Role of RNA polymerase IV in plant small RNA metabolism

Xiaoyu Zhang, Ian R. Henderson, Cheng Lu, Pamela J. Green, and Steven E. Jacobsen

PNAS 2007;104;4536-4541; originally published online Mar 5, 2007;
doi:10.1073/pnas.0611456104

Role of RNA polymerase IV in plant small RNA metabolism

Xiaoyu Zhang[†], Ian R. Henderson[†], Cheng Lu[‡], Pamela J. Green[‡], and Steven E. Jacobsen[†§¶]

[†]Department of Molecular, Cell, and Developmental Biology, University of California, Los Angeles, CA 90095; [‡]Department of Plant and Soil Sciences and Delaware Biotechnology Institute, University of Delaware, Newark, DE 19711; and [§]Howard Hughes Medical Institute, University of California, Los Angeles, CA 90095

Edited by Susan R. Wessler, University of Georgia, Athens, GA, and approved January 22, 2007 (received for review December 22, 2006)

In addition to the three RNA polymerases (RNAP I–III) shared by all eukaryotic organisms, plant genomes encode a fourth RNAP (RNAP IV) that appears to be specialized in the production of siRNAs. Available data support a model in which dsRNAs are generated by RNAP IV and RNA-dependent RNAP 2 (RDR2) and processed by DICER (DCL) enzymes into 21- to 24-nt siRNAs, which are associated with different ARGONAUTE (AGO) proteins for transcriptional or posttranscriptional gene silencing. However, it is not yet clear what fraction of genomic siRNA production is RNAP IV-dependent, and to what extent these siRNAs are preferentially processed by certain DCL(s) or associated with specific AGOs for distinct downstream functions. To address these questions on a genome-wide scale, we sequenced ≈335,000 siRNAs from wild-type and RNAP IV mutant *Arabidopsis* plants by using 454 technology. The results show that RNAP IV is required for the production of >90% of all siRNAs, which are faithfully produced from a discrete set of genomic loci. Comparisons of these siRNAs with those accumulated in *rdr2* and *dcl2 dcl3 dcl4* and those associated with AGO1 and AGO4 provide important information regarding the processing, channeling, and functions of plant siRNAs. We also describe a class of RNAP IV-independent siRNAs produced from endogenous single-stranded hairpin RNA precursors.

Arabidopsis | epigenetic | gene silencing | DNA methylation | RNA interference

Small RNAs (sRNAs) are essential components of most eukaryotic genomes and play important roles in many biological processes. In *Arabidopsis thaliana*, sRNAs are 21–24 nt long and function in both transcriptional gene silencing by directing DNA and histone methylation and posttranscriptional gene silencing through inhibition of translation and degradation of target mRNAs (for reviews, see refs. 1–6). Four distinct types of sRNAs have been identified in plants. MicroRNAs (miRNAs) and transacting siRNAs (tasiRNAs) are primarily involved in regulating gene expression and plant development, and siRNAs play a major role in defending the genome against the proliferation of invading viruses and endogenous transposable elements. The function of the fourth type of sRNAs, natural-antisense siRNAs (nat-siRNAs), is not entirely clear but is likely related to plant stress responses (1–6).

There are major differences in the mechanisms responsible for the production, processing, and channeling of different types of sRNAs. miRNA precursors are single-stranded hairpin RNAs transcribed by RNA polymerase II (RNAP II), which are processed by DCL1 into mostly 21-nt sRNAs and then primarily associated with ARGONAUTE1 (AGO1) (7–9). The precursors for tasiRNAs are dsRNAs produced by RNAP II and RNA-dependent RNAP 6 (RDR6), which are processed by DCL4 into 21 nucleotides and require AGO7 for their downstream functions (10–12). nat-siRNAs are derived from dsRNAs formed between sense–antisense pairing of overlapping RNAP II transcripts, and the AGO protein involved has yet to be identified (13, 14). Finally, the production of siRNAs is known to involve RNAP IV, RNA-dependent RNAP 2 (RDR2), and all four DCLs (15–20), and siRNAs are primarily incorporated into AGO4 but also into other AGOs, such as AGO1 (8, 9, 21).

RNAP IV is a recently identified class of RNAP that is specific to plant genomes. Unlike RNAP I, II, and III, RNAP IV appears to be specialized in siRNA metabolism, because *nrpd1a*, *nrpd1b*, or *nrpd2a* mutants are phenotypically normal but defective in siRNA production at all endogenous loci tested (16–19). RNAP IV exists in two distinct forms, one consisting of the subunits Nuclear RNA Polymerase D 1a (NRPD1a) and NRPD2a and the other composed of NRPD1b and NRPD2a. It has been proposed that the NRPD1a/NRPD2a form functions together with RDR2 in the production of siRNA precursors, whereas the NRPD1b/NRPD2a form is involved in the targeting of DNA methylation by siRNAs (RNA-directed DNA methylation, RdDM) (17, 19, 21–23). However, many questions concerning the functioning of RNAP IV remain unanswered, and the role of RNAP IV in siRNA production on a genome-wide scale remains unknown. It is also unclear to what extent RNAP IV acts together with RDR2 and the four DCL enzymes in *Arabidopsis* (15, 20, 24–27) or with downstream effectors such as AGO4 (28, 29) or AGO1 (9).

To address these questions on a genome-wide scale, we compared the siRNAs accumulated in wild-type and *nrpd* mutant plants through the cloning and sequencing of large quantities of sRNAs by using 454 technology. We found that RNAP IV is required for the production of >90% of all siRNAs. In addition, the siRNA profiles of wild type and *nrpd* mutants were compared with those of *rdr2* and *dcl2 dcl3 dcl4*, as well as those associated with AGO1 and AGO4 (9, 20, 30). The most striking result from these comparisons was the strong similarity among the profiles of siRNAs that depend on RNAP IV and RDR2. We also identified a class of RNAP IV-independent endogenous siRNAs derived from single-stranded hairpin precursors that were found to persist in both *nrpd* and *rdr2* mutants. These results strongly support the notion that RNAP IV functions together with RDR2 in the synthesis of double-stranded siRNA precursors. Finally, by reintroducing wild-type copies of the RNAP IV genes into previously mutant backgrounds, we found that the profiles of siRNAs were reestablished in a remarkably faithful manner, suggesting that RNAP IV may be recruited to a specific set of genomic loci in the absence of prior siRNA signals.

