One precursor One siRNA model for Pol IV-dependent RNA directed DNA methylation.

Jixian Zhai[1,#], Sylvain Bischof[1,#], Haifeng Wang[2,1], Suhua Feng[1] , Tzuu-fen Lee[3], Chong Teng[3], Xinyuan Chen[4], Soo Young Park[5], Linshan Liu[5], Javier Gallego-Bartolome[1,6], Wanlu Liu[1], Ian R. Henderson[1,$], Blake C. Meyers[3], Israel Ausin[2,*] and Steven E. Jacobsen[1,7,*]

[1] Department of Molecular, Cell and Developmental Biology, University of California at Los Angeles, Los Angeles, CA 90095, USA;

[2] Haixia Institute of Science and Technology (HIST), Fujian Agriculture and Forestry University, Fuzhou 350002, China;

[3] Department of Plant and Soil Sciences and Delaware Biotechnology Institute, University of Delaware, Newark, DE 19716, USA;

[4] Department of Microbiology, Immunology, and Molecular Genetics, [5] Department of Chemistry and Biochemistry, University of California at Los Angeles, Los Angeles, CA 90095, USA;

[6] Plant Biology Laboratory & Howard Hughes Medical Institute, The Salk Institute for Biological Studies, La Jolla, CA 92037, USA;

[7] Howard Hughes Medical Institute, University of California at Los Angeles, Los Angeles, CA 90095, USA.

[#] Co-first authors

[$] Present address: Department of Plant Sciences, University of Cambridge, Cambridge CB2 3EA, United Kingdom.

[*] Correspondence: israelausin@fafu.edu.cn (I.A.), jacobsen@ucla.edu (S.E.J.)

SUMMARY

RNA-directed DNA methylation in *Arabidopsis thaliana* is driven by the plant-specific RNA Polymerase IV (Pol IV), which is believed to transcribe precursor RNAs that give rise to 24-nt small interfering RNAs (siRNAs) that target DNA methylation. However, very little is known about the mechanisms of Pol IV action or the nature of Pol IV transcripts. Here, we describe Pol IV-dependent RNAs (P4RNAs) from wild-type Arabidopsis that are surprisingly short in length (30 to 40 nucleotides) and that mirror 24-nucleotide siRNAs in distribution, abundance, strand bias, and 5'-adenine preference. P4RNAs exhibit transcription-start-sites (TSSs) similar to Pol II products, and have unique features such as 5'-monophosphates and 3'-misincorporated nucleotides. The 3'-misincorporation preferentially occurs at methylated cytosines on the template DNA strand, suggesting a co-transcriptional feedback to siRNA biogenesis by DNA methylation to reinforce silencing locally. These results highlight an unusual mechanism of Pol IV transcription and suggest a "one precursor, one siRNA" model for the biogenesis of 24-nt siRNAs in Arabidopsis.

INTRODUCTION

In Arabidopsis, 24-nucleotide (nt) siRNAs are the triggers for RNA-directed DNA methylation (RdDM), which plays central roles in repressing transposable elements (TEs) and maintaining genome integrity (Law and Jacobsen, 2010; Matzke and Mosher, 2014). The current model for 24-nt siRNA biogenesis is composed of several sequential steps. First, Pol IV recognizes heterochromatic regions, in part via SAWADEE HOMEODOMAIN HOMOLOG 1 (SHH1) (Law et al., 2013), and transcribes precursor RNAs. These precursor RNAs are then thought to be processed by RNA-DEPENDENT RNA POLYMERASE 2 (RDR2) to form double-stranded RNAs (dsRNAs). The dsRNAs are primarily cleaved by DICER-LIKE 3 (DCL3) to produce 24-nt siRNAs (Matzke and Mosher, 2014). Although Pol IV is clearly required for the biogenesis of 24-nt siRNAs (Herr et al., 2005; Kanno et al., 2005; Onodera et al., 2005; Pontier et al., 2005) and was shown to be transcriptionally active *in vitro* (Haag et al., 2012), the nature and characteristics of Pol IV-dependent transcripts remain poorly understood. The low transcription level at silent Pol IV loci and the efficient downstream processing by dicers has made Pol IV transcripts difficult to study. Recently, regions containing Pol IV-dependent transcripts were described in a *dcl2/3/4* triple mutant that compromises downstream processing of siRNAs (Li et al., 2015). In that study, with the conventional assumption that Pol IV transcribes long precursors, RNA was fragmented before cloning and sequenced reads were assembled into contiguous regions that corresponded to well-known siRNA producing regions (Li et al., 2015). In the present study, we sought to identify the very low abundance Pol IV-dependent transcripts in a wild-type background, and to precisely characterize their start and end positions. We utilized a new method for

