlorophyll a apoprotein A1



LOC_Os01g35184	CBL-interacting serine/threonine-protein kinase
LOC_Os02g31040	dihydrolipoyllysine-residue succinyltransferase component of 2-oxoglutarate dehydrogenase complex, mitochondrial precursor
LOC_Os02g54040	hypothetical protein
LOC_Os06g39710	photosystem II P680 chlorophyll A apoprotein
LOC_Os09g04680	photosystem II P680 chlorophyll A apoprotein
LOC_Os09g04440	ATP binding protein
LOC_Os09g29404	glycogen operon protein glgX
LOC_Os11g01660	hypothetical protein

Figure 4: Many-to-many relationship in trans-NAT networks with bipartite structures.

3 Methods

3.1 Datasets

Genomic sequences and annotation information of the version 5 of *O. sativa* genome were downloaded from the TIGR Plant Genomics Database (`www.tigr.org`).

We included three sets of genome-matched small RNAs. The first set of small RNAs were cloned and sequenced in this study with the 454 high-throughput sequencing technique (see Section 3.3). Our small RNA sequence data have been deposited into NCBI GEO database with access number GSE12317. The second set of 11,809 sequences were downloaded from CSRDB (Cereal small RNA Database, `http://sundarlab.ucdavis.edu/smrnas/`) [21]. The third set of MPSS 320,531 17-bp signature sequences were retrieved from University of Delaware (`http://mpss.udel.edu/rice/?`) [33]. This dataset contain six libraries constructed in different development stages or with certain stress treatments.

3.2 Search for cis-NATs and trans-NATs

Putative cis-NATs were identified on the basis of *O. sativa* genome annotation. If a pair of overlapping genes were located on opposite strands at the same loci of the genome and the overlapped region was longer than 30nt, they were consider to be a cis-NAT pairs.

Trans-NATs were identified by pair-wise alignment of transcripts to search for transcript pairs with high sequence complementarity to one another. In this work, a pair of transcripts from different genomic loci were considered as a trans-NAT if they satisfied the following two criteria: they have a continuous perfect paring region longer than 100 nt, and their overlapping region can form RNA-RNA duplex. We applied a computational tool, DINAMelt [12, 29], to inspect whether the overlapping regions of a trans-NAT pair can melt into RNA-RNA duplex in silico. Some protein-coding genes contain transposons in their introns. In order to exclude the effect of intronic transposon on the identification of trans-NATs, for these genes, we performed the trans-NAT search on the coding sequences, which do not include any introns and untranslated regions.

3.3 Small RNA sequence profiling and analysis

Four-week old rice seedlings were dehydrated for 12 h (drought stress), treated with 150 mM NaCl for 24 hrs (salt stress), or grown under normal condition (control). Three corresponding small RNA libraries were constructed from these three samples as described previously [46]. Briefly, total RNA was isolated from the frozen seedlings by using the Trizol (Invitrogen, USA) according to the manufacturer's instructions. Low molecular weight RNA was enriched by NaCl and PEG precipitation. About 100 ug low molecular weight RNA was separated on a denaturing 15% polyacrylamide gel. Labeled RNA oligonucleotides with 18 and 26 nt were used as size standards. The RNAs from 18 to 26 nt were excised, purified and ligated with adaptors. Reverse transcription was preformed after ligation with adapters, followed by PCR amplification [46]. These three libraries were sequenced with the 454 high-throughput pyro-sequencing platform [28] at 454 Life Sciences. A total of 714,202 raw sequence reads from 3 independent libraries were parsed to remove 5' and 3'-adaptors. 54,016, 174,530 and 102,876 sequence reads that match TIGR

version 5.0 rice genome sequences (`www.tigr.org`) were obtained from drought, salt and control libraries, respectively. After removing sequences that could map to rRNA, tRNA and sn/snoRNA, we obtained 58,781, 43,003 and 80,990 unique small RNAs that match perfectly with the rice genome from control, salt and drought libraries, respectively. The estimated breakdown products were varied among the 3 libraries and it was in the range of 5-8%. Note that the 454 sequencing techniques may result in small insertions and a small number of mismatches. If we allowed one unmatched nucleotide, we obtained 3912, 3,334 and 6,382 more mapped reads in control, drought and salt libraries, respectively. The remaining sequences that could not mapped to the *O. sativa* genome were discarded. Possible reasons for the appearance of the unmapped sequences are as follows: these sequences might contain too many sequence errors, they might be affected by RNA editing, or be derived from the unsequenced genomic regions or splicing junction sites of cDNAs that are not characterized, or they are simply contaminants.