Results and Discussion

Characterization of sRNA Diversity by Large-Scale 454 Sequencing. To infer the function of RNAP IV, we characterized and compared the

Author contributions: X.Z. and S.E.J. designed research; X.Z., I.R.H., and C.L. performed research; P.J.G. contributed new reagents/analytic tools; X.Z. analyzed data; X.Z. wrote the paper; and S.E.J. edited the paper.

The authors declare no conflict of interest.

This article is a PNAS direct submission.

Freely available online through the PNAS open access option.

Abbreviations: sRNA, small RNA; miRNA, microRNA; tasiRNA, transacting siRNA; nat-siRNA, natural–antisense siRNA; RNAP, RNA polymerase; PM, perfectly matched *Arabidopsis* sequences; MPSS, massively parallel signature sequencing; RISC, RNA-induced silencing complex; RDR2, RNA-dependent RNAP 2.

[¶]To whom correspondence should be addressed. E-mail: jacobsen@ucla.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0611456104/DC1.

Table 1. Summary of sRNA sequences used in this study

	Wild type	*nrpd1a/1b*	*nrpd2a/2b*	F_1*
RAW sequences	76,772	106,905	72,604	78,293
Perfectly matched to genome	56,170 (73.2%)†	78,067 (73.0%)†	49,017 (67.5%)†	55,070 (70.3%)†
Filtered out‡	404 (0.5%)†	2,361 (0.2%)†	1,145 (0.2%)†	932 (0.1%)†
miRNAs	10,408 (18.7%)§	54,501 (72.0%)§	32,082 (67.0%)§	8,416 (15.5%)§
tasiRNA	799 (1.4%)§	4,491 (5.9%)§	2,561 (5.3%)§	959 (1.8%)§
siRNAs¶	44,559 (79.9%)§	16,714 (22.1%)§	13,229 (27.6%)§	44,763 (82.7%)§

*From the cross *nrpd1a/1b* X *nrpd2a/2b.*
†As percentage of raw sequences.
‡Matched abundant cellular RNAs such as tRNAs.
§As percentage of perfect matches to the genome and excluding those that were filtered out.
¶May contain unidentified miRNAs and tasiRNAs.

sRNA populations accumulated in wild-type and *nrpd* mutant plants through the cloning and sequencing of large numbers of sRNAs by using 454 technology (Table 1). sRNAs reads (76,772) were generated from wild-type inflorescences, of which 56,170 (>73%) perfectly matched the *Arabidopsis* genomic sequence over their entire length (PM). A small fraction of PM sRNAs (404 reads; ≈0.7%) matched abundant cellular RNAs (e.g., tRNAs) and were eliminated from further analyses, because they could represent degradation products. Of the remaining PM reads, 10,408 were miRNAs (18.7%), 799 matched tasiRNAs (1.4%), and 44,559 were primarily siRNAs (79.9%).

The 454 data set from the wild-type Columbia strain generated here was ≈7-fold larger than a previous 454 sRNA study (20), making many detailed analyses possible. The coverage of this data set was evaluated by comparing it to the 721,044 wild-type inflorescence sRNA reads generated by using a different method, massively parallel signature sequencing (MPSS), which is higher throughput but does not provide information about the size of the siRNA, because it generates only 17-bp sequences (31). All four tasiRNAs and 31 of the 35 miRNAs present in the MPSS data set were found in our 454 data set; we also recovered two additional low-copy miRNA families that were not in the MPSS data set. With regard to siRNAs, 5,044 of the 5,363 moderate or dense siRNA clusters (>94%) defined by the MPSS data were represented by the 454 data set. These results suggest that the 454 data set generated here provides a reasonable representation of the sRNA population in wild-type inflorescences.

One major technological advantage of 454 sequencing compared with MPSS is its ability to sequence through the entirety of cloned sRNAs, thus revealing the length of each sRNA and providing important clues regarding its origin and biological function. In *Arabidopsis*, different DCL enzymes usually produce sRNAs with distinct lengths. In general, DCL1 produces 21-mer miRNAs and nat-siRNAs, DCL2 produces 22 mers, DCL3 produces 24-mer siRNAs, and DCL4 produces 21-mer tasiRNAs (5, 13, 15, 20, 24–26, 32, 33). We therefore focused our analyses on these three size classes, 21, 22, and 24 mers (see *Methods*). As shown in Fig. 1, in wild type, the ratio of 21:22:24 mers is ≈1:0.35:1.99 for all sRNAs and 1:1.25:7.59 for siRNAs. Thus, the vast majority of siRNAs in wild type are 24 mers.

Consistent with their role in silencing transposons and other repetitive sequences, siRNAs of all three size classes showed a marked enrichment in heterochromatic regions where transposons and other repeats cluster [supporting information (SI) Fig. 4]. All three sizes were also found to be depleted from genes with known functions (SI Fig. 5). Interestingly, different size classes appeared to be preferentially associated with different types of repeats. In particular, 24 mers were more frequently associated with dispersed repeats than with tandem and inverted repeats, but 21 and 22 mers were more frequently associated with inverted and dispersed repeats than with tandem repeats (SI Fig. 6).

