Identification of MIR390a precursor processing-defective mutants in Arabidopsis by direct genome sequencing

Josh T. Cuperus\textsuperscript{a,b}, Taiowa A. Montgomery\textsuperscript{a,b}, Noah Fahlgren\textsuperscript{a,b}, Russell T. Burke\textsuperscript{b}, Tiffany Townsend\textsuperscript{b}, Christopher M. Sullivan\textsuperscript{b,c}, and James C. Carrington\textsuperscript{b,c,1}

\textsuperscript{a}Molecular and Cellular Biology Program; \textsuperscript{b}Department of Botany and Plant Pathology; and \textsuperscript{c}Center for Genome Research and Biocomputing, Oregon State University, Corvallis, OR 97331

Contributed by James C. Carrington, November 16, 2009 (sent for review November 2, 2009)

Transacting siRNA (tasiRNA) biogenesis in Arabidopsis is initiated by microRNA (miRNA)–guided cleavage of primary transcripts. In the case of TAS3 tasiRNA formation, ARGONAUTE7 (AGO7)–miR390 complexes interact with primary transcripts at two sites, resulting in recruitment of RNA-DEPENDENT RNA POLYMERASE6 (DCL4) for dsRNA biosynthesis. An extensive screen for Arabidopsis mutants with specific defects in TAS3 tasiRNA biogenesis or function was done. This yielded numerous ago7 mutants, one dcl4 mutant, and two mutants that accumulated low levels of miR390. A direct genome sequencing-based approach to both map and rapidly identify one of the latter mutant alleles was developed. This revealed a G-to-A point mutation (miR390a-1) that was calculated to stabilize a relatively nonpaired region near the base of the MIR390a foldback, resulting in misprocessing of the miR390/ miR390* duplex and subsequent reduced TAS3 tasiRNA levels. Directed substitutions, as well as analysis of variation at paralogous miR390-generating loci (MIR390a and MIR390b), indicated that base pair properties and nucleotide identity within a region 4–6 bases below the miR390/miR390* duplex region contributed to the efficiency and accuracy of precursor processing.

Results and Discussion

Screen for TAS3-Based syn-tasiRNA–Deficient Mutants. TAS3a-based synthetic (syn)-tasiRNAs with complementarity to the PDS mRNA provide a visual readout for tasiRNA activity in transgenic Arabidopsis (15). The 35S::TAS3aPDS-1 construct yields tandem syn-tasiRNAs from the 5′ D7[+] and 5′ D8[+] positions in place of siRNA2141 and siRNA2142, also known as tasi-ARFs (11, 12). These repress mRNAs encoding several AUXIN RESPONSE FACTORS, including ARF3 and ARF4, regulation of which is essential for proper developmental timing and lateral organ development (15, 17, 18) (Fig. L4). In wild-type (Col-0) plants expressing 35S::TAS3aPDS-1, photobleaching emanates from the midrib and major veins, with the phenotype most prominent when viewed from the adaxial side of leaves (Fig. 1B) (15). Syn-tasiRNA accumulation and photobleaching are suppressed in plants containing loss-of-function dcl4-2, dcl4-3, and zip-1 (AGO7-defective) mutants (15). A screen for mutants with TAS3 tasiRNA specific defects was done using the syn-tasiRNA line. Besides loss of photobleaching, mutants with TAS3-specific defects were predicted to have 1) low or no syn-tasiRNA and endogenous tasi-ARF (siRNA2142), 2) normal levels of TAS1 tasiRNA (siR255), 3) normal levels of miRNA, such as mir171, that do not function in the TAS3 pathway, and 4) an accelerated vegetative phase change (AVPC) phenotype, which is associated with loss of TAS3 tasiRNA (15, 17–20). TAS3 pathway-specific mutants were not expected to have severe developmental defects, as would be expected for general loss-of-miRNA function mutants (21, 22). The AVPC phenotype is characterized by downward-curled rosette leaves, giving the appearance of a narrow leaf phenotype, and early development of abaxial trichomes (19). In all, 200 pools of seedlings from the M2 generation were screened. A total of 355

Author contributions: J.T.C., T.A.M., and J.C.C. designed research; J.T.C., T.A.M., N.F., R.T.B., and T.T. performed research; N.F. and C.M.S. analyzed data; and J.T.C. and J.C.C. wrote the paper.

The authors declare no conflict of interest.

Published under a Creative Commons Attribution 3.0 Unported License.