To provide candidate datasets of small RNAs that match protein-coding genes, we also removed small RNAs that match transposons, retrotransposons and repeats in the TIGR rice genome annotation. We then extracted five sets of small RNAs from the candidate datasets for each of the three libraries: small RNAs that match to any protein-coding genes; small RNAs mapped to the full-length sequences of any genes involved in cis-NATs and trans-NATs, respectively; small RNAs matched to the overlapping regions of cis- and trans-NATs, respectively. The last set of small RNAs were considered as cis-NAT or trans-NAT-derived small RNAs.

3.4 Enrichment of endogenous small RNAs in the overlapping regions of NATs

For each NAT (i.e. a cis-NATs or a trans-NATs), we computed the densities of small RNA loci in the overlapping region and along the whole regions of the two NAT genes as follows: we first counted the number of unique small RNAs, N_o, mapping to the overlapping region and the total number of unique small RNAs, N_g, matching the two genes. Then we measured the length of the overlapping region, L_o, and the sum of the length of the two genes, L_g. Finally, the ratios N_o/L_o and N_g/L_g were considered to be small RNA locus densities in the overlapping regions and the whole regions of the NAT genes, respectively. For each dataset described in Section 3.1 and each library constructed in this study (see Section 3.3), we computed the average densities in the overlapping regions (A_o) and along whole regions of the NAT genes (A_g) that spawn small RNAs. The ratio A_o/A_g was used as the enrichment score.

The significance of the enrichment of small RNAs in the overlapping regions of NATs was quantified by the probability that the enrichment is more than that in arbitrary regions (with the same length of the NAT overlap regions) of a set of randomly chosen gene pairs; the smaller the probability, the more statistically significant the enrichment. This probability was taken as the p-value of the enrichment. To estimate this p-value, we adopted a randomization procedure. Briefly, we first computed the enrichment score (ratio of A_o/A_g) and the average length of the overlapping regions, l, of a given set of n pairs of NATs. Secondly, we arbitrarily chose n pairs of protein-coding genes in the genome. For each pair of genes in the arbitrary set, we randomly chose a region of length l from each gene in the pair, and treated it as the "overlapping region". Then, following the steps described above, we computed the enrichment score of the arbitrary set, which constituted one sample of the randomization procedure. We collected a large number of such samples, specifically, ten thousands in our study. We then estimated the p-value by

the frequency that a sample has a bigger enrichment score than that of the given set of NATs.

3.5 Functional gene analysis

The statistical significance of enrichment of a GO term t was measured by a cumulative hypergeometric test [3]. Given M genes in a genome (e.g., *O. sativa*), assume that N of these M genes have a particular property (e.g., involved in salt-specific trans-NATs), m of the M genes in the genome contain term t in their annotations, and n of the N genes (which are involved in salt or drought-specific trans-NATs) have term t in their annotations. We calculated a p value for the statistical significance of the GO term t as the probability under which we would expect at least n genes to have t if we randomly selected m genes from the given M genes in the genome. Specifically, the p value was computed as follows:

$$P(m, n, M, N) = \sum_{n \le x \le \min\{m,N\}} \frac{C_x^N C_{m-x}^{M-N}}{C_N^M},$$

where $C_m^M = \frac{M!}{m!(M-m)!}$.

Acknowledgment

The research was supported in part by NSF grants IIS-0535257 and DBI-0743797, a grant from the Alzheimer's Association, and a grant from Monsanto Corporation to W. Zhang, and NIH grant R01GM059138 and USDA grant 2008-35100-04518 to J-K. Zhu.

References

[1] A. Agresti. *An Introduction to Categorical Data Analysis*. Wiley & Sons, 1996.

[2] E. Allen, Z. Xie, A.M. Gustafson, and J.C. Carrington. microRNA-directed phasing during trans-acting siRNA biogenesis in plants. *Cell*, 121(2):207–21, Apr 2005.

[3] D.G. Altman. *Practical Statistics for Medical Research*. Chapman & Hall/CRC, 1991.

[4] M.J. Axtell, C. Jan, R. Rajagopalan, and D.P. Bartel. A two-hit trigger for siRNA biogenesis in plants. *Cell*, 127(3):565–77, Nov 2006.

[5] D.P. Bartel. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, 116(2):281–97, Jan 2004.

[6] I. Behm-Ansmant and E. Izaurralde. Quality control of gene expression: a stepwise assembly pathway for the surveillance complex that triggers nonsense-mediated mRNA decay. *Genes Dev*, 20(4):391–8, Feb 2006.

