Elucidation of the Small RNA Component of the Transcriptome

Cheng Lu,[1] Shivakundan Singh Tej,[1] Shujun Luo,[4] Christian D. Haudenschild,[4] Blake C. Meyers,[1,2]* Pamela J. Green[1,2,3]*

Small RNAs play important regulatory roles in most eukaryotes, but only a small proportion of these molecules have been identified. We sequenced more than two million small RNAs from seedlings and the inflorescence of the model plant *Arabidopsis thaliana*. Known and new microRNAs (miRNAs) were among the most abundant of the nonredundant set of more than 75,000 sequences, whereas more than half represented lower abundance small interfering RNAs (siRNAs) that match repetitive sequences, intergenic regions, and genes. Individual or clusters of highly regulated small RNAs were readily observed. Targets of antisense RNA or miRNA did not appear to be preferentially associated with siRNAs. Many genomic regions previously considered featureless were found to be sites of numerous small RNAs.

Small RNAs [21 to 24 nucleotides (nt)] function to silence genes by multiple mechanisms and are present in diverse eukaryotic organisms. Among these molecules, small interfering RNAs (siRNAs) and microRNAs (miRNAs) are the two major types, and both are produced by RNase III–like enzymes called DICERs (*1*, *2*). Whereas siRNAs are processed from longer double-stranded RNA molecules and represent both strands of the RNA, miRNA molecules originate from "hairpin" precursors transcribed from one strand of distinct genomic loci. Existing methods do not sequence deeply enough to sample the full complexity of small RNAs in plant and animal systems, nor do they quantify small RNA abundances.

To investigate the complexity of small RNAs, we adapted massively parallel signature sequencing (MPSS) for these molecules (fig. S1). MPSS sequences hundreds of thousands of molecules per reaction and provides quantitative information. Briefly, small RNA molecules are isolated by size fractionation on a polyacrylamide gel; RNA adapters are sequentially ligated to the 5′ and 3′ ends; and reverse transcriptase generates the first strand of cDNA, which is amplified and used as the template for MPSS (*3*). We generated libraries using small RNA of *Arabidopsis* inflorescence or seedlings, resulting in 721,044 (67,528 distinct) and 686,124 (27,833 distinct) 17-nucleotide sequences or "signatures," respectively (Table 1A; see SOM for the second round of sequencing on seedlings in Table 1B). For the two libraries, 77% of the total distinct small RNA sequences matched the genome [the Institute for Genomic Research (TIGR) version 5.0] (*4*), representing 84% of the nearly 1.5 million total signatures (Table 1A) and exceeding by more than 10-fold the total distinct sequences from all previous *Arabidopsis* studies (*5*). The unmatched signatures may be derived from genomic gaps such as ribosomal RNA (rRNA) repeats or centromeres or may result from sequencing errors (*6*). Signatures matching to rRNAs, transfer RNAs (tRNAs), small nucleolar RNAs (snoRNAs), or small nuclear RNAs (snRNAs) made up 5.9% of the inflorescence library and 31.9% of the seedling library (table S1), lower levels compared with those previously reported (*7*, *8*). Even after removing these RNAs from consideration, the inflorescence library was proportionally more complex (Table 1A). The increased levels and diversity of small RNAs in inflorescence could reflect stronger silencing of transposons in the germline tissue, similar to that of *Caenorhabditis elegans* (*9*). Of the distinct signatures in the inflorescence and seedling libraries, 68.7 and 52.4%, respectively, matched unique sites in the genome (table S2).

We examined the distribution of small RNAs on the five *Arabidopsis* chromosomes and compared this with repeat and mRNA abundance distributions (Fig. 1, A and B; and fig. S2). The small RNAs from both libraries were highly concentrated in the pericentromeric regions of each chromosome, similar to the repeats. In contrast, mRNA levels were greatest in the euchromatic regions (fig. S2). The small RNA data for these and other specific genomic locations are best examined via the Web (*10*); the site provides detailed information about each signature that can be accessed by clicking on the corresponding triangle.

More than half of the genomic sequences matching the small RNAs in the two libraries were transposons or retrotransposons (table S3a). However, the small RNAs matching these sequences accounted for less than half the number of distinct small RNAs (Fig. 2) because more than 80% of these predicted siRNAs matched multiple locations in the genome. The corresponding small RNA signatures were predominantly found at moderate abundances (11 to 100 TPQ, transcripts per quarter million). At least half of the 11,324 retrotransposon or transposon-related sequences in the *Arabidopsis* genome had matches to small RNAs in each library, and small RNAs matched to 41% of 572 pseudogenes (table S3a).