This article contains supporting information online at www.pnas.org/cgi/content/full/0913203107/DCSupplemental.
hygromycin-resistant (transgene-containing) individuals with a reduced-photobleaching phenotype were recovered (Fig. 1C).

Of the mutants, 95 had relatively strong or severe vegetative defects, and 44 had no visible morphological defects. Most of the severely affected plants (class I) were dwarfed, had serrated leaves, and resembled known miRNA-debilitated mutants, such as those with hen1, hyll, se, or strong hypomorphic ago1 alleles (2, 8, 23) (Fig. S1). Among seven class I mutants analyzed, each had reduced levels of miR171 and TAS1 siR255, indicating that they were generally deficient in miRNA accumulation or activity (Fig. S1). Class I mutants were not analyzed further.

As exemplified by mutant 104a5, 216 mutants had an AVPC phenotype (Fig. 1B and C). Approximately 64% (138) or 25% (55) of these mutants lacked the TAS1 tasiRNA siR255 altogether, or produced siR255-related small RNA that migrated during electrophoresis as a 22-nt RNA, respectively, indicating that they possessed general (TAS3 and TAS1) tasiRNA defects (Fig. 1C and D). These 193 plants with general tasiRNA defects comprised class II mutants. Complementation analyses of a random sampling of 15 mutants, including 90b5 (Fig. 1D), revealed that the complete loss-of-siR255 subgroup within class II was dominated by rdr6 (10) and zgs3 (5) mutants (Table S1). Fourteen mutants produced 21-to-22 nt, size-shifted TAS1 siR255, and 14 of 14 of these possessed dcl4 defects based on complementation tests (Table S1). Loss of DCL4 is known to

Fig. 1. Syn-tasiRNA strategy, mutant screen, and characterization of class II and III mutants (A) Organization of syn-tasiRNA construct 35S:TAS3aPDS-1. The mIR390/miR390 target and syn-tasiRNA/PDS target sequences are shown in the expanded diagrams. (B) Photobleached phenotype of 35S:TAS3aPDS-1-transformed Col-0 and a class II mutant (104a5) recovered from the screen are shown next to nontransformed Col-0 (wt). (C) Flowchart of the screen for TAS3 tasiRNA-defective mutants using 35S:TAS3aPDS-1-transformed plants. (D and E) Representative images and select small RNA blot profiles from parental 35S:TAS3aPDS-1-transformed Col-0 (T) plants, reference mutants (rdr6-15 and zip1), and class II and class III mutants. Small RNA data using each radiolabeled probe in each panel were from the same blot. All small RNA, except for TAS3 siR2142 in the 70b1 mutant, comigrated with the 21-nt standard. (F) Mean (n = 3) relative level ± SD of TAS3 siR2142, miR390, miR171, and TAS1 siR255 (TAS3aPDS-1 = 1.0) in several mutants.
result in 22-nt size-shifted tasiRNA, because of the surrogate activity of DCL2 (24–26).

Only 12% (26) of plants possessed an AVPC phenotype, normal levels of 21-nt siR255, and normal levels of miR171 (Fig. 1C). Nearly all of these, which were designated as class III mutants, possessed low or undetectable levels of TAJ3 siR2142. Among the class III mutants, complementation analysis revealed 23 independent ago7 mutants, 14 of which were subjected to ago7 allele sequencing. Most of the ago7 alleles contained substituents affecting the PIWI domain, whereas single mutants with mid-domain or N-terminal domain substitutions were substitutions affecting the RNaseIII domain (Fig. 2). The 70b1 dcl4 allele contained a nonconserved Gly-to-Arg substitution affecting a region between the PAZ domain and the PIWI domain, whereas single mutants with mid-domain or N-terminal domain substitutions were identified (Table S1). A TAJ3-specific dcl4 mutant (70b1), in which TAJ3 siR2142-related small RNA, but not TAJ3 siR255, was shifted to 22 nt, was recovered, although minor reductions of both TAJ3 and TAJ3 tasiRNA were detected (Fig. 1E and Fig. S2). The 70b1 dcl4 allele contained a nonconserved Gly-to-Arg substitution affecting a region between the PAZ domain and first RNaseIII domain (Table S1).

Two recessive mutants, 52b2 and 87a3, could not be assigned to any of the complementation groups tested through crosses to zip1, rdr6-15, sgs3-11, and dcl4-2. These mutants had similar, moderate AVPC phenotypes (Fig. 1E). The 52b2 mutant accumulated significantly reduced levels siR2142 (44.9% compared with Col-0; P < 0.0028), but normal levels of miR171 and siR255 (Fig. 1E and F). Interestingly, both 52b2 and 87a3 had low levels of miR390 and TAS1 and TAS3 siRNA, but not miR390 (Fig. 1E), with quantitative blot assays revealing a significant difference (P < 0.0001) between 52b2 and Col-0 plants (Fig. 1F).