[7] E. Berezikov, F. Thuemmler, L.W. van Laake, I. Kondova, R. Bontrop, E. Cuppen, and R.H. Plasterk. Diversity of microRNAs in human and chimpanzee brain. *Nat Genet*, 38(12):1375–7, Dec 2006.

[8] O. Borsani, J. Zhu, P.E. Verslues, R. Sunkar, and J.K. Zhu. Endogenous siRNAs derived from a pair of natural cis-antisense transcripts regulate salt tolerance in Arabidopsis. *Cell*, 123(7):1279–91, Dec 2005.

[9] P. Brodersen and O. Voinnet. The diversity of RNA silencing pathways in plants. *Trends Genet*, 22(5):268–80, May 2006.

[10] M. Buhler, W. Haas, S.P. Gygi, and D. Moazed. RNAi-dependent and -independent RNA turnover mechanisms contribute to heterochromatic gene silencing. *Cell*, 129(4):707–21, May 2007.

[11] B. Czech, C.D. Malone, R. Zhou, A. Stark, C. Schlingeheyde, M. Dus, N. Perrimon, M. Kellis, J.A. Wohlschlegel, R. Sachidanandam, G.J. Hannon, and J. Brennecke. An endogenous small interfering RNA pathway in Drosophila. *Nature*, 453(7196):798–802, Jun 2008.

[12] R.A. Dimitrov and M. Zuker. Prediction of hybridization and melting for double-stranded nucleic acids. *Biophys J*, 87(1):215–26, Jul 2004.

[13] B. Dujon. The yeast genome project: what did we learn? *Trends Genet*, 12(7):263–70, Jul 1996.

[14] N. Fahlgren, M.D. Howell, K.D. Kasschau, E.J. Chapman, C.M. Sullivan, J.S. Cumbie, S.A. Givan, T.F. Law, S.R. Grant, J.L. Dangl, and J.C. Carrington. High-throughput sequencing of Arabidopsis microRNAs: evidence for frequent birth and death of MIRNA genes. *PLoS ONE*, 2(2):e219, 2007.

[15] C.M. Farrell and L.N. Lukens. Naturally occurring antisense transcripts are present in chick embryo chondrocytes simultaneously with the down-regulation of the alpha 1 (I) collagen gene. *J Biol Chem*, 270(7):3400–8, Feb 1995.

[16] M. Ghildiyal, H. Seitz, M.D. Horwich, C. Li, T. Du, S. Lee, J. Xu, E.L. Kittler, M.L. Zapp, Z. Weng, and P.D. Zamore. Endogenous siRNAs derived from transposons and mRNAs in Drosophila somatic cells. *Science*, 320(5879):1077–81, May 2008.

[17] G.J. Hannon. RNA interference. *Nature*, 418(6894):244–51, Jul 2002.

[18] S.R. Henz, J.S. Cumbie, K.D. Kasschau, J.U. Lohmann, J.C. Carrington, D. Weigel, and M. Schmid. Distinct expression patterns of natural antisense transcripts in Arabidopsis. *Plant Physiol*, 144(3):1247–55, Jul 2007.

[19] C.H. Jen, I. Michalopoulos, D.R. Westhead, and P. Meyer. Natural antisense transcripts with coding capacity in Arabidopsis may have a regulatory role that is not linked to double-stranded RNA degradation. *Genome Biol*, 6(6):R51, 2005.

[20] H. Jin, V. Vacic, T. Girke, S. Lonardi, and J.K. Zhu. Small RNAs and the regulation of cis-natural antisense transcripts in Arabidopsis. *BMC Mol Biol*, 9:6, 2008.

[21] C. Johnson, L. Bowman, A.T. Adai, V. Vance, and V. Sundaresan. CSRDB: a small RNA integrated database and browser resource for cereals. *Nucleic Acids Res*, 35(Database issue):D829–33, Jan 2007.

[22] M.W. Jones-Rhoades, D.P. Bartel, and B. Bartel. MicroRNAS and their regulatory roles in plants. *Annu Rev Plant Biol*, 57:19–53, 2006.

[23] K.D. Kasschau, N. Fahlgren, E.J. Chapman, C.M. Sullivan, J.S. Cumbie, S.A. Givan, and J.C. Carrington. Genome-wide profiling and analysis of Arabidopsis siRNAs. *PLoS Biol*, 5(3):e57, Mar 2007.

[24] S. Katiyar-Agarwal, S. Gao, A. Vivian-Smith, and H. Jin. A novel class of bacteria-induced small RNAs in Arabidopsis. *Genes Dev*, 21(23):3123–34, Dec 2007.