RNA sequencing named PATH (Parallel Analysis of Tail and Head) that utilizes RNA adapters to capture both ends of any RNA greater than 27-nt with a 5' monophosphate and a 3' hydroxyl. Using PATH, we efficiently cloned Pol IV-dependent RNAs (P4RNAs) and found that they were only on the order of 30 to 40 nt in length. These P4RNAs mirror 24-nt Pol IV-dependent siRNAs in distribution, abundance, strand bias, and 5'-adenine preference, suggesting that they are direct precursors that often determine both the start position and strandedness of siRNAs. We also observed extensive incorporation of 3'-nontemplated nucleotides preferentially at methylated cytosines on the template DNA strand, suggesting that Pol IV misincorporates and terminates at DNA methylated sites. Our results support a "one precursor, one siRNA" model for the biogenesis of Pol IV-dependent 24-nt siRNAs in Arabidopsis that creates a positive feedback loop between siRNA biogenesis and DNA methylation.

RESULTS

Identification of short RNAs derived from Pol IV-dependent 24-nt siRNA loci.

We utilized an RNA cloning scheme called PATH that uses RNA ligation and gel-based size selection to capture RNAs larger than 27 nt that contain a 5' monophosphate and a 3' hydroxyl (Fig 1a, see Methods). From a wild-type Columbia (Col) library, around 162 million PATH reads were obtained, of which ~50 million could be mapped uniquely to the Arabidopsis genome (Table S1), and PATH reads with a minimal length of 27 nt were used for further analysis. Although the majority of PATH reads matched to structural RNAs (t/r/sn/snoRNA) (Table S1), a large number of reads mapped to previously defined Pol IV siRNA loci (Law et al., 2013) (Fig. 1a and Table

S2). In addition to their co-localization, the abundances of siRNAs and PATH reads were also highly positively correlated (Fig. 1b). At Pol IV siRNA loci, in addition to the 24-nt siRNA peak that is still present in PATH libraries (Fig. 1c), we observed a secondary peak of PATH reads (P4RNAs) that primarily ranged from 30 to 40 nt (Fig. 1c). The size distribution of P4RNAs was different from other PATH reads in the libraries such as reads matching tRNAs and snoRNAs (Fig. S1a). The short length of P4RNAs fits well with the specific preference of DCL3 for short (30 to 50 base pairs, bp) dsRNA substrates (Nagano et al., 2014), consistent with the hypothesis that these short P4RNAs serve as the precursors for siRNA biogenesis.

Biogenesis and processing of Pol IV RNAs

To investigate the biogenesis of P4RNAs, we constructed small RNA and PATH libraries from null mutants of *NRPD1* (encoding the largest subunit of Pol IV), *NRPD*/*E2* (encoding the shared second-largest subunit of both Pol IV and Pol V), *RDR2*, as well as from the double mutants of *DCL3 NRPD1* and *DCL3 RDR2*. We found that both siRNAs and P4RNAs were eliminated in all these mutant backgrounds, indicating a complete dependency of P4RNA biogenesis on both Pol IV and RDR2 (Fig. 2a). Although the loss of P4RNAs in *rdr2* mutants is somewhat unexpected given the assumed downstream function of RDR2 (Haag et al., 2012), this result is consistent with a previous report (Li et al., 2015), and suggests that *in vivo*, Pol IV and RDR2 activities are tightly coupled. These results are also consistent with the earlier observation that RDR2 is in tight association with Pol IV subunits *in vivo* (Haag et al., 2012; Law et al., 2011).

To test if P4RNA biogenesis is dependent on downstream processing events by dicer proteins, we carried out additional siRNA and PATH library analysis from plants triply mutant for *DCL2*, *3*, and *4*. Consistent with a recent report detecting Pol IV dependent transcripts in a *dcl2/3/4* mutant background (Li et al., 2015), we observed a dramatically increased accumulation of P4RNAs in the triple mutant *dcl2/3/4* (Fig. 2b). In contrast, siRNAs levels were dramatically decreased in *dcl2/3/4* at almost all loci (Fig. 2b, c) and their sizes were shifted from predominantly 24 nt to predominantly 21 nt (Fig. S1b). In addition, we observed that the abundances of the remaining siRNAs in *dcl2/3/4* were tightly correlated with the abundances of siRNAs in wild-type (Fig 2c). We also found that the abundances of P4RNAs in *dcl2/3/4* were strongly positively correlated with the abundances of siRNAs in Col (Fig. S1c). As one example, P4RNAs and siRNAs showed a very similar enrichment at the boundaries of long transposable elements (Fig. 2d). The accumulation of P4RNAs but reduction of siRNAs in *dcl2/3/4* mutants are inconsistent with the interpretation that P4RNAs are merely longer, misprocessed siRNAs, and suggests that P4RNAs are indeed the precursors of Pol IV-dependent siRNAs.