Identification of mir390a-1 by Pooled Genome Sequencing. In principle, direct genome sequencing of a mutant genome using high-throughput sequencing (HTS) technology can identify sites of mutation. However, each EMS-mutagenized genome can possess hundreds or thousands of changes in addition to the causal mutation. We developed a strategy for direct sequencing of a bulk segregating population of genomes for identification of the causative 52b2 mutation. A segregating F2 population from a cross between 52b2 (Col-0 background) and the polymorphic accession Ler was prepared, and 93 homozygous plants with both AVPC and low photobleaching phenotypes were identified. DNA from the 93 plants was pooled and subjected to high-throughput sequencing, which provided 221,000,000 36-base reads.

A pipeline, Mapping and Assembly with Short Sequences (MASS; Fig. 2A), was devised to map and assemble sequence data. Approximately 143,000 SNPs (27) were used to identify and quantify Col-0- and Ler-specific reads from repeat-filtered sequences. The ratios of summed Col-0 SNPs/summed Ler SNPs were calculated in 100,000 base windows (20,000 base scroll) across the Arabidopsis genome. A major peak of enriched Col-0 SNPs was identified on chromosome II (Fig. 2B). In addition, several minor peaks of Col-0-enriched SNPs were identified around pericentromeric regions. The basis for these minor peaks was not determined conclusively, although the peaks may reflect miscalled SNPs that do not exist in Ler. A 1.52-Mb region encompassing the major Col-0–enriched peak was assembled with the program Mapping and Assembly with Quality (MAQ) (28) using all high-quality sequencing reads, revealing five

Fig. 2. High-throughput sequencing of the 52b2 mutant genome and identification of the causal mutation (A) Flowchart of sequence-based mapping and mutation identification using a bulk segregant population. (B) Scrolling window plot of ratios (Col-0/Ler) of total SNPs detected in the bulk segregant sequence dataset. (C) A 152-Mb interval spanning the major Col-0–enriched region of chromosome 2 is illustrated. Each nucleotide position that deviates from the reference genome position is indicated by an arrow. The complete or partial sequences of mapped reads from a 50-base segment (chromosome 2 16069100–16069149) from the mir390a locus is shown in the expanded portion. (D) Restoration of photobleaching phenotype in 52b2 mutant plants by transformation with a wild-type mir390a transgene. (E) Foldback sequence and predicted structure from wild-type mir390a and mutant 52b2 mir390a-1 alleles. The position corresponding to the mutation is indicated by arrows. For comparative purposes, four foldback domains were assigned, as indicated by the brackets.
G-to-A or C-to-T changes that were consistent with EMS-induced mutation (Table S2). One mutation affected the sequence at the base of the foldback from MIR390a (Fig. 2 C and D). Portions of the 87a3 mutant genome were sequenced manually across the loci corresponding to the G-to-A or C-to-T positions in 52b2, revealing exactly the same mir390a mutation but wild-type Col-0 sequences at each of the other positions. Thus, 52b2 and 87a3 were independent mutants containing the same mir390a allele (mir390a-1). A genomic fragment containing wild-type MIR390a was introduced into 52b2 mutant plants. This restored photobleaching to the 52b2 mutant line and partially suppressed the leaf curling phenotype (Fig. 2E), confirming that the mir390a-1 mutation was causal.

The sequencing-based approach that identified the mir390a-1 mutation should be broadly applicable to identification of other markerless (e.g., EMS-induced) mutations. The major benefit of the approach is the simultaneous mapping and sequencing at a genome-wide level. The ability to score all known polymorphisms in individuals from the mapping population affords tremendous marker density, and the MASS pipeline provides a straightforward route to identification of a small number of candidate genes within a relatively small interval of 1–2 Mb. Similar high-throughput sequencing-based approaches for identification of casual mutations were presented recently (29–31).

The mir390a-1 mutation affects position 94 (G94-to-A94 substitution) from the 5′ end of the predicted foldback. Using both mFOLD and RNAFold (32, 33), G94 in the wild-type sequence was predicted to be nonpaired, or to base pair with U12 with low probability, in the “C region” of the foldback below the mir390/mir390′ segment (Fig. 2D). In mir390a-1, A94 was predicted to base pair with high probability to U12 (Fig. 2D). Due to the distance of the mutation away from the mir390/mir390′ end of the predicted foldback, the foldback is disrupted and contains only the 21-nt miR390*, as well as the differences in the C region between MIR390a and MIR390b (discussed below), this position may have significant information content, which correlates with variations in small RNA uniformity or processing accuracy at both ends of each sequence (Fig. 3B and C; SI Methods). The sequencing-based approach using Nicotiana benthamiana plants (34).