[25] Y. Kawamura, K. Saito, T. Kin, Y. Ono, K. Asai, T. Sunohara, T.N. Okada, M.C. Siomi, and H. Siomi. Drosophila endogenous small RNAs bind to Argonaute 2 in somatic cells. *Nature*, 453(7196):793–7, Jun 2008.

[26] B. Lehner, G. Williams, R.D. Campbell, and C.M. Sanderson. Antisense transcripts in the human genome. *Trends Genet*, 18(2):63–5, Feb 2002.

[27] C. Lu, K. Kulkarni, F.F. Souret, R. MuthuValliappan, S.S. Tej, R.S. Poethig, I.R. Henderson, S.E. Jacobsen, W. Wang, P.J. Green, and B.C. Meyers. MicroRNAs and other small RNAs enriched in the Arabidopsis RNA-dependent RNA polymerase-2 mutant. *Genome Res*, 16(10):1276–88, Oct 2006.

[28] M. Margulies, M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L.A. Bemben, J. Berka, M.S. Braverman, Y.J. Chen, Z. Chen, S.B. Dewell, L. Du, J.M. Fierro, X.V. Gomes, B.C. Godwin, W. He, S. Helgesen, C.H. Ho, G.P. Irzyk, S.C. Jando, M.L. Alenquer, T.P. Jarvie, K.B. Jirage, J.B. Kim, J.R. Knight, J.R. Lanza, J.H. Leamon, S.M. Lefkowitz, M. Lei, J. Li, K.L. Lohman, H. Lu, V.B. Makhijani, K.E. McDade, M.P. McKenna, E.W. Myers, E. Nickerson, J.R. Nobile, R. Plant, B.P. Puc, M.T. Ronan, G.T. Roth, G.J. Sarkis, J.F. Simons, J.W. Simpson, M. Srinivasan, K.R. Tartaro, A. Tomasz, K.A. Vogt, G.A. Volkmer, S.H. Wang, Y. Wang, M.P. Weiner, P. Yu, R.F. Begley, and J.M. Rothberg. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–80, Sep 2005.

[29] N.R. Markham and M. Zuker. DINAMelt web server for nucleic acid melting prediction. *Nucleic Acids Res*, 33(Web Server issue):W577–81, Jul 2005.

[30] S. Mi, T. Cai, Y. Hu, Y. Chen, E. Hodges, F. Ni, L. Wu, S. Li, H. Zhou, C. Long, S. Chen, G.J. Hannon, and Y. Qi. Sorting of small RNAs into Arabidopsis argonaute complexes is directed by the 5' terminal nucleotide. *Cell*, 133(1):116–27, Apr 2008.

[31] D. Moazed, M. Buhler, S.M. Buker, S.U. Colmenares, E.L. Gerace, S.A. Gerber, E.J. Hong, M.R. Motamedi, A. Verdel, J. Villen, and S.P. Gygi. Studies on the mechanism of RNAi-dependent heterochromatin assembly. *Cold Spring Harb Symp Quant Biol*, 71:461–71, 2006.

[32] T.A. Montgomery, M.D. Howell, J.T. Cuperus, D. Li, J.E. Hansen, A.L. Alexander, E.J. Chapman, N. Fahlgren, E. Allen, and J.C. Carrington. Specificity of ARGONAUTE7-miR390 interaction and dual functionality in TAS3 trans-acting siRNA formation. *Cell*, 133(1):128–41, Apr 2008.

[33] K. Nobuta, R.C. Venu, C. Lu, A. Belo, K. Vemaraju, K. Kulkarni, W. Wang, M. Pillay, P.J. Green, G.L. Wang, and B.C. Meyers. An expression atlas of rice mRNAs and small RNAs. *Nat Biotechnol*, 25(4):473–7, Apr 2007.

[34] K.A. O'Donnell and J.D. Boeke. Mighty Piwis defend the germline against genome intruders. *Cell*, 129(1):37–44, Apr 2007.

[35] K. Okamura, S. Balla, R. Martin, N. Liu, and E.C. Lai. Two distinct mechanisms generate endogenous siRNAs from bidirectional transcription in Drosophila melanogaster. *Nat Struct Mol Biol*, 15(6):581–90, Jun 2008.

[36] K. Okamura, W.J. Chung, J.G. Ruby, H. Guo, D.P. Bartel, and E.C. Lai. The Drosophila hairpin RNA pathway generates endogenous short interfering RNAs. *Nature*, 453(7196):803–6, Jun 2008.