We also analyzed siRNA and PATH libraries from other *DCL* mutant combinations (*dcl2/4, dcl3, dcl2/3,* and *dcl3/4*) and compared these with *dcl2/3/4*. As previously reported (Henderson et al., 2006), all *dcl3* mutant combinations showed a shift in siRNA size from 24 nt to 21 and/or 22 nt (Fig. 3a, top panel). PATH sequencing indicated that the over-accumulation of P4RNAs was negligible in *dcl2/4*, higher in *dcl3, dcl2/3,* and *dcl3/4*, and highest in *dcl2/3/4*, indicating that the P4RNA processing by DCL2 and DCL4 is less efficient than by DCL3 (Fig. 3a, bottom panel). As a secondary

confirmation of the accumulation of P4RNAs in *dcl3* mutants, we examined small RNA northern blots for the presence of 30- to 40-nt signals. Using LNA probes to six different siRNA producing loci we indeed observed a RNA smear with larger sizes, along with an accumulation pattern (highest in *dcl2/3/4*) that matched the trend of the P4RNA level in our PATH libraries (Fig. 3b, c) (Henderson et al., 2006). In addition, we used a reverse transcription quantitative PCR (real-time RT-PCR) approach to verify the very short nature of the P4RNAs. Because the peak of P4RNA abundance is 30 - 40 nt, whereas longer P4RNAs are much less abundant (Fig. 1c), we reasoned that primers spaced about 40 nt from each other should amplify P4RNAs much more efficiently than those spaced slightly further apart. After normalizing for efficiency of DNA amplification, we observed a 100-fold higher level of PCR product when comparing a 36 bp versus a 58 bp amplicon at one locus, and 50-fold higher level of PCR product when comparing a 39 bp versus an 81 bp amplicon at a second locus (Fig. S2a).

To directly test whether the longer P4RNAs might simply be misprocessed siRNAs that are loaded into AGO4, we performed AGO4 RNA-IP (RIP) in the *dcl2/3/4* genetic background using antibodies against endogenous AGO4, and then characterized AGO4-associated RNAs with high-throughput sequencing. The AGO4 RIP result showed that even though there is a massive accumulation of longer P4RNAs and a reduction of 24nt siRNAs, AGO4 still selectively binds to the remaining 22-24 nt siRNAs but not the longer P4RNAs (Fig. S2c), strongly supporting our model that P4RNAs are the precursors of the 24-nt siRNAs.

Even though double-stranded RNAs are thought to be intermediates of siRNA biogenesis, it is known that at many loci in the genome, siRNAs are predominantly found

on one strand of the genome but not the other (Lister et al., 2008; Zhong et al., 2014). In addition, these strand-biased clusters of siRNAs correspond to strand biased DNA methylation (Lister et al., 2008; Zhong et al., 2014). At these strand-biased siRNA clusters, we also observed strongly strand-biased P4RNAs in both Col and *dcl2/3/4* (Fig. 1d and S1d). These results suggest that our PATH libraries are primarily composed of single-stranded, Pol IV-derived strands, but not the RDR2-derived second strand. In support of this hypothesis, we used RT-PCR to amplify these strand-biased clusters, and indeed found that Pol IV/RDR2-dependent transcripts could be amplified equally from both strands (Fig. S3). This suggests that RDR2 strands are likely present *in vivo*, but are not cloned by the specific PATH sequencing technique employed in this study. The observation that P4RNAs and siRNAs show the same strandedness suggests that the Pol IV-derived strands, rather than the RDR2-derived strands, are strongly favored to become the final 24-nt siRNA products.

In summary, results from siRNAs and PATH libraries of mutants deficient in siRNA biogenesis or processing, together with the shared strandedness of P4RNAs and siRNAs, suggest that the 30 to 40 nt P4RNAs serve as precursors to Pol IV dependent siRNAs.

Pol IV transcription initiates at Pol II like TSSs and favors 5'-adenine.

Next we investigated the nature of the 5' ends of P4RNAs. Given that our cloning method only captures RNAs with 5-monophosphates (Fig. 1a), we sought to measure the proportion of P4RNAs containing this type of 5' end. To address this, we used Terminator exonuclease to preferentially digest RNAs with a 5'-monophosphate. We

subsequently measured the abundance of the remaining P4RNAs by real-time RT-PCR. Consistent with a recent study (Li et al., 2015), Terminator treatment degraded the majority of P4RNAs at all loci tested (Fig. S2b). Thus, while it is possible that subpopulations of P4RNAs have other end structures such as 5'-triphosphates, 5'-caps, or 5'-hydroxyl groups, it appears that the majority of P4RNAs contain 5'-monophosphates.