The relative number of distinct small RNAs per megabase of sequence was lower for genes than for other genomic features (Fig. 2, table S3, a and b). About two-thirds of the genes that were matched had relatively few small RNAs (1 to 10 TPQ). These low-abundance signatures could represent perfectly matched miRNAs, or siRNAs targeted to silenced genes, unannotated pseudogenes, unannotated repeats, or other unknown sources of siRNAs (Fig. 1C). Matching the small RNAs to genes in different GO functional categories indicated that the small RNAs were well distributed among a broad range of cellular processes and molecular functions (table S4). A comparison of mRNA and small RNA MPSS data for highly expressed genes suggested that

Table 1. Summary statistics for small RNA MPSS libraries. The signatures sequenced for each library reflect the sum of two sequencing reactions. "Distinct" refers to the number of different sequences found within the set. "Total" refers to the union of the different libraries. "Genome matches" refers to distinct signatures that perfectly match to at least one location in the genome, and includes signatures matching to tRNAs, rRNAs, snRNAs or snoRNAs.

No.	Library	Signatures sequenced	Distinct signatures	Genome matches
	A. Inflorescence and seedling signatures			
1	Inflorescence	721,044	67,528	56,920
2	Seedling	686,124	27,833	17,101
Total of rows 1 and 2		1,407,168	91,445	70,633
	B. Additional signatures from a second round of sequencing from seedlings			
3	Seedling	802,978	33,640	20,379
4	Combined seedling	1,489,102	42,062	24,650
Total of all libraries		2,210,146	104,800	77,434

[1]Delaware Biotechnology Institute, [2]Department of Plant and Soil Sciences, and [3]College of Marine Studies, University of Delaware, Newark, DE 19711, USA. [4]Solexa, Inc., 25861 Industrial Boulevard, Hayward, CA 94545, USA.

*To whom correspondence should be addressed. E-mail: meyers@dbi.udel.edu (B.C.M.); green@dbi.udel.edu (P.J.G.)

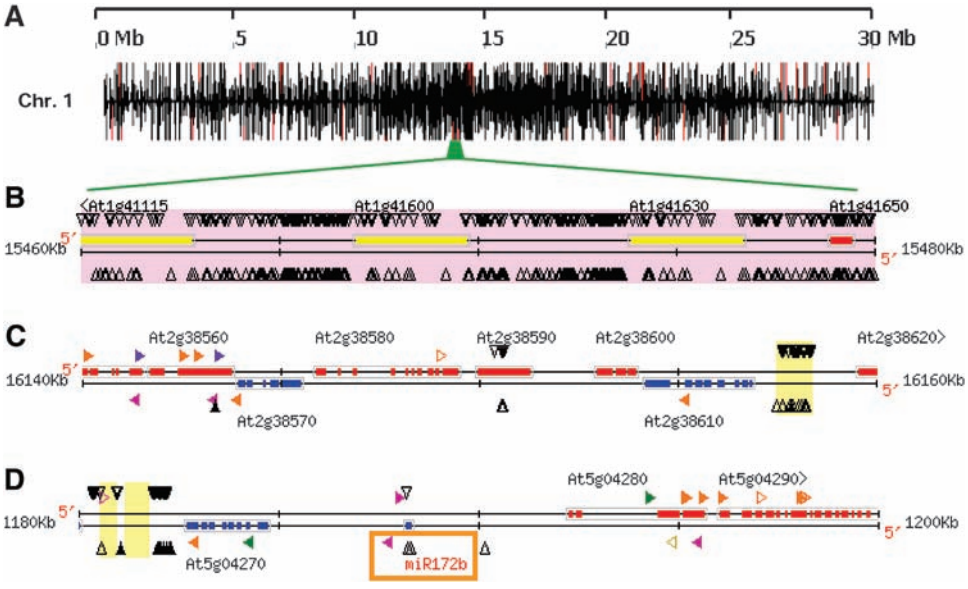


Fig. 1. Small RNAs map to numerous chromosomal locations. (**A**) Inflorescence small RNAs matched to chromosome 1. The height of the vertical lines indicates the abundance of the small RNA. Maximum height of black bar, >25 TPQ; red bar maximum >125 TPQ. (**B**) A pericentromeric region from Chr. 1. Retrotransposon-related sequences identified by RepeatMasker are highlighted in pink, and this entire region was found to be repetitive, including the spaces between annotated retrotransposons. Black triangles above or below the matching strands, small RNAs; hollow triangles, signatures mapping to more than one location; red or blue boxes, exons on top or bottom strands, respectively; colored triangles, poly(A) (MPSS from polyadenylated RNA) MPSS signatures; retrotransposons, thin yellow bars. (**C**) A typical genic region; most small RNAs map to intergenic regions which are often unannotated transposon-related sequences. Yellow shading, DNA transposon-related sequences identified by RepeatMasker. (**D**) An intergenic region of chr. 5. Orange box, small RNAs and poly(A) MPSS signatures that correspond to mir172.

the small RNA data contained only a very low level of degradation products of the longer mRNAs.

