Supplemental Experimental Procedures

Data sets

The genomic sequence of *Arabidopsis thaliana* was downloaded from NCBI (August 13, 2001 release)(Arabidopsis Genome Initiative, 2000), and that of *Oryza sativa* L. ssp. *indica* was downloaded from the Beijing Rice Genomics Institute (1st draft)(Yu et al., 2002). Repetitive elements in genomic sequence were identified using RepeatMasker (Smit, A.F.A. and Green, P., http://www.repeatmasker.org/) with the default repeat library for *Arabidopsis* and the version 2 *Oryza* repeats from TIGR (http://www.tigr.org/) for *Oryza*. Sequences of *Arabidopsis* mRNA transcripts were downloaded from TAIR (http://www.arabidopsis.org/,4/17/2003 release). For *Oryza* mRNAs, version 12 of the Rice Gene Index (TIGR) was used. Homologous *Oryza* mRNAs were defined for each *Arabidopsis* mRNA as the *Oryza* mRNAs with the top 5 tblastn (Gish, W. (1996-2004) http://blast.wustl.edu) scores to the protein encoded by the *Arabidopsis* mRNA, with a maximum E value of 10^{-30}.

MIRcheck

MIRcheck is a script that compares a 20mer within a predicted secondary structure against a set of parameters that describe the majority of previously known plant miRNAs and their hairpins. To pass MIRcheck, a 20mer must meet all the following requirements.

1) All predicted base pairs must be in the same direction (i.e. all nucleotides pairing to the 20mer must either all be 5_ of the 20mer or all be 3_ of the 20mer). 2) No more than 4 of the 20 nt may be unpaired, with no more than 2 adjacent unpaired nt. 3) The length of the hairpin, measured as the number of nucleotides containing the 20mer, the loop, and

the nucleotides predicted to pair to the 20mer (including an equal number of terminal unpaired nt, if any) must be at least 60. 4) No more than 1 of the 20nt may be asymmetrically unpaired. 5) The pairing must extend 4nt beyond the 20mer. This means that at least 1 24mer containing the 20mer must meet the above requirements for directionality, number, and pattern of base pairing. In addition, the 24mer* (here defined as the sequence base pairing to the 24mer, including an equal number of terminal unpaired nt, if any), must not by longer than 27 nt, with no more than 5 nt unpaired in total and no more than 3 adjacent unpaired nt. Finally, at least 1 nt in either the 24mer or 24mer* must unpaired.

Identification of conserved 20mers in miRNA-like foldbacks

Imperfect inverted repeats were found in the unmasked *Arabidopsis* and *Oryza* genomes using EINVERTED (EMBOSS, http://www.hgmp.mrc.ac.uk/Software/EMBOSS/index.html), with the parameters threshold = 40, match score = 3, mismatch score = 3, gap penalty = 40, and maximum repeat length = 240. Each inverted repeats was extended by 10 nt on each side, the secondary structure of the inverted repeat predicted by RNAfold (Hofacker et al., 1994) and each 20mer in the inverted repeats was evaluated by MIRcheck. 20mers that overlapped with repetitive elements were removed from the analysis, as were 20mers for which 12/20 nt were any one nucleotide or 17/20 were any pair of nucleotides. Patscan (Dsouza et al., 1997) was used to find near matches (0-2 nt changes) between *Arabidopsis* and *Oryza* 20mers that passed MIRcheck and were on the same arm (5_ or 3_) of their presumptive hairpins.

Testing robustness of hairpin folding

Perfect matches to the 20mers in *At*Set2 and *Os*Set2 were found in their respective genomes using Patscan (Dsouza et al., 1997). For each genomic match 240 nt of flanking sequence on each side was joined to the 20mer and the RNAfold-predicted secondary structure was tested for a miRNA-like hairpin by MIRcheck, using less stringent cutoffs (no more then 6 unpaired nt in the putative miRNA or miRNA*, basepairing must extend 3 nt beyond the 20mer). 20mers for which more than 50% of more of genomic matches overlapped masked sequence elements were removed from the analysis. In addition, *Arabidopsis* 20mers for which less than 50% of intergenic matches had miRNA-like hairpins as defined by MIRcheck were removed from the analysis.