[37] N. Osato, H. Yamada, K. Satoh, H. Ooka, M. Yamamoto, K. Suzuki, J. Kawai, P. Carninci, Y. Ohtomo, K. Murakami, K. Matsubara, S. Kikuchi, and Y. Hayashizaki. Antisense transcripts with rice full-length cDNAs. *Genome Biol*, 5(1):R5, 2003.

[38] O. Pontes, C.F. Li, P.C. Nunes, J. Haag, T. Ream, A. Vitins, S.E. Jacobsen, and C.S. Pikaard. The Arabidopsis chromatin-modifying nuclear siRNA pathway involves a nucleolar RNA processing center. *Cell*, 126(1):79–92, Jul 2006.

[39] R. Rajagopalan, H. Vaucheret, J. Trejo, and D.P. Bartel. A diverse and evolutionarily fluid set of microRNAs in Arabidopsis thaliana. *Genes Dev*, 20(24):3407–25, Dec 2006.

[40] O. Rosok and M. Sioud. Systematic identification of sense-antisense transcripts in mammalian cells. *Nat Biotechnol*, 22(1):104–8, Jan 2004.

[41] J.G. Ruby, C. Jan, C. Player, M.J. Axtell, W. Lee, C. Nusbaum, H. Ge, and D.P. Bartel. Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans. *Cell*, 127(6):1193–207, Dec 2006.

[42] J.G. Ruby, A. Stark, W.K. Johnston, M. Kellis, D.P. Bartel, and E.C. Lai. Evolution, biogenesis, expression, and target predictions of a substantially expanded set of Drosophila microRNAs. *Genome Res*, 17(12):1850–64, Dec 2007.

[43] S. Katiyar-Agarwal and R. Morgan and D. Dahlbeck and O. Borsani and J.r. Villegas A and J.K. Zhu and B.J. Staskawicz and H. Jin. A pathogen-inducible endogenous sirna in plant immunity. *Proc Natl Acad Sci U S A*, 103(47):18002–7, Nov 2006.

[44] J. Shendure and G.M. Church. Computational discovery of sense-antisense transcription in the human and mouse genomes. *Genome Biol*, 3(9):RESEARCH0044, Aug 2002.

[45] S. Steigele and K. Nieselt. Open reading frames provide a rich pool of potential natural antisense transcripts in fungal genomes. *Nucleic Acids Res*, 33(16):5034–44, 2005.

[46] R. Sunkar, X. Zhou, Y. Zheng, W. Zhang, and J.K. Zhu. Identification of novel and candidate miRNAs in rice by high throughput sequencing. *BMC Plant Biol*, 8:25, 2008.

[47] The Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*, 32, 2004.

[48] T.A. Volpe, C. Kidner, I.M. Hall, G. Teng, S.I. Grewal, and R.A. Martienssen. Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *Science*, 297(5588):1833–7, Sep 2002.

[49] H. Wang, N.H. Chua, and X.J. Wang. Prediction of trans-antisense transcripts in Arabidopsis thaliana. *Genome Biol*, 7(10):R92, 2006.

[50] X.J. Wang, T. Gaasterland, and N.H. Chua. Genome-wide prediction and identification of cis-natural antisense transcripts in Arabidopsis thaliana. *Genome Biol*, 6(4):R30, 2005.

[51] L. Williams, C.C. Carles, K.S. Osmont, and J.C. Fletcher. 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 U S A*, 102(27):9703–8, Jul 2005.

[52] Z. Xie, L.K. Johansen, A.M. Gustafson, K.D. Kasschau, A.D. Lellis, D. Zilberman, S.E. Jacobsen, and J.C. Carrington. Genetic and functional diversification of small RNA pathways in plants. *PLoS Biol*, 2(5):E104, May 2004.

[53] K. Yamada, J. Lim, and et al. Empirical analysis of transcriptional activity in the Arabidopsis genome. *Science*, 302(5646):842–6, Oct 2003.

[54] R. Yelin, D. Dahary, R. Sorek, E.Y. Levanon, O. Goldstein, A. Shoshan, A. Diber, S. Biton, Y. Tamir, R. Khosravi, S. Nemzer, E. Pinner, S. Walach, J. Bernstein, K. Savitsky, and G. Rotman. Widespread occurrence of antisense transcription in the human genome. *Nat Biotechnol*, 21(4):379–86, Apr 2003.

[55] Y. Zhang, X.S. Liu, Q.R. Liu, and L. Wei. Genome-wide in silico identification and analysis of cis natural antisense transcripts (cis-NATs) in ten species. *Nucleic Acids Res*, 34(12):3465–75, 2006.