Considering these differences and the dependence of siRNA production on distinct DCL enzymes, 21-, 22-, and 24-mer siRNAs were analyzed separately. In addition, we used a proximity-based algorithm to group siRNAs into clusters (i.e., genome regions corresponding to multiple closely spaced siRNAs; see *Methods*) (31). These clusters may represent "sites of action," where siRNAs were produced. In this way, 686 21-mer clusters, 952 22-mer clusters, and 5,703 24-mer clusters were defined. Interestingly, the majority of 21- and 22-mer clusters, as well as a substantial fraction of 24-mer clusters, overlapped with each other (≈84%, ≈85%, and ≈21%, respectively; SI Fig. 7), suggesting that multiple siRNA-producing machineries (e.g., multiple DCLs) may coexist and/or function together at numerous loci genome wide.

siRNA Clusters Derived from Single-Stranded Hairpin RNA Precursors. One interesting feature of siRNA clusters that has not been systematically examined in previous studies is their "strandedness"; that is, whether considerable numbers of clusters exist where all siRNAs can be mapped to only one strand of the DNA. This is of particular interest to this study, because the derivation of siRNAs from both strands suggests that such siRNAs are processed from dsRNA substrates produced by RNAP IV and RDR2 or through the pairing of sense–antisense RNAP II transcripts. In contrast, the production of a cluster of siRNAs from only one strand would suggest single-stranded hairpin RNAs as DCL substrates. To address this question, we first identified all siRNA reads that matched only one genomic location ("unique PMs"), and thus their origins could be unambiguously determined. Next, we examined each siRNA cluster with ≥10 unique PMs and defined a cluster as single-stranded if the vast majority of the unique PMs (>90%) were derived from the same strand. As listed in SI Table 3, single-stranded siRNA clusters could be readily identified for all three

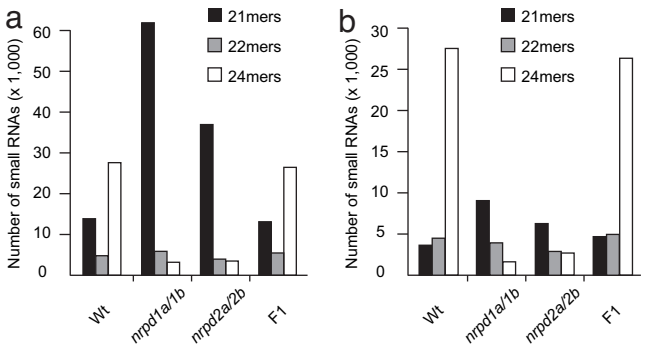

Fig. 1. The lengths of sRNAs that perfectly matched the *Arabidopsis* genomic sequence in wild-type, *nrpd1a/1b*, *nrpd2a/2b*, and the F_1 progenies from a cross between *nrpd1a/1b* and *nrpd2a/2b*. (*a*) All sRNAs. (*b*) siRNAs (miRNAs and tasiRNAs were excluded). *y* axis, the number of individual sRNA sequences of a certain size.

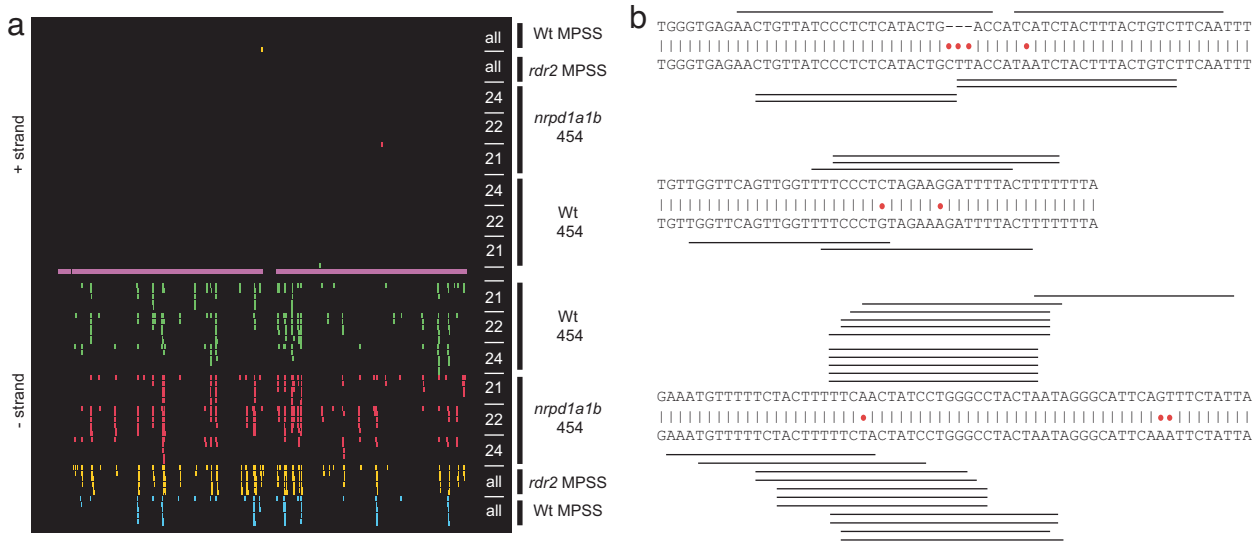

Fig. 2. The locus *IR71* as an example of siRNAs produced from single-stranded hairpin RNA precursors. (*a*) All individual siRNAs at *IR71* that matched a unique position in the genome isolated from wild type (454 or MPSS), *nrpd1a/1b* (454), and *rdr2* (MPSS). Vertical color bars, individual siRNAs; pink horizontal bars, the two arms of the inverted repeat. siRNAs shown above the inverted repeats matched the Watson strand, and those below matched the Crick strand of the genome. (*b*) Examples of siRNAs produced from the imperfectly matched regions (red dots) of the predicted single-stranded hairpin of *IR71*. Horizontal bars, siRNAs that match the left (above the alignment) or right arm (below the alignment).

sizes. Notably, a much higher fraction of 21-mer unique PM clusters (13 of 29; 44.8%) were found to be single-stranded than 22-mer unique PM clusters (6 of 53; 11.3%) or 24-mer unique PM clusters (8 of 325; 2.5%). These results likely represent a conservative estimate of the abundance of single-stranded siRNA clusters genome-wide, because most siRNA clusters correspond to repetitive sequences, which would not have met these conservative criteria (containing ≥10 unique PMs).