Identification of miRNA-like patterns of conservation between putative homologs

Pairwise comparisons were made between putatively homologous *Arabidopsis* and *Oryza* miRNA hairpins to identify those pairs with greatest conservation at the ends of the hairpins. For this purpose, each hairpin sequence was defined as the sequence containing the putative miRNA and putative miRNA* plus 10 flanking nucleotides on each side. For each pair of potentially homologous hairpins the *Arabidopsis* hairpin was broken up into overlapping 20mers, and the best local alignment between each *Arabidopsis* 20mer and the *Oryza* hairpin was found using LALIGN (Pearson, W.R., ftp://ftp.virginia.edu/pub/fasta/). An *Arabidopsis*:*Oryza* pair was considered to pass if it met all of the following requirements: a) the best local alignment to a 20mer from the vicinity of the putative *Arabidopsis* miRNA (defined as the 20mers starting within 3 nt of

the start of the of putative miRNA) mapped to within 10 nt of the start of putative *Oryza* miRNA, b) the best local alignment to a 20mer from the vicinity of the putative *Arabidopsis* miRNA* (defined as the 20mers starting within 3 nt of the start of the of putative miRNA*) mapped to within 10 nt of the start of putative *Oryza* miRNA*, c) the best local alignment between a 20mer from the vicinity of the putative *Arabidopsis* miRNA had at least as many matching nucleotides as the best local alignment to a 20mer from the vicinity of the putative *Arabidopsis* miRNA*, and d) the best local alignment to a 20mer from the vicinity of the putative *Arabidopsis* miRNA* had more matching nucleotides than the best local alignment to any 20mer from the loop region of the *Arabidopsis* hairpin.

Grouping 20mers into families of related sequences

Overlapping genomic 20mers corresponding to *At*Set5 and *Os*Set5 were joined together into longer regions with miRNA-encoding potential. Families of related potential miRNAs were defined both by sequence similarity as defined by blastn (Gish, W. (1996-2004) http://blast.wustl.edu) and by having genomic loci that correspond to sense or antisense versions of the same putative hairpin.

Identification of conserved pairing to mRNAs

Complementary sites to miRNAs in mRNA datasets were found using 3 searches: 1) A search for ungapped antisense matches using Patscan, in which G:U pairs are counted as _ a mispair and up to 4 total mispairs are allowed, 2) a perl script that searches for antisense matches with up to 1 insertion and up to 2 mispairs (G:U counted as mispair),

and 3) a perl script that searches for antisense matches with up to 1 deletion and up to 2 mispairs (G:U counted as mispair). Complementary sites were given a score S = MP + 2*(INDEL) + _(GU), where MP equals the number of mispairs (not counting G:U pairs), INDEL equals the number of insertion and deletions, and GU equals the number of G:U pairs. The results of these 3 searches were combined and made non-redundant. Complementary sites were considered to be conserved at a certain score cutoff if both an *Arabidopsis* miRNA and at least one homologous *Oryza* miRNA had complementary sites to homologous mRNAs from *Arabidopsis* and *Oryza*, respectively, that both met the score cutoff.

Identification of miRNAs and conserved miRNA complementary sites in ESTs

The April 5, 2004 release of non-human, non-mouse ESTs was downloaded from NCBI. Near matches to 20mers from *Arabidopsis* and *Oryza* miRNAs (0-1 nt changes) were found with Patscan. Matches that passed MIRcheck with the putative miRNA on the same arm of the hairpin as the *Arabidopsis* or *Oryza* miRNA were counted as potential miRNA homologs. Near matches in ESTs to 20mers comprising conserved, validated Near matches in ESTs to validated conserved *Arabidopsis* miRNA complementary sites (see Table 3) were found in the same manner. For each EST with a potential miRNA complementary site, putative homologs in the set of *Arabidopsis* proteins were found using blastx (maximum E value = 10^{-6}). Those cases in which near matches to *Arabidopsis* miRNA complementary sites were found in ESTs homologous to the corresponding *Arabidopsis* protein were considered to be potentially conserved miRNA complementary sites.

Probes for Northern blots

The sequences of the probes used for Northern blots were GATCAATGCGATCCCTTTGGA (miR393), AGGAGGTGGACAGAATGCCAA (miR394), AGTTCAAGAAAGCTGTGGAA (miR396a), GCAGGGTGACCTGAGAACA (miR398b), GAGTTCCCCCAAACTCTTCAG (miR395a), GTGCTCACTCTCTTCTGTCA (miR156), and TAGAGCTCCCTTCAATCCAAA (miR159).