The number of distinct small RNAs that matched to intergenic regions exceeded the numbers that matched to genes, pseudogenes, transposons, or retrotransposons (Fig. 2), an observation that cannot be explained by the fraction of the genome that these entities comprise. Inflorescence small RNAs matching intergenic regions were about four times as complex as those from seedlings, and this complexity difference was evident to a slightly lesser extent for small RNAs in other categories (Fig. 2; table S3, a and b). Small RNAs in the intergenic regions potentially represent miRNAs or siRNAs from unannotated repeats such as tandem or inverted genomic repeats (*11*). We observed good correlations for tandem repeats (r = 0.5986) and inverted repeats (r = 0.4955) and small RNAs, based on comparisons of the repeat scores (representing the size and percent sequence identity) versus the total numbers of matching small RNA signatures for each repeat.

Repetitive sources of siRNAs should produce numerous small RNAs that match nearby sequences, whereas each miRNA derives from a specific sequence within the corresponding *miR* gene(s). We developed a proximity-based algorithm to build clusters of small RNAs, so that clusters with overlapping genomic locations could be compared across libraries. Moreover, the characteristics of these clusters may help differentiate novel miRNAs from siRNAs, as sparse clusters may characterize miRNAs and dense clusters may characterize siRNAs.

Genes matched by small RNAs contained an average of one sparse cluster (table S3c).

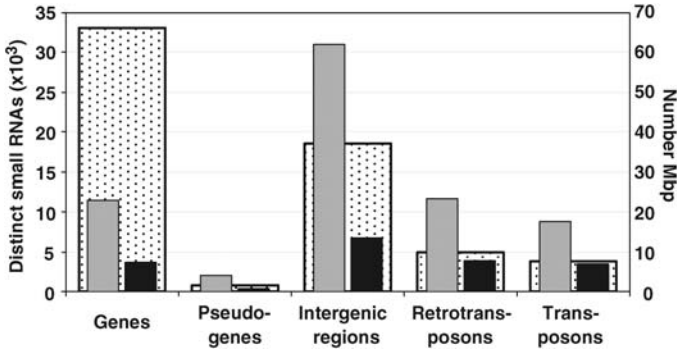
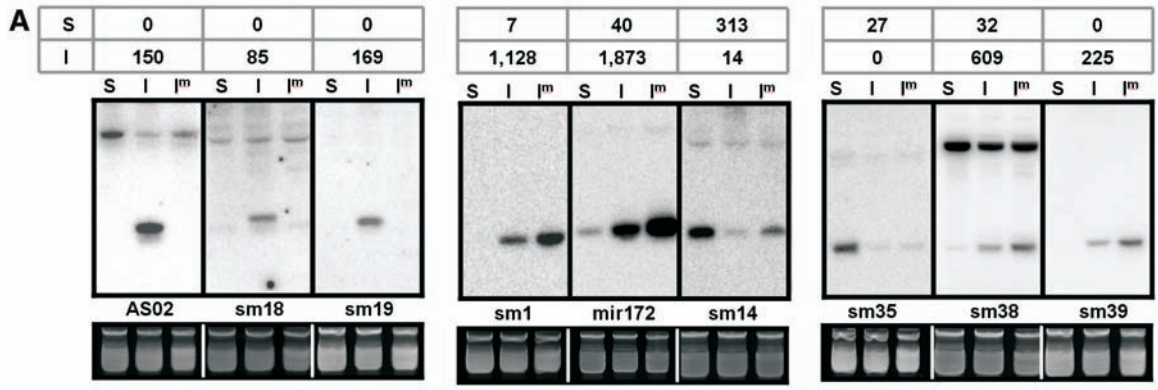


Fig. 2. Small RNAs matching classes of genomic features. Stippled bars indicate the total number of base pairs of the *Arabidopsis* genome (scale on the right) that are found in the indicated genomic features. Retrotransposon and transposon categories are from RepeatMasker. Gray vertical bars, total number of distinct small RNAs matched from the inflorescence library; black vertical bars, total number of distinct small RNAs matched from the seedling library; the scale for distinct small RNAs is on the left.