One example of a single-stranded siRNA cluster is *INVERTED REPEAT 71* (*IR71*), a large inverted repeat where all four DCLs are involved in siRNA production (20). As shown in Fig. 2*a*, virtually all unique PMs found at this locus were derived from the Crick strand of the genome, including 59 of the 60 21 mers, all 106 22 mers, and all 35 24 mers (Fig. 2*a*). In addition, all 712 siRNAs from the wild-type MPSS data set that mapped uniquely to this locus were found to be derived from the Crick strand, and virtually all unique PMs isolated from mutants such as *rdr2* and *nrpd1a/1b* (2,670 of 2,672 and 257 of 258, respectively) at this locus were also from the same strand (Fig. 2*a*). Considering that the unique PMs spanned both arms of the inverted repeats, these results strongly suggest that all siRNAs from *IR71* were derived from a ≈7-kb-long hairpin RNA precursor.

The two arms of *IR71* share 98.9% nucleotide sequence identity, and all of the unique PMs mapped to regions of the inverted repeat with one or occasionally two mismatches between the two arms (Fig. 2*b*). This result revealed an interesting property of the DCL enzymes. That is, similar to DCL1, DCL2, DCL3 and DCL4 can also produce siRNAs by "dicing" a hairpin RNA precursor in regions that contain mismatches. There are, however, two major differences between the processing of miRNA precursors and long-hairpin RNAs. First, miRNA precursors are processed exclusively by DCL1 into a single size class (almost always 21 mers), as shown by previous genetic studies and numerous miRNA Northern blot analyses (1–6, 20), as well as the fact that nearly all miRNAs from our data set were 21 mers (not shown). In contrast, all four DCLs are involved in siRNA production at *IR71* (20). Second, one strand of the DCL1 product (miRNA) is loaded into AGO1-containing RNA-induced silencing complex (RISC) and accumulates while the other strand (miRNA*) is degraded. In contrast, at *IR71*, siRNAs derived from both strands of imperfectly matched regions accumulate to similar levels (Fig. 2). Taken together, these results uncover similarities in the biochemical properties of the four DCLs but also suggest the involvement of additional factors in distinguishing miRNA precursors from other single-stranded hairpin RNAs.

DNA Methylation in siRNAs Clusters. siRNAs target *de novo* DNA methylation in *Arabidopsis* (17, 19, 28, 34–36), and the majority of siRNA clusters identified by using MPSS correspond to regions containing DNA methylation (37). However, it is not clear whether 21, 22, and 24 mers have a similar role in directing DNA methylation. To address this question, we first identified all siRNA clusters of a single size class (i.e., those that did not overlap with another cluster of a different size) and then determined the fractions of these clusters that colocalized with DNA methylated regions. As shown in SI Fig. 8, siRNAs of all three size classes were found to colocalize with methylated regions with frequencies that were higher than the genome average. However, a much higher fraction of 24- than 22-mer clusters were methylated, and 21-mer clusters were the least methylated (≈41.8% for 21 mers, ≈63.8% for 22 mers, and ≈90.1% for 24 mers, compared with the genome average level of methylation of ≈18.9%). This is in agreement with the previous finding that DCL3 products (24 mers) play a major role in RdDM, but DCL2 and DCL4 products can also direct DNA methylation at some loci (20).

RNAP IV Plays a Pivotal Role in siRNA Biogenesis. To explore the function of RNAP IV on a genome-wide scale, we cloned and sequenced 106,905 sRNAs from the inflorescences of *nrpd1a/1b* double mutant plants, of which 75,706 were PMs. A total of 54,501 reads were found to be miRNAs, and 4,491 were tasiRNAs. Neither the ratio of miRNA to tasiRNA (≈13.0 for wild type and ≈12.1 for *nrpd1a/1b*) nor the relative abundance of individual miRNA or tasiRNA families was affected in this mutant (SI Fig. 9). This is consistent with previous results examining individual miRNAs or tasiRNAs by using Northern blots, as well as the fact that the *nrpd1a/1b* mutant is phenotypically normal (16–19). Taken together, these results suggest that RNAP IV is not required for the production or function of miRNAs and tasiRNAs.

Comparison of the remaining 16,714 sRNAs from *nrpd1a/1b* to

Table 2. Association of RNAP IV-dependent and independent siRNAs with AGO1 and AGO4

siRNA	Total (%)*	RNAP IV independent (%)†
Total		
Wild type	35,473	3,438 (9.7)
AGO1	8,544	3,791 (44.4)
AGO4	13,738	1,102 (8.0)
21 mers		
Wild type	3,609 (10.2)	1,427 (39.5)
AGO1	6,046 (70.8)	3,048 (50.4)
AGO4	473 (3.4)	64 (13.5)
22 mers		
Wild type	4,459 (12.6)	984 (22.1)
AGO1	2,472 (28.9)	742 (30.0)
AGO4	1,843 (13.4)	185 (10.0)
24 mers		
Wild type	27,405 (77.3)	1,027 (3.7)
AGO1	26 (0.3)	1 (3.8)
AGO4	11,422 (83.1)	853 (7.5)

*The total of individual 21-, 22-, and 24-mer siRNAs. Numbers in parentheses show the percentage of siRNAs that belonged to the corresponding size class.
†Derived from siRNA clusters in *nrpd1a/1b*. Numbers in parentheses show the percentage of total siRNAs that are RNAP IV-independent.

those from wild type revealed several major differences. First, there is a vast reduction in siRNA abundance for all three size classes. As normalized by the number of miRNAs and tasiRNAs, the sequencing of the sRNAs from *nrpd1a/1b* was approximately ≈5.2- to 5.6-fold deeper than wild type, yet ≈2.7-fold fewer siRNAs were recovered. Thus, there was a ≈14- to 15-fold reduction in the abundance of siRNAs in *nrpd1a/1b* compared with wild type, indicating that the production of ≈93% of all siRNAs in wild-type plants required RNAP IV activity. Consistent with this estimate, only ≈9.7% of individual siRNA reads from wild type mapped to siRNA clusters that persisted in *nrpd1a/1b* (Table 2). Second, a smaller fraction of 21 mers (≈54.6%) were lost in *nrpd1a/1b* than 22 mers (≈84.0%), whereas nearly all 24 mers (≈98.9%) were eliminated. As a consequence, the most abundant size class was found to be 21 mers, whereas 24 mers were sparse; the ratio of 21:22:24 mers became 1:0.43:0.18 (in contrast to 1:1.25:7.59 for wild type; Fig. 1). Third, in *nrpd1a/1b*, although the chromosomal locations of 22 and 24 mers remained largely heterochromatic, the distribution of 21 mers was more sporadic (SI Fig. 4). Finally, in *nrpd1a/1b*, significantly smaller fractions of siRNAs from all three sizes classes were found to be derived from dispersed or tandem repeats, but a marked increase was observed in the fractions of 22 and 24 mers associated with inverted repeats (SI Fig. 6).