Primers for PCR validation of miRNAs

The primers which amplified the corresponding cDNAs were AGGATCAATGCGATCCC (miR393), GAAGGAGGTGGACAGAA (miR394), GGAGGTGGACAGAATGC (miR394), GGAGTTCCCCCAAACAC (miR395a), AAGTTCAAGAAAGCTGTG (miR396b), ATCAACGATGCACTCAA (miR397b), AGGGGTGACCTGAGAAC (miR398a), and AGGGCAACTCTCCTTTG (miR399).

Primers for 5_ RACE mapping of miRNA cleavage sites

gene	nesting gene specific primer	nested gene specific primer
At1g12820	GGTGCATCTTTTCTAGTCCCAACCACTGTT	GTCCCAACCACTGTTCGGTAGAGGTAAAGT
At3g26810	CTAATTCGGGAAAGACACACTAACGGAAGA	AAAGACACACTAACGGAAGACGACCAATCA
At3g62980	TTGATGGTTATATGAAGCAGCCGAGATTCA	ATTGGCGACCGAGCTCTCTCATGTTCCTAA
At4g03190	ACAAATGGAAGGTGAATCTCTCCCACAACA	AGACCATGCGATCCCTTTGGATGT
At3g23690	AGCTCTTGTTACTCTCCTGCGTTGTTGGTT	AAATTGACCTGGATATGGAGGGAATCTGA
At1g27340	CAAAAAGAAACCAACCGTGTGTGAACCGTT	CCAACCGTGTGTGAACCGTTCGAATACAAC
At5g43780	GAGTATCCGTCATAAGAAGCGCATGACCAT	ACCATTATGCACCGGGTTCCTAAGCTGGAA
At2g22840	GATGAGTCCGAATTCAAAGGGATTATTGCT	CCTTCGACTGTAAGTTCATCGTGGCAGGAA
At2g36400	CCTCTTCGTTGCCTGTGGACAGCTTCA	GAGCGAAGAACGAGGCCAATCGTCAA
At2g45480	CTAAAGACGGACGAATGACATTTTCACAGG	TGGCACTAGGCAGTGAAGCAATGGTAACTA
At4g24150	ATGGCATCATCTTCTTTCGTAAACCACCTC	CATATGGAGAGGAGGCAGACGGAAGAGTAG
At4g37740	TAACCGGAGATTCCTTGGGTTGTAAGTTGA	CGGTGGAAGGATATATCGAAACTCCTCTGT
At5g53660	GTAGACATCGGAGCAAGCCAGCTTGACA	TAAGGACCATACCCGTGATGATCCTCACTA
At2g29130	ATTATCAATTGTTCCTTGGCCCGTGAAGTA	GCCAACATGTAGAAGGTTGCATTTGGATAA
At2g38080	AGGAGGAGAGTGTTCCCGAGTAATGAACAG	GTCCATGAAAGGAGAAGCGGTTACAAGGTA
At5g60020	TCAAAAGTTCCTTGACCTGTGACGTATGGT	AGTCATGAAGAAGGAGGCACTCGGATAACT
At1g08830	ATTAGCATCCTCAGGGGCACCGTGTGT	TGGCAATCAGTGATTGTGAAGGTGGCAGTT
At2g28190	GGTCAGACTAAGCTCATGGCCACCCTTT	AGGTCATCCTTAAGCTCGTGAACCACAAAG

At3g15640	ATAACGAAAGAAGATGCGGAAATGGCTGAT	CTAGGGTTTTGAGCTGAGAGGAGACGATTC
At3g05690	GTTTGAGAAACACACAAAGCCAGAGAGAGC	AGTGACCAAACACATCATTGTTCTTCACCA

Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408, 796-815.

Dsouza, M., Larsen, N., and Overbeek, R. (1997). Searching for patterns in genomic data. Trends Genet 13, 497-498.

Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, S., Tacker, M., and Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. Monatshefte f Chemie 125, 166-188.

Vazquez, F., Gasciolli, V., Crete, P., and Vaucheret, H. (2004). The nuclear dsRNA binding protein HYL1 is required for microRNA accumulation and plant development, but not posttranscriptional transgene silencing. Curr Biol 14, 346-351.

Yu, J., Hu, S., Wang, J., Wong, G. K., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., et al. (2002). A draft sequence of the rice genome (Oryza sativa L. ssp. indica). Science 296, 79-92**.**