Because our method did not include a fragmentation step that is typical in RNA-seq library protocols (Li et al., 2015), it was possible to detect the 5' nucleotide of P4RNA reads. We observed a strong enrichment of T/C (Y) at the -1 position (the nucleotide immediately upstream of the first nucleotide of the P4RNA read) and A/G (R) at the +1 position (beginning nucleotide of the read) (Fig. 4a). Further, the four possible Y/R dinucleotides at the -1/+1 positions were by far the most enriched dinucleotides at the 5' end of P4RNAs (Fig. 4b). This pattern is very similar to that known for the transcriptional start sites (TSSs) of RNA Polymerase II (Pol II) in plants and other organisms and it is referred to as the "Y/R rule" (Cumbie et al., 2015; Nechaev et al., 2010; Yamamoto et al., 2007). This result suggests that Pol IV has retained this preference from its evolutionary ancestor Pol II (Ream et al., 2009), and that the 5' ends of P4RNAs likely represent Pol IV transcriptional start sites. The short TSS like sequences at the 5' ends of P4RNA, along with their very short nature does not support previous models in which Pol IV initiates transcription solely at the nucleosome-depleted promoter regions near the ends of transposons to produce long transcripts (Li et al., 2015). Instead, our results suggest that Pol IV can initiate transcription at many positions that resemble the Y/R features of Pol II TSSs, transcribing many short P4RNAs along the length of transposons.

Because Pol IV transcripts feature Pol II-like TSSs, we also performed genome-wide profiling of Pol II occupancy in Arabidopsis via ChIP-seq in wild-type and different mutant backgrounds (*nrpd1* and *dcl2/3/4*), and compared this with Pol IV ChIP-seq (Law et al., 2013). Our results showed that Pol II does not appear to access Pol IV loci even in the absence of Pol IV (*nrpd1*) (Fig. S4a). Furthermore, whole-genome bisulfite sequencing and small RNA sequencing of floral tissues from the weak Pol II mutant (*nrpb2-3*) did not reveal obvious changes in either DNA methylation or siRNA biogenesis (Fig. S4b, c), which is consistent with our previous analysis of *nrpb2-3* using leaf tissue (Stroud et al., 2013). Therefore despite the proposed crosstalk between Pol II and the RdDM pathway (Zheng et al., 2009), our data suggest that Pol II and Pol IV occupy distinct territories on the genome.

Arabidopsis 24-nt siRNAs are primarily loaded into AGO4, and are strongly biased toward having a 5'-adenine, which was previously shown to involve an AGO4 loading preference (Havecker et al., 2010; Mi et al., 2008). Interestingly, we found that P4RNAs, like siRNAs, also show a strong enrichment for 5'-adenine (Fig. 4c, and S5a, b). Because the size of P4RNAs is approximately 30 to 40 nt, on average only one 24-nt siRNA duplex could be processed from each of these P4RNA precursors. This fact, coupled with the shared 5' adenine preference and the shared strand preference, suggests that 24-nt siRNAs are preferentially cleaved from the 5' portion of P4RNAs. Consistent with this hypothesis, DCL3 was shown to prefer short double-stranded RNAs (30 to 50 bp) that contain a 5' adenine (Nagano et al., 2014). Therefore, our results favor a scenario in which the 5'-adenine preference of P4RNAs likely contributes to the 5'-adenine preference of Pol IV siRNAs. In addition, our results provide a plausible explanation as

to why AGO4 evolved to bind siRNAs with 5' adenine. Taken together, the short size and 5'-A feature of Pol IV transcripts may help to channel their processing to DCL3 rather than dicer proteins in other silencing pathways, and thus lead to production of predominantly 24-nt siRNAs at Pol IV transcribed loci.

Pol IV transcription preferentially terminates at methylated cytosines with misincorporated nucleotides.

We analyzed the sequence composition of P4RNA reads with perfect match to genome and found enrichment for A, C, and U at last three positions of the 3' end (Fig. S5c). It is not known whether Pol IV tends to cease transcription at this sequence, or whether it might transcribe a longer RNA that is then processed by an unknown endonuclease. We found a similar compositional bias at the 3' end of Pol IV-dependent siRNAs (Fig. S5d), although the magnitude of the biases were lower for siRNAs, suggesting that P4RNAs are processed at some level at their 3' ends to produce siRNAs. Despite the ACU enrichment at the 3' end of siRNAs, they still show enrichment for 5' adenine, which is likely explained by the AGO4 preference for loading siRNAs with a 5' adenine. These results provides additional evidence that the P4RNAs described here are indeed the precursors of siRNAs, and it is again consistent with (1) the shared strandedness of P4RNAs and siRNAs, and (2) the hypothesis that the Pol IV strand, rather than the RDR2 strand, is favored as the final siRNA.