The siRNA clusters remaining in *nrpd1a/1b* were examined further for their strandedness and for structural characteristics of the corresponding genomic regions. For this analysis, we focused on clusters containing ≥10 or more unique PMs, as described above. All 22- and 24-mer clusters (seven and three, respectively), as well as the majority of 21-mer clusters (19 of 29; 65.5%) in *nrpd1a/1b* that met this criterion were found to be strand-specific and to localize within inverted repeats (SI Table 4). Four of the remaining 10 21-mer clusters were also strand-specific and processed from single-stranded RNA with secondary structures that resembled miRNA precursors (SI Fig. 10); these loci were not initially identified as inverted repeats, because the matched regions were relatively short. Additionally, most strand-specific siRNA clusters present in wild type were retained in *nrpd1a/1b* (SI Table 3). Taken together, these results suggest that the RNAP IV activity is required for the production of the vast majority of all sRNAs in the *Arabidopsis* genome except for miRNAs, tasiRNAs, and siRNAs produced from single-stranded hairpin RNA precursors. This is consistent with the fact that all individual endogenous siRNAs tested by RNA blot analysis in previous studies depend on RNAP IV (16–19, 23), but siRNAs derived from a highly expressed inverted-repeat transgene driven by a Pol II promoter (which may resemble the single-stranded hairpin RNA precursors described here) were found to be RNAP IV-independent (17).

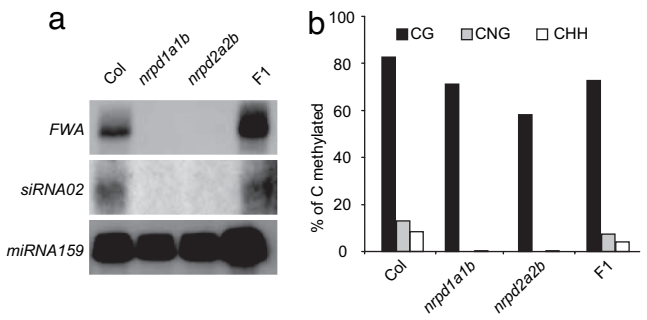

Fig. 3. Immediate restoration of siRNA production and RNA-directed DNA methylation. (*a*) Northern blot analysis of siRNAs derived from the *FWA* and *siRNA02* loci (16, 19, 38). miRNA 159 is shown as control. (*b*) Bisulfite sequencing analysis of DNA methylation at *MEA-ISR*.

The correlation of RNAP IV-independent siRNA clusters and DNA methylation was examined in a manner similar to that described above. siRNA clusters of all three size classes were found to be methylated at levels significantly higher than genome average (≈40.0% for 21 mers, ≈53.1% for 22 mers, and ≈77.8% for 24 mers). Furthermore, RNA-directed DNA methylation was indeed previously shown at *IR71*, a locus shown in this study to produce RNAP IV-independent single stranded siRNAs (20). These results suggest that RNAP IV-independent siRNAs can also direct DNA methylation, which is consistent with the finding that a considerable number of RNAP IV-independent siRNAs are associated with AGO4 (see below).

Robustness of RNAP IV Function at Defined Genomic Loci. Although RNAP IV is critically important in siRNA production, little is known about the early events in this process. For example, it is unknown whether the RNAP IV complexes are localized to specific genomic loci, or whether the prior existence of siRNAs is required to guide such localization in a self-reinforcing manner. The requirement of multiple subunits for RNAP IV function allowed the use of genetic strategies to address these questions. Considering that the RNAP IV activity is lost in either *nrpd1a/1b* or *nrpd2a/2b* (refs. 16–19, and see below), it was of interest to determine to what extent siRNA production can resume when the wild-type functions of all subunits are restored. We therefore crossed the *nrpd1a/1b* and *nrpd2a/2b* mutants together and examined the siRNA profiles of the progeny (F_1 plants) containing a wild-type copy of each gene.

We sequenced 72,604 and 78,293 sRNAs isolated from the inflorescences of *nrpd2a/2b* and the F_1 plants, respectively (Table 1). Consistent with the requirement of NRPD2a for NRPD1a/1b activities, the siRNA profile of *nrpd2a/2b* was found to be virtually the same as *nrpd1a/1b*. Remarkably, the F_1 plants shared a nearly identical siRNA profile with wild type. Approximately 93.8% of 21-mer clusters, ≈98.2% of 22-mer clusters, and ≈98.4% 24-mer clusters in wild type were also present in the F_1. Conversely, ≈97.3% of 21-mer clusters, ≈93.2% of 22-mer clusters, and ≈99.2% of 24 mers in the F_1 were also present in wild type. The subtle differences could be due to sampling, because clusters present only in wild type or F_1 were of relatively low abundance (not shown). These results were further validated by Northern blot analyses at two loci (Fig. 3*a*).

To determine whether the function of siRNAs in directing DNA methylation is also restored, we examined the DNA methylation status of *MEDEA–INTERGENIC SUBTELOMERIC REPEAT*

MEA-ISR) by using genomic bisulfite sequencing. *MEA-ISR* was chosen for this analysis, because virtually all non-CG methylation at this locus depends on the presence of siRNAs (20, 35). As shown in Fig. 3*b*, both CG and non-CG sites were methylated in wild type, whereas non-CG methylation was eliminated in *nrpd1a/1b* or *nrpd2a/2b*. Significantly, non-CG methylation was restored in F_1 plants to near wild-type levels.