35S:MIR390a, 35S:mir390a-1, and 35S:MIR390b were expressed individually to analyze miR390 biogenesis and accumulation, or coexpressed with 35S:TA3aPDS-2 (syn-tasiRNA) and 35S:HA-AGO7 to test for TASS tasiRNA initiation activity. Compared with 35S:MIR390a, 35S:mir390a-1 yielded miR390 at 28.3% (P < 3.02 × 10−5) or 28.6% (P < 0.002) when expressed individually or with the other TASS tasiRNA components, respectively (Fig. 3A). These data suggest that processing of the mir390a-1 and MIR390b foldbacks occurs inefficiently.

To analyze processing accuracy of MIR390a, mir390a-1, and MIR390b foldbacks, small RNA libraries from triplicate samples were subjected to high-throughput sequencing analysis after transient expression in N. benthamiana. Reads were first normalized based on library size and spike-in standards (35). Reads from within 29-nt windows, centered around the middle of the annotated mir390 or mir390* sequences, were analyzed for size, 5′ position, and 3′ position. The information content of each dataset was used to calculate Shannon’s entropy (H) (36, 37), providing measures of small RNA uniformity or processing accuracy at both ends of each sequence (Fig. 3B and C; SI Methods). 35S:MIR390a yielded predominantly 21-nt, canonical miR390 with highly uniform 5′ and 3′ ends, and moderately heterogeneous 20–21 nt miR390* sequences.
with uniform 5′ ends but with 3′ ends from two major positions (Fig. 3B, Fig. S3 and Table S3). In contrast, 3′SS:mir390a-1 yielded 5′, 3′- and size-heterogeneous miR390 and miR390* sequences, with only 45.6% ± 22.6% of mir390-related sequences containing accurately processed 5′ and 3′ ends (Fig. 3B and Fig. S3). This was reflected in high H values for each 3′SS:mir390a–derived miR390 and miR390* parameter (Fig. 3C). MIR390b yielded sequences with intermediate processing accuracy. Both ends of miR390, and the 5′ end of miR390*, exhibited more heterogeneity than the comparable ends of sequences from MIR390a (Fig. 3 B and C). Combined with the syn-tasiRNA biogenesis data (Fig. 3A), these experimental findings indicate that the mir390a-1 mutation affects both processing efficiency and accuracy, resulting in low levels of functional miR390. The findings also indicate that MIR390b possesses the properties of a low-efficiency mutant allele. Natural variation affecting foldback structure and miRNA biogenesis has been shown previously in plants and animals (38, 39).

**Mutational Analysis of the MIR390a Foldback.** The G-to-A substitution in the mir390a-1 mutant could conceivably debilitate processing because of a change in foldback base pairing, loss of a base determinant, or both. Computational analysis of predicted foldback variants suggested that the mir390a-1 structure possessed a higher probability of base pairing between U12 and A94, compared with the probability of pairing between U12 and G94 in the wild-type foldback (Fig. 4A). This was reflected in a lower calculated entropy at both positions in the mir390a-1 foldback (Fig. 4A) (36, 37). The MIR390b predicted foldback, with even more extensive base pairing, yielded lower calculated positional entropies at nearly all bases in region C (Fig. 4A) (36, 37). Seven 3′SS: MIR390a mutants with substitutions at either position 94 and/or position 12 were constructed (Table S4). Including mir390a-1, the series resulted in foldbacks containing all possible single-base substitutions at both positions, and two combinations of dual-base substitutions (Fig. 4A). In addition, the sequences comprising MIR390a region C were substituted for the approximate equivalent sequences from MIR390b. Predicted foldback structures, positional entropies and miR390 biogenesis levels in a transient assay were determined.