We also performed an analysis in which we allowed multiple mismatches during genome mapping of P4RNAs. Interestingly, we found that more than half of the P4RNAs contained one or two non-templated nucleotides at their 3'-ends (Fig. 5a, b). In contrast,

in the same PATH libraries, reads derived from Pol II-transcribed coding regions, or from microRNA processing intermediates, had very few mismatches, and these mismatches were not localized to the 3' ends (Fig. S6a, b). To further rule out the possibility that P4RNA 3' non-templated nucleotides are due to lower quality of sequencing toward the end of the read or result from incomplete trimming of adapter sequence, we analyzed the second read (read2) from the paired-end sequencing. On read2, the beginning nucleotide corresponds to the 3' end nucleotide of RNA, where base quality is high and there is no trimming step involved, and we observed the same high-level of non-templated nucleotides at the 3'-end of P4RNA (Fig. S6c).

The 3' end base composition of P4RNAs containing non-templated nucleotides was quite different from those with a perfect genome match (Fig. S5c, e), suggesting that P4RNAs with non-templated nucleotides terminate by a different mechanism than those without. On the other hand, the 5' end base composition of reads with non-templated nucleotides was very similar to those without, suggesting that both classes may share a similar transcription initiation mechanism by Pol IV (Fig. S5f, g). All four nucleotides were found among the non-templated nucleotides, although there was some preference for guanines (Fig. S7a). In addition, we observed different preferences for non-templated nucleotides depending on the sequence that should have been present, suggesting that the preference for a particular non-templated nucleotide is determined by the sequence of the DNA template for Pol IV (Fig. S7b, c). Because all four nucleotides were present, and because different incorrect nucleotides were present depending on the template DNA sequence, it seems most likely that these nucleotides arise from misincorporation during Pol IV transcription rather than from the activity of a terminal transferase.

Because misincorporation of nucleotides occurred most frequently at positions corresponding to guanines (which would be cytosines on the DNA template) (Fig. 5a), we hypothesized that misincorporation might be caused by in part by cytosine DNA methylation. The most highly methylated sequences in the Arabidopsis genome are CG dinucleotides (Cokus et al., 2008; Lister et al., 2008), and Pol IV siRNA loci targeted by RdDM are usually heavily methylated at most CG sites (Stroud et al., 2013) (Fig. S4b). Agreeing with the DNA methylation hypothesis, we found that CG dinucleotides (G being the last nucleotide of the RNA) exhibited by far the highest enrichment of misincorporation amongst the 16 possible dinucleotide sequences (Fig. 5c). In addition, CG dinucleotides were strongly enriched at the 3'-end of P4RNAs that exhibited misincorporation (Fig. 5d). The second group of most commonly methylated sequences in the genome and at Pol IV siRNA loci are CHG sites (Fig. S4b) (Cokus et al., 2008; Lister et al., 2008). A trinucleotide analysis showed that all of the trinucleotides showing the strongest tendency for misincorporation were those that contained CG sites (Fig. S7d). In addition, CHG sites showed a strong tendency for misincorporation, which was higher than AHG, THG, or GHG sites (Fig. S7d).

To directly test whether the loss of methylation can alter the pattern of 3' end misincorporation, we analyzed the *ddm1 (decrease in DNA methylation 1)* mutant that exhibits a severe loss of methylation in heterochromatin (Matzke and Mosher, 2014). Despite the significant loss of methylation, many silent loci are still producing 24-nt siRNAs in *ddm1* (Colome-Tatche et al., 2012), suggesting that Pol IV is still largely functional in *ddm1*. Because P4RNAs are elevated in *dcl3* mutants, we utilized a small RNA dataset from *ddm1 dcl3* double mutant (RNAs smaller than 200 nt) (McCue et al.,

2015) and compared these with a similar dataset from the *dcl3* single mutant. We first focused our analysis on a set of strong *ddm1* hypomethylated CG DMRs (Differentially Methylated Regions) (Fig. 6a, see methods) from whole genome bisulfite sequencing data of *ddm1* (Creasey et al., 2014). We found that the enrichment of P4RNA 3'-misincorporation at CG dinucleotides was eliminated in *ddm1 dcl3* compared to the *dcl3* single mutant (Fig. 6b). We also examined trinucleotide enrichments in *ddm1 dcl3* at a set of CHG DMRs (Fig. S7e), and observed a significant reduction at CHG but not at AHG, THG, or GHG (Fig. S7f). Moreover, P4RNAs in *ddm1/dcl3* were slightly longer than those in *dcl3* single mutant (Fig. S7g). Take together, these results suggest that DNA methylation itself is contributing to the pattern of 3'-misincorporation.