The immediate and full restoration of the production and function of siRNAs in F_1 plants could be explained in at least two ways. First, it is possible that the recruitment of the NRPD1a/2a complex to specific genomic loci to initiate siRNA production is extremely efficient and reproducible. If so, certain signal(s) might persist on the chromosomes in the absence of siRNAs or siRNA-directed DNA methylation. CG DNA methylation may be a plausible candidate mark; however, at the *FWA* locus, DNA methylation does not seem to be required for the recruitment of RNAP IV activity (38). A second possibility is that a component of the NRPD complexes could remain associated with chromatin in the mutants used in this study. For instance, although NRPD2a is unstable in *nrpd1a/1b*, and NRPD1b is unstable in *nrpd2a/2b*, NRPD1a remains roughly at wild-type level in *nrpd2a/2b* (19). It is, therefore, possible that NRPD1a is still bound to its sites of action in *nrpd2a/2b*, and siRNA production resumes when NRPD2a is restored. In either case, these results strongly suggest that the NRPD1a/2a complex is localized or recruited reproducibly to specific loci in the genome, and this targeting does not appear to require the prior existence of siRNAs or the DOMAINS REARRANGED METHYLASE (DRM)-dependent DNA methylation that depends on siRNAs.

Similar Roles of RNAP IV and RDR2 in siRNA Biogenesis. If RNAP IV and RDR2 function together to generate dsRNAs as siRNA precursors, the loss of the RNAP IV and RDR2 activities should result in similar defects in siRNA biogenesis. To test this on a genome-wide scale, we compared the siRNAs accumulated in *nrpd1a/1b* and *rdr2*. A large number of sRNAs (≈916,000) were recently generated from *rdr2* by using MPSS, and analyses of these sRNAs showed a marked decrease in the abundance of siRNAs and enrichments of miRNA and tasiRNAs (30). We found that nearly all siRNA clusters identified in *nrpd1a/1b* were also found as siRNA clusters in *rdr2*, including 175 of 182 21-mer clusters (≈96.2%), 91 of 93 22-mer clusters (≈97.8%), and 96 of 97 24-mer clusters (≈99.0%). Additionally, ≈98.5% of all RNAP IV-dependent clusters (i.e., those present in wild type but not in *nrpd1a/1b*) were found to be lost in *rdr2*. The high level of correlation despite the differences in sequencing methods strongly suggests that, in support of the model above, the *nrpd1a/1b* and *rdr2* mutants display largely the same defects in siRNA genesis.

Relationship Between RNAP IV and DICER Functions in siRNA Biogenesis. We analyzed 11,427 sRNA sequences from *dcl2 dcl3 dcl4* in a previous study, of which 1,586 were siRNAs produced by DCL1, the only remaining DICER enzyme in this mutant background (20). We compared these siRNAs to those identified here from *nrpd1a/1b* (primarily processed from single-stranded hairpin RNAs) to determine the dependence of the siRNA clusters in *dcl2 dcl3 dcl4* on RNAP IV. For this analysis, we focused on relatively abundant siRNA clusters in *dcl2 dcl3 dcl4* (those with ≥10 siRNAs) to avoid sampling artifacts caused by clustering of relatively sparse siRNAs. We found that 28 of the 31 clusters in *dcl2 dcl3 dcl4* were also present in *nrpd1a/1b*. Therefore, in *dcl2 dcl3 dcl4,* the major role of DCL1 in siRNA biogenesis appears to be the processing of single-stranded hairpin RNAs produced in an RNAP IV-independent manner. The remaining three 3 *dcl2 dcl3 dcl4* clusters were present in wild type as clusters of all three sizes with siRNAs matching both strands but were entirely missing from *nrpd1a/1b*, suggesting they were likely produced from RNAP IV-dependent dsRNAs. At all three loci, only 21-mer clusters remained in *dcl2 dcl3 dcl4* (SI Fig. 11). It thus appears that DCL1, in rare cases, can process RNAP IV-dependent dsRNA substrates.

Relationship Between RNAP IV and AGO Functions in siRNA Biogenesis. The sRNAs generated by the DCL enzymes are incorporated into RISCs containing different AGO proteins to perform different downstream functions. Specifically, miRNAs are incorporated into AGO1-containing RISC (8), siRNAs are incorporated into AGO4-containing RISC (9, 23), and the normal functions of tasiRNAs require AGO7 (26, 39, 40). A large number of sRNAs associated with AGO1 or AGO4 have recently been reported (9). To determine whether the siRNAs produced by RNAP IV are preferentially associated with AGO1 or AGO4, we analyzed the relative abundance of siRNAs in AGO1 and AGO4 that were derived from RNAP IV-dependent or independent clusters. As shown in Table 2, a significantly larger fraction of AGO1-associated siRNAs (≈44.4%) were derived from RNAP IV-independent clusters than of the total siRNAs in wild type (≈9.7%). In contrast, AGO4 exhibited a slight preference for RNAP IV-dependent siRNAs (Table 2), suggesting that the majority of siRNAs produced by RNAP IV are incorporated into AGO4. Furthermore, a detailed comparison revealed that AGO1 was preferentially associated with 21- and 22- mer (but not 24-mer) RNAP IV-independent siRNAs. In contrast, AGO4 was associated with significantly smaller fractions of 21 and 22 mers, but a higher fraction of 24 mers that were RNAP IV-independent. Thus the association of RNAP IV-independent siRNAs with either AGO1 or AGO4 appeared to be affected by their lengths. These results suggest that the origins of siRNA precursors (e.g., dsRNAs or single-stranded hairpin RNAs) may not be the primary determinant for which AGO they are associated with. Instead, the particular DCL enzymes processing these precursors or the lengths of the resulting siRNAs may play more important roles in determining their association with particular RISCs and their downstream functions.