Each substitution at position 94 (mir390a-1, mir390a-94U, and mir390a-94C) resulted in significantly (P < 0.003) lower miR390 levels compared with wild-type MIR390a, although the mir390a-94U and mir390a-94C defects were only modest (Fig. 4B). Unlike mir390a-1, mir390a-94U, and mir390a-94C mutations were not predicted to base pair with U12 (Fig. 4A). These data generally reinforce a role for G94, as either a single base-determinant or a high-entropy, weak-base-pair partner with U12. Among the position 12 substitutions, mir390a-12C was significantly (P < 8.8 × 10−5) debilitated for miR390 biogenesis and was predicted to form a low-entropy base pair with G94, mir390a-12A retained both A12 and G94 in a predicted nonpaired configuration and yielded wild-type levels of miR390 (Fig. 4). These position 12 mutants lend support to the idea that a nonpaired or weakly paired G94 contributes to miR390 biogenesis. In contrast, mir390a-12G was predicted to adopt a fold involving low-entropy, highly base-paired 12G and G94 positions, but led to wild-type levels of miR390 (Fig. 4). However, the mir390a-12G local stem structure was predicted to include novel, high-entropy asymmetric bulges that differed from the comparable positions from MIR390a (Fig. 4A).

Among the double mutants, mir390a-12C94A contained the A94 mutation from mir390a-1 and a base-pair–disrupting change at position 12 (Fig. 4A). This mutant was highly debilitated for miR390 biogenesis, indicating that the mir390a-1 defect (A94) was not due solely to the increased base pair configuration between positions 12 and 94 (Fig. 4). Interestingly, the mir390a-12G94U mutant foldback, which contained the G and U positions from wild-type MIR390a reversed, yielded nearly wild-type levels of miR390 (Fig. 4).

Substitution of the base of the MIR390a stem with that from MIR390b led to significant (P < 6.73 × 10−6) debilitation of miR390 formation (Fig. 4 B and C). The mutant foldback region C was predicted to contain the same low-entropy, highly base-paired configuration as predicted for MIR390b foldback (Fig. 4A). However, the functional significance of the MIR390b locus remains unclear.

The relatively high diversity of sequences, sizes, and secondary structures of plant MIRNA foldbacks (40) means that processing determinants are not particularly obvious. Based on in vitro processing assays with MIR167b foldbacks, DCL1 is sufficient to catalyze ATP-dependent pri- and premiRNA transcript processing, although only a minority of such products possess accurate 5′ and 3′ ends (6). The dsRNA binding motifs of DCL1 may provide a basal function for foldback recognition. However, inclusion of both SE and HY1 in these reactions increases the rate and accuracy of processing (6, 41). This may indicate that SE and HY1 function as accessory factors that position DCL1 accurately on substrates through interaction with one or more foldback structural features. We propose that the inaccurate and inefficient processing of the mir390a-1 foldback is due to loss of interaction with key factors.
promoting miRNA biogenesis. In particular, it is attractive to consider G94 in a flexible, high-entropy context as a recognition determinant for HYL1 and/or SE. Both SE and HYL1 promote miR390 accumulation in vivo (23, 42). Importantly, the effects of the \textit{mir390a-1} mutation on foldback processing in transient assays are very similar to the effects of \textit{MIR167b} foldback processing in the absence of SE and HYL1 in vitro (6). It seems unlikely, however, that the base pair position G94 is the sole determinant for such interactions, as there is high sequence and structural diversity at this position among foldbacks from conserved \textit{MiRNA} families. By analogy with the Drosha-Pasha/DGCR8 complex interacting with the base of animal foldbacks (43), features defining the junction between the base of the stem and the nonpaired region outside of the paired stem junction ~15 nucleotides from the miRNA/miRNA* duplex in accurate processing of many Arabidopsis miRNAs.

**Methods**

References for \textit{rdr6-15}, \textit{dcl4-2}, \textit{sgs3-11}, \textit{hen1-1}, \textit{hy1-2}, \textit{se-2}, \textit{hst-5}, \textit{dcl1-7}, and \textit{zip-1} alleles were described (12). Detailed descriptions, protocols, and references for transgenic plant materials, RNA blot assays, transient expression assays in \textit{N. benthamiana}, Arabidopsis mutagenesis and genetic screen, sequencing and analysis of small RNA populations, and the MASS pipeline are provided in \textit{SI Methods}.

**ACKNOWLEDGMENTS**

We thank Sarah Dvorak for assistance with the genetics screen and technical assistance. We also thank Detlef Weigel, Korbinian Schneeberger, and Richard Clark for helpful discussion, genomic sequencing test data sets, and SNP data; and we thank Nina Fedoroff, Mike Axtell, Detlef Weigel, and Javier Palatnik for sharing data and manuscripts prior to publication. This work was supported by grants from the National Science Foundation (MCB-0618433), National Institutes of Health (AI43288) and United States Department of Agriculture–National Research Initiative (2006-35301-17420).

---