If P4RNAs are the precursors of siRNAs and DCL3 can cleave to some extent from the 3'-end of P4RNA, siRNAs should also contain some level of misincorporated bases at their 3' ends. By allowing for mismatches during genome mapping, we indeed found that siRNAs contain non-templated nucleotides that were enriched at the 3' end (Fig. S6d). However, the proportion of siRNAs with 3'-mismatches (~1%) was far lower than that of P4RNAs (~50%, Fig. 5b), suggesting that far fewer siRNAs are processed from the 3' ends of P4RNAs than from the 5' ends, possibly due to the preference of 5'-adenine by DCL3 (Nagano et al., 2014). We also observed enrichment of CG sites at the 3' end of siRNAs that showed misincorporation (Fig. S6e), again consistent with some level of processing of the 3' end of P4RNAs into siRNAs.

Since DCL3 appears to process from both ends of the double-stranded Pol IV/RDR2-derived RNA, and since P4RNAs are enriched for adenines at their 5' ends and misincorporated nucleotides at their 3' ends (Fig. 7a), two predictions are that siRNAs

with 5’ adenines should have lower than average 3’ misincorporation, and siRNAs with 3’ misincorporation should have lower than average 5’ adenine content. Indeed, we found that siRNAs with 5’ adenines had 50% lower misincorporation than siRNAs with other 5’ nucleotides (Fig. 7b), and siRNAs with misincorporated nucleotides showed a lower 5’ adenine content than those with perfect matches to the genome (Fig. 7c). These results further support that P4RNAs containing misincorporated 3’ nucleotides are processed into siRNAs. In summary, our results support that three different mechanisms can contribute the formation of the 3’ end of P4RNAs - termination or 3’ end processing at sequences enriched for ACU at the last three positions of the P4RNA, termination associated with misincorporation at methylated cytosines, and termination associated with misincorporation at other nucleotides.

DISCUSSION

The results of this study support the general scheme that P4RNAs are first transcribed by Pol IV, made double stranded by RDR2, and diced by DCL3 and other dicers to make the siRNAs which are loaded into AGO4. DCL3 cleavage produces an siRNA duplex with symmetric structure (Matzke and Mosher, 2014), but only one strand (the guide strand) is retained in AGO4, while the other strand (the passenger strand) is cleaved by AGO4 and degraded (Ye et al., 2012). Little is known about guide-strand selection of 24-nt siRNAs, and it remains unclear why, at many loci, siRNAs predominately match to one strand, and direct strand-specific methylation (Zhong et al., 2014). Our analysis revealed that the strand biases for P4RNAs and siRNAs are highly positively correlated (Fig. 1d and Fig. S1d), suggesting that the siRNAs derived from the

Pol IV strand rather than the RDR2-synthesized complementary strand, are favored as the guide-strand siRNA. This bias appears to be partially accomplished by a strong bias for adenine to be present as the first nucleotide of P4RNAs, by the preference of DCL3 for 5' adenines for siRNA processing, and by the preference of AGO4 for loading siRNAs with a 5' adenine (Fig 7a). P4RNAs also appear to be processed to some extent at their 3' ends to yield siRNAs that have similar 3' end signatures as are present in P4RNAs, including either an enrichment for ACU sequences, or misincorporated nucleotides. In addition, Pol IV transcripts are so short that on average only one siRNA will arise from each P4RNA. Finally, the short nature of Pol IV transcripts, coupled with the preference of DCL3 for short dsRNAs, may serve to channel Pol IV/RDR2 products into DCL3 rather than other dicers, such as DCL4 that prefers long dsRNAs (Nagano et al., 2014). These results imply a "one precursor, one siRNA" model for processing of siRNAs from Pol IV to RDR2 to DCL3 to AGO4 (Fig. 7a). Based on the patterns of TSSs, strand-bias and 3'-misincorporation we conclude that the majority of PATH reads at siRNA loci are P4RNA, but it is still possible that some of these PATH reads are RDR2-transcribed as they are genetically dependent on both RDR2 and Pol IV.

Because Pol IV transcripts are so short, the sites of Pol IV transcriptional initiation and termination are largely determining the positions of siRNAs in the genome. Production of short transcripts by Pol IV could provide for siRNA biogenesis while avoiding the risk of transcribing full-length transposable elements. In addition, the short length of Pol IV transcripts may help to prevent spreading of RdDM to flanking regions, and allow specific silencing of transposons that are close to genes.