In contrast, many transposons contained more than one cluster, typically dense. In the intergenic, unannotated regions of the *Arabidopsis* genome, more than 4600 clusters of small RNAs were identified in the inflorescence library alone, which suggests a previously unrecognized activity for a large proportion of the intergenic space. A comparison of genes with and without antisense transcripts for small RNAs indicated no correlation (table S7), consistent with and extending previous arguments against small RNA involvement in general antisense control (*12*, *13*). Nevertheless, the impact of small RNAs may be far greater than this analysis of perfectly matching signatures reflects, because small RNAs are active against imperfectly matched targets (*14*, *15*), and such interactions may be numerous (table S5).

Of the 4067 genes matched by small RNAs, 693 (17%) contained small RNAs found in only one of the two libraries (table S6). Four times as many of these sequences were specific to inflorescence as to seedling (SOM file 2), which may reflect a greater variety of specialized cell types or an increased use of small RNAs in all cell types within the inflorescence. We selected representative known miRNAs or new small RNAs for validation by RNA gel-blot analysis. Figure 3A includes examples of signatures that were specific to or highly preferential for inflorescence or seedlings (signal in only one library, or >100-fold greater in one library). These include AS02, which is a known siRNA (*16*), and five new small RNAs (sm18, sm19, sm1, sm35, and sm39). Clear differential expression recapitulating the MPSS results was detected for all probes. We also examined several small RNAs that exhibited 10- to 100-fold differences in accumulation, represented in Fig. 3A by mir172, a known miRNA, sm14, and sm38. The correlation between RNA gel blots and MPSS was strong, but not always perfectly proportional, particularly for small differences or low abundances (see SOM).

Most siRNAs in *Arabidopsis* are dependent on the RNA-dependent RNA polymerase, RDR2 and are absent in an *rdr2* mutant (*16*). RNA isolated from the inflorescence of the *rdr2* mutant was also included in our



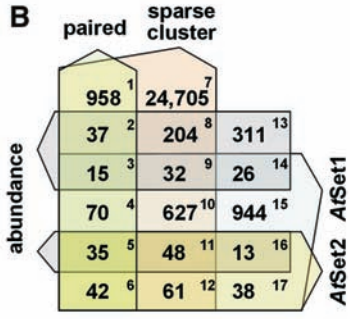


Fig. 3. Regulation and identification of new miRNA candidates. (**A**) Differential control of miRNAs and siRNA. RNA gel blots of low-molecular-weight RNA isolated from inflorescence tissues (I) and 2-week-old seedlings (S) were probed with labeled oligonucleotides. The blots also included RNA from inflorescence tissues of the *rdr2* mutant (I^m). The normalized abundance level from the MPSS data for each small RNA is listed above the blots, and ethidium bromide staining of the 5S/tRNA region of the gels is shown below. (**B**) Five-way Venn diagram of selection criteria for miRNAs. The number of distinct signatures matching the criteria is indicated in each box numbered in upper right as mentioned in the text and table S8. The figure excludes 39,622 distinct signatures that did not pass any of the criteria. See SOM and text for descriptions of the paired, sparse, and abundance filters and application of the *At*Set1 and *At*Set2 data sets (*20*). Probes for the right-hand blots in (A) are from box 3.

blots (I^m in Fig. 3A and fig. S3) to help distinguish new siRNAs and miRNAs. As expected, siRNA AS02 is lacking in this mutant as reported previously (*16*), as are sm18 and sm19 (Fig. 3A, left). The other new small RNAs in Fig. 3A and fig. S3 are not diminished in *rdr2*. These are presumably miRNAs, although it is possible that they belong to a specialized class of siRNAs dependent on another RDR such as RDR6 (*16*). Indeed, of these RDR2-independent small RNAs, several derive from regions that can form typical pre-miRNA hairpin structures (see fig. S4 for examples) and, thus, fit the requirements for annotation as new miRNAs (*17*).

Most of the miRNAs known at the time of this analysis (*18*) were found in our data set (77 of 92). Signatures exactly matching 73 miRNAs accounted for ~40% of the total abundance of genome-matched signatures from the inflorescence library, and 72 known miRNAs accounted for ~62% of seedling signatures (SOM file 1, Fig. 1D). We examined 61 known or predicted mRNA targets of *Arabidopsis* miRNAs for evidence of transitivity, the production of secondary siRNAs that match a target gene outside the sequence originally targeted. Although transitivity is common for transgene miRNA targets (*19*), most endogenous targets had no matching small RNAs other than miRNAs, or the only matching small RNAs were few, of very low abundance, or corresponded to repeats (see SOM).