Conclusions

sRNA data can be downloaded or visualized along with DNA methylation and related data from http://epigenomics.mcdb.ucla.edu/smallRNAs; all sRNAs described here are also included in SI Datasets 1–5. Our analyses of large numbers of sRNA sequences from wild type and several mutants have provided important insights into the role of RNAP IV in sRNA metabolism and function on a genome-wide scale. First, we found that RNAP IV is required for the production of the vast majority of all siRNAs; however, we also discovered a considerable number of endogenous siRNAs produced from single-stranded hairpin RNAs in an RNAP IV-independent manner. All four DCLs appear to be involved in this process by "dicing" hairpin RNA precursors even in regions that do not perfectly match. This observation uncovered a previously unknown biochemical property of DCL2, DCL3, and DCL4, thus raising the interesting question of what distinguishes a normal hairpin RNA (processed by all four DCLs) from a miRNA precursor (processed by DCL1 only). It is also interesting to consider that, because miRNAs are critically important in regulating plant development, their precursors may have evolved to be specifically recognized and processed by DCL1 such that miRNAs are accurately generated. In contrast, other single-stranded hairpin RNAs with no developmental functions or evolutionary constraints are more likely to be promiscuously recognized and processed by all four DCLs. Second, the nearly identical sRNA profiles of *nrpd* and *rdr2* mutants suggest that RNAP IV and RDR2 function together to produce dsRNAs, and that other RDR genes cannot substitute RDR2 in this process. Third, DCL1 primarily processes single-stranded hairpin RNAs (including miRNA precursors) but can occasionally process RNAP IV-dependent dsRNAs. Interestingly, this latter case resembles the production of nat-siRNAs with regard to the requirement for RNAP IV and DCL1 (13). Fourth, we found that RNAP IV-dependent and independent siRNAs are preferen-

tially associated with AGO4 and AGO1, respectively. In addition to their origins, the lengths of the siRNAs or the particular DCLs that generate them also seem to be important in determining which AGO protein they are associated with.

The primary function of siRNAs in plants is to defend against the proliferation of endogenous transposons or invading viruses through transcriptional or posttranscriptional gene silencing. Results presented here, as well as those from previous studies focused on individual loci, clearly indicate that significant functional redundancies exist at multiple steps. For example, multiple DCL enzymes function at the same genomic loci, and some DCLs can substitute for other DCLs to a certain degree (15, 20, 24–27). Further, the loading of certain types of siRNAs into a specific AGO may be strongly preferred but not absolute (9). Considering some viruses have evolved to inhibit specific components of the RNAi pathway, promiscuities in the processing and channeling and redundancies in the downstream functions of siRNAs may be advantageous to the host genome by providing a residual level of defense when the "default" pathway is compromised (41). In contrast to these redundant and plastic situations, we found that RNAP IV is strictly required for the production of all siRNAs except those processed from single-stranded hairpin RNAs. One interpretation of this finding is that the products of other RNAPs cannot efficiently serve as templates for RDR2 to produce dsRNAs. The basis for this distinction could lie in the recruitment of RNAP IV, its interaction with other proteins, or unique properties of the RNA it produces. In this regard, perhaps one of the most intriguing findings from this study is that RNAP IV-dependent siRNAs and their function in directing non-CG DNA methylation are immediately restored when wild-type NRPD proteins are reintroduced into mutants. This suggests that the recruitment and localization of RNAP IV to genomic regions are specific and robust. Future studies aimed at understanding the mechanisms responsible for the recruitment of RNAP IV to these specific loci should be enlightening.

Methods

Plant Materials. All *Arabidopsis* plants used in this study are of the Columbia (Col-0) accession. The *nrpd1a*/*1b* and *nrpd2a*/*2b* double mutants have been described (18, 19).

siRNA Isolation, Cloning, and 454 Sequencing. Plants were grown on soil under continuous light, and all genotypes were grown side by side to minimize potential variations caused by environmental factors. Floral tissues, including inflorescence meristem, floral buds, and open flowers, were used for this study. siRNA isolation, gel purification, cloning, and sequencing were performed as described (20, 30).

Bioinformatic Analyses of siRNA Sequences. Analyses of sRNA sequences were performed as described (20, 30). Only 21-, 22-, and 24-mer PMs were analyzed further (Table 1). Although a significant number of 23 mers were also recovered, they did not appear to represent a unique size class, because nearly all 23 mers overlapped with 24-mer clusters. siRNA clusters were defined similarly to that recently published (three or more siRNA reads that were <500 bp apart) (31).

sRNA, RNA Blot Analysis, and Bisulfite Sequencing. sRNA extraction, RNA blot analysis, and bisulfite sequencing were performed as described (20).

We thank Drs. Craig Pikaard (Washington University, St. Louis, MO) and Thierry Lagrange (Université de Perpignan, Perpignan, France) for providing the *nrpd1a*/*1b* and *nrpd2a*/*2b* mutants, Dr. Blake Meyers (University of Delaware) for providing the *rdr2* sRNA sequences, Drs. Yijun Qi (National Institute of Biological Sciences, Beijing, China) and Greg Hannon (Cold Spring Harbor Research Laboratory, Cold Spring Harbor, NY) for providing sRNA sequences associated with AGO1 and AGO4, and Drs. Matteo Pellegrini and Shawn Cokus (University of California, Los Angeles) for implementing the web browser. Research in the laboratory of P.J.G. is supported by National Science Foundation Grant 0548569. X.Z. is also supported by a postdoctoral fellowship from the Jonsson Cancer Center Foundation. I.R.H. was supported by a European Molecular Biology Organization long-term postdoctoral fellowship. Research in the laboratory of S.E.J. is supported by National Institutes of Health Grant GM60398 and a grant from the National Institutes of Health ENCODE Program HG003523. S.E.J. is an investigator of the Howard Hughes Medical Institute.