Our proposed model for biogenesis of 24-nt siRNAs in Arabidopsis share characteristics with that of 21U-RNAs (or piRNAs) in *C. elegans.* 21U-RNAs are 21-nt in length and begin with a 5' uridine, and they are autonomously expressed from thousands of loci dispersed in two broad regions of chromosome IV (Batista et al., 2008; Ruby et al., 2006). Like P4RNAs, precursors of 21U-RNAs, which are ~26 nucleotide Pol II transcripts, are terminated by an unknown mechanism to produce unusually short transcripts (Gu et al., 2012). However, unlike P4RNAs, 21U-RNA precursors are processed at their 5' ends to remove two nucleotides, and transcription units usually contain a conserved motif 42-nt upstream of the mature piRNA (Batista et al., 2008; Gu et al., 2012; Ruby et al., 2006). In contrast we did not find evidence for sequence conservation upstream of P4RNA start sites (Fig. S4d), and instead Pol IV appears to be recruited by epigenetic signals such as the binding of H3K9 methylation through the Pol IV interacting protein SHH1 (Law et al., 2013). Furthermore, while there is clear evidence for transposon-derived secondary siRNA production through RNA-dependent RNA Polymerase (RdRP) activity in *C. elegans* (22G-RNA) (Lee et al., 2012), as well as Zucchini-dependent, secondary phased piRNAs in mammals (Han et al., 2015; Mohn et al., 2015), there is little evidence to support the possibility of secondary siRNA production in the RdDM pathway.

The 5' ends of P4RNAs are very similar in sequence composition to that of Pol II transcripts, showing strong enrichment for Y/R dinucleotides at the -1/+1 positions. However, as opposed to Pol II transcripts that begin with trimethylguanosine caps, the majority of P4RNAs have 5' monophosphates. Since RNA polymerases normally start transcribing with a triphosphate containing nucleotide, it is unclear how Pol IV

transcripts acquire a 5' monophosphate. It is possible that Pol IV is able to initiate transcription with a nucleoside monophosphate as previously reported for some RNA polymerase (Martin and Coleman, 1989; Ranjith-Kumar et al., 2002), or the P4RNAs may be initially capped after which a de-capping enzyme converts the trimethylguanosine cap to a 5' monophosphate, or an unknown polyphosphatase-like enzyme may directly convert the 5′-triphosphate to a 5′-monophosphate. Additionally, since 5' monophosphate containing RNAs are often a target of 5' to 3' exonucleases, there may be mechanisms to protect P4RNA 5' ends until subsequent processing steps are completed. It is also unclear why P4RNA evolved to have 5' monophosphates rather than cap structures, but one possibility is that this helps the cell avoid inadvertently mistaking a P4RNA for a Pol II transcript in order to avoid translation of transposon RNAs.

Our observations suggest that a component of the mechanism by which Pol IV transcription terminates is that DNA methylation on the template strand causes misincorporation of inappropriate bases. The mechanism by which this happens is unclear, but it is known that DNA methylation can cause transcriptional elongation defects in *Neurospora* (Rountree and Selker, 1997). The preferential termination of Pol IV transcription near DNA methylation would promote siRNA generation near sites of preexisting DNA methylation, thereby creating a self-reinforcing loop in which siRNAs direct DNA methylation targeting and DNA methylation helps direct the location of further siRNA production.

Experimental Procedures

Biological materials

The mutant alleles of *nrpd1-4*, *nrpb2-3*, *rdr2-1, nrp(d/e)2*, *rdr2/dcl3*, and combinations of *dcl2/3/4* used in this study were in the background of *Arabidopsis thaliana* ecotype Columbia-0 and have been previously described (Henderson et al., 2006; Li et al., 2008; Pontier et al., 2005; Xie et al., 2004; Zheng et al., 2009). The *dcl3/nrpd1* double mutant was obtained by crossing *dcl3-1* with *nrpd1-4*. Plants were grown in a growth chamber with 16 hour of light or greenhouse condition for five weeks. Immature inflorescence tissues including inflorescence meristem and early stages floral buds (up to stage 11/12) were collected.

Sequencing of small RNA, PATH mRNA, and BS-seq libraries

Total RNA was first treated with RiboMinus™ Plant Kit for RNA-Seq (Invitrogen A10838-08) to remove rRNA, followed by size selection of RNA on a 15% UREA TBE Polyacrylamide gel (Invitrogen, EC6885BOX). Gels containing RNA with size between 15- to 27-nt were kept for small RNA library, while gels containing 28- to ~300-nt RNA were kept for PATH library. After gel elution, library construction for both sRNA and PATH was done using the Illumina TruSeq Small RNA Sample Preparation Kit (RS-200-0012), except that at the final size selection step, PCR products were separated on a 6% TBE Polyacrylamide gel (Invitrogen, EC6265BOX) and selected for the range from 120- to ~1000-bp for PATH library. Gel-eluted PCR products with different TruSeq index sequences were pooled and sent for Illumina sequencing. The