To enrich for new miRNAs, we developed a set of filters that captured the majority of the 77 known miRNAs present in the small RNA MPSS data. Our abundance and sparse cluster filters captured 71 and 58, respectively, and the "paired" filter, designed to identify small RNAs near another small RNA that could be a miRNA (the complementary molecule produced from the opposite arm of the miRNA precursor), identified 39 known miRNAs (table S8). These filters were applied in combination with hairpin folding (*At*Set1) and rice conservation (*At*Set2) data sets (*20*) to generate the five-way Venn diagram in Fig. 3B. Among those that form hairpins, the sequences in box 3 were retained by all of the three filters and represent good candidates for novel miRNAs (right, Fig. 3A). None were lacking in the inflorescence of the *rdr2* mutant as expected. This is also true for representatives of box 9 retained by the sparse and abundance filters and box 2, which had paired and sparse configurations but was not identified by the *At*Set1 filter (fig. S3b). The absence of these sequences in *At*Set2 indicates that filters based on MPSS data can enhance miRNA prediction capability even when cross-species conservation is lacking.

Our data indicate that the small RNA component of the genome and its regulatory role is more extensive and complex than previously demonstrated. Many regions of the genome considered inactive or featureless were found in our analyses to be sites of considerable small RNA activity. Insight into the functional basis for this complexity will result from detailed analyses of the *Arabidopsis* small RNAs and application of this approach in diverse treatments, small RNA mutants, and other species.

References and Notes

1. E. Bernstein, A. A. Caudy, S. M. Hammond, G. J. Hannon, *Nature* **409**, 363 (2001).
2. A. Grishok *et al.*, *Cell* **106**, 23 (2001).
3. S. Brenner *et al.*, *Nat. Biotechnol.* **18**, 630 (2000).
4. J. R. Wortman *et al.*, *Plant Physiol.* **132**, 461 (2003).
5. A. M. Gustafson *et al.*, *Nucleic Acids Res.* **33**, D637 (2005).
6. B. C. Meyers *et al.*, *Genome Res.* **14**, 1641 (2004).
7. W. Park, J. Li, R. Song, J. Messing, X. Chen, *Curr. Biol.* **12**, 1484 (2002).
8. R. Sunkar, J. K. Zhu, *Plant Cell* **16**, 2001 (2004).
9. T. Sijen, R. H. Plasterk, *Nature* **426**, 310 (2003).
10. http://mpss.udel.edu/at
11. R. A. Martienssen, *Nat. Genet.* **35**, 213 (2003).
12. C. H. Jen, I. Michalopoulos, D. R. Westhead, P. Meyer, *Genome Biol.* **6**, R51 (2005).
13. X. J. Wang, T. Gaasterland, N. H. Chua, *Genome Biol.* **6**, R30 (2005).
14. A. L. Jackson, P. S. Linsley, *Trends Genet.* **20**, 521 (2004).
15. L. P. Lim *et al.*, *Nature* **433**, 769 (2005).
16. Z. Xie *et al.*, *PLoS Biol.* **2**, E104 (2004).
17. V. Ambros *et al.*, *RNA* **9**, 277 (2003).
18. S. Griffiths-Jones, *Nucleic Acids Res.* **32**, D109 (2004).
19. E. A. Parizotto, P. Dunoyer, N. Rahm, C. Himber, O. Voinnet, *Genes Dev.* **18**, 2237 (2004).
20. M. W. Jones-Rhoades, D. P. Bartel, *Mol. Cell* **14**, 787 (2004).
21. We are grateful to M. Nakano and D. Lee for the web interface, J. Carrington for the *rdr2* mutant, F. Souret and B.-C. Yoo for comments on the manuscript, and S. Jacobsen for helpful discussions. This work was supported primarily by NSF SGER #0439186, with additional support from the NSF Plant Genome Program (B.C.M.), DOE DE-FG02-04ER15541 (P.J.G.), and NIH P20 RR16472-04.

Supporting Online Material

www.sciencemag.org/cgi/content/full/309/5740/1567/DC1
SOM Text
Materials and Methods
Figs. S1 to S9
Tables S1 to S4
References and Notes
Data files 1 and 2

7 April 2005; accepted 28 July 2005
10.1126/science.1114112