1. Wassenegger M (2005) *Cell* 122:13–16.
2. Jones-Rhoades MW, Bartel DP, Bartel B (2006) *Annu Rev Plant Biol* 57:19–53.
3. Zhang B, Pan X, Cobb GP, Anderson TA (2006) *Dev Biol* 289:3–16.
4. Meins F, Jr, Si-Ammour A, Blevins T (2005) *Annu Rev Cell Dev Biol* 21:297–318.
5. Vaucheret H (2006) *Genes Dev* 20:759–771.
6. Vazquez F (2006) *Trends Plant Sci* 11:460–468.
7. Papp I, Mette MF, Aufsatz W, Daxinger L, Schauer SE, Ray A, van der Winden J, Matzke M, Matzke AJ (2003) *Plant Physiol* 132:1382–1390.
8. Baumberger N, Baulcombe DC (2005) *Proc Natl Acad Sci USA* 102:11928–11933.
9. Qi Y, He X, Wang XJ, Kohany O, Jurka J, Hannon GJ (2006) *Nature* 443:1008–1012.
10. Allen E, Xie Z, Gustafson AM, Carrington JC (2005) *Cell* 121:207–221.
11. Peragine A, Yoshikawa M, Wu G, Albrecht HL, Poethig RS (2004) *Genes Dev* 18:2368–2379.
12. Vazquez F, Vaucheret H, Rajagopalan R, Lepers C, Gasciolli V, Mallory AC, Hilbert JL, Bartel DP, Crete P (2004) *Mol Cell* 16:69–79.
13. Borsani O, Zhu J, Verslues PE, Sunkar R, Zhu JK (2005) *Cell* 123:1279–1291.
14. Katiyar-Agarwal S, Morgan R, Dahlbeck D, Borsani O, Villegas A, Jr, Zhu JK, Staskawicz BJ, Jin H (2006) *Proc Natl Acad Sci USA* 103:18002–18007.
15. Xie Z, Johansen LK, Gustafson AM, Kasschau KD, Lellis AD, Zilberman D, Jacobsen SE, Carrington JC (2004) *PLoS Biol* 2:E104.
16. Herr AJ, Jensen MB, Dalmay T, Baulcombe DC (2005) *Science* 308:118–120.
17. Kanno T, Huettel B, Mette MF, Aufsatz W, Jaligot E, Daxinger L, Kreil DP, Matzke M, Matzke AJ (2005) *Nat Genet* 37:761–765.
18. Onodera Y, Haag JR, Ream T, Nunes PC, Pontes O, Pikaard CS (2005) *Cell* 120:613–622.
19. Pontier D, Yahubyan G, Vega D, Bulski A, Saez-Vasquez J, Hakimi MA, Lerbs-Mache S, Colot V, Lagrange T (2005) *Genes Dev* 19:2030–2040.
20. Henderson IR, Zhang X, Lu C, Johnson L, Meyers BC, Green PJ, Jacobsen SE (2006) *Nat Genet* 38:721–725.
21. Li CF, Pontes O, El-Shami M, Henderson IR, Bernatavichute YV, Chan SW, Lagrange T, Pikaard CS, Jacobsen SE (2006) *Cell* 126:93–106.
22. Vaucheret H (2005) *Nat Genet* 37:659–660.
23. Pontes O, Li CF, Nunes PC, Haag J, Ream T, Vitins A, Jacobsen SE, Pikaard CS (2006) *Cell* 126:79–92.
24. Gasciolli V, Mallory AC, Bartel DP, Vaucheret H (2005) *Curr Biol* 15:1494–1500.
25. Dunoyer P, Himber C, Voinnet O (2005) *Nat Genet* 37:1356–1360.
26. Yoshikawa M, Peragine A, Park MY, Poethig RS (2005) *Genes Dev* 19:2164–2175.
27. Bouche N, Lauressergues D, Gasciolli V, Vaucheret H (2006) *EMBO J* 25:3347–3356.
28. Zilberman D, Cao X, Jacobsen SE (2003) *Science* 299:716–719.
29. Zilberman D, Cao X, Johansen LK, Xie Z, Carrington JC, Jacobsen SE (2004) *Curr Biol* 14:1214–1220.
30. Lu C, Kulkarni K, Souret FF, MuthuValliappan R, Tej SS, Poethig RS, Henderson IR, Jacobsen SE, Wang W, Green PJ, *et al.* (2006) *Genome Res* 16:1276–1288.
31. Lu C, Tej SS, Luo S, Haudenschild CD, Meyers BC, Green PJ (2005) *Science* 309:1567–1569.
32. Schauer SE, Jacobsen SE, Meinke DW, Ray A (2002) *Trends Plants Sci* 7:487–491.
33. Xie Z, Allen E, Wilken A, Carrington JC (2005) *Proc Natl Acad Sci USA* 102:12984–12989.
34. Dalmay T, Hamilton A, Rudd S, Angell S, Baulcombe DC (2000) *Cell* 101:543–553.
35. Chan SW, Zilberman D, Xie Z, Johansen LK, Carrington JC, Jacobsen SE (2004) *Science* 303:1336.
36. Chan S, Henderson W-L, Zhang IR, Shah X, Chien G, JS-C & Jacobsen SE (2006) *PLoS Genet* 2(6):e83.
37. Zhang X, Yazaki J, Sundaresan A, Cokus S, Chan SW, Chen H, Henderson IR, Shinn P, Pellegrini M, Jacobsen SE, *et al*. (2006) *Cell* 126:1189–1201.
38. Chan SW, Zhang X, Bernatavichute YV, Jacobsen SE (2006) *PLoS Biol* 4(11):e363.
39. Xu L, Yang L, Pi L, Liu Q, Ling Q, Wang H, Poethig RS, Huang H (2006) *Plant Cell Physiol* 47:853–863.
40. Adenot X, Elmayan T, Lauressergues D, Boutet S, Bouche N, Gasciolli V, Vaucheret H (2006) *Curr Biol* 16:927–932.
41. Deleris A, Gallego-Bartolome J, Bao J, Kasschau KD, Carrington JC, Voinnet O (2006) *Science* 313:68–71.