mRNA library was constructed using the Illumina TruSeq RNA Sample Preparation Kit (RS-122-2001) according to the standard manual. PATH libraries were sequenced using either paired-end mode with length of read1 being 120-bp and length of read2 being 30-bp (PE120+30), or single-end 100-bp (SE100); while sRNA and mRNA libraries were sequenced with single-end 50-bp (SE50). For BS-seq, DNA was isolated using the DNeasy Plant Mini Kit (Qiagen #69104) according to manufacturer instructions and quantified using the Qubit dsDNA High Sensitivity Kit (Life Technologies #Q32851). Libraries were constructed with 30ng DNA using the Ovation Ultralow Methyl-Seq Library Systems (NuGEN #0335). Bisulfite conversion was done using the EpiTect Bisulfite Kit (Qiagen # 59104). BS-seq Libraries were sequenced at a length of 50 bp. All sequencing was carried out on Illumina HiSeq machines at the Broad Stem Cell Research Center (BSCRC) sequencing core at University of California, Los Angeles.

Realtime RT-PCR and RNA gel blot

Realtime RT-PCR to detect P4RNAs in various genotypes was performed as the following: Total RNA were extracted from 100 mg of flowers using Trizol (Ambion) with an extra step of 24:1 Chloroform : isoamyl alcohol to remove remaining phenol prior to isopropanol precipitation, the resuspended RNAs were cleaned up using Quick-RNA miniprep (Zymo research, USA) for further purification and complete gDNA removal. Then 5ug of the purified RNA from each sample was used for RT reaction with SuperScript III first-strand kit (Invitrogen, USA). 1uL of the RT reaction was used for real time PCR using iQ SYBR green supermix (Biorad, USA). We used terminator exonuclease (Epicenter, USA) for the removal of RNA with 5'-monophosphate: 5ug of

total RNA were treated for an hour at 30 degree with Terminator, and the control, non-treated, RNA with the same buffer and 50% glycerol instead of terminator exonuclease enzyme. After Terminator treatment the RNA were used for RT reaction as described above. Primer information can be found in Table S3.

RNA gel blotting was performed as previously described (Henderson et al., 2006) with LNA probes for detecting transposon regions and regular RNA probes for detecting 5S locus. Probe information can be found in Table S3.

Pol II ChIP-seq and AGO4-RIP

ChIP-seq was performed as described previously (Johnson et al., 2014). 5 ug Pol II antibodies (Abcam #ab817) were used for each ChIP. Libraries for Pol II ChIP-seq were generated using the Ovation Ultralow DR Multiplex System (NuGen #0330) and sequenced at a length of 50 bp. ChIP-seq data was visualized using ngsplot (Shen et al., 2014).

AGO4-RIP was performed as previously described (Ji et al., 2011) with commercial AGO4 antibody (Agrisera #AS09 617). RNA isolated from RIP was used for library construction with Illumina TruSeq Small RNA Sample Preparation Kit (RS-122-2001) according to the standard manual.

Bioinformatic analysis

Data handling

In general, qseq files received from the sequencing core were demultiplexed with an in-house Perl script and converted to fastq files for downstream analysis. For small RNA and PATH data, original reads were first trimmed using Cutadapt (v1.4), then mapped to the reference TAIR10 genome using Bowtie (Langmead et al., 2009) allowing only one unique hit (-m 1). We allowed zero mismatch for small RNA mapping (-v 0) and up to three mismatches for PATH mapping (-v 3). mRNA data were mapped using Tophat (Trapnell et al., 2009) allowing two mismatches and only one unique hit.

Analysis of siRNAs and P4RNAs

We used a list of Pol IV dependent siRNA loci that is previously described(Law et al., 2013) and manually inspected and filtered out a few loci that are tRNA related (listed in Table S2). In brief, these are 200 base pair bins where siRNAs were significantly reduced in a Pol IV mutant compared to two replicates of wild-type controls (FRD < 10^{-10}). More details can be found in the "Identification of siRNA clusters" section of Methods in the Law et al. paper (Law et al., 2013). Our definition of P4RNAs is any 27+ nt PATH reads derived from these ~7000 previously defined Pol IV dependent 24-nt siRNA loci. For the abundance calculation, sRNA reads with the length between 18 and 26 nt, and PATH reads with the length greater or equal to 27 nt were included (Table S1). 5'-Adenine preference was analyzed by calculating the nucleotide composition at each position (counting from 5'end) for Pol IV siRNAs and P4RNAs. Percentage of 3'-nontemplated nucleotides in P4RNAs was calculated at each position by analyzing the mapping results with a custom Perl script with the focus on the "MD:Z" column in sam format. The length of the 3'-misincorporation was done using the seed mapping option of bowtie to

map the first 24 nt of P4RNA perfectly and then allow up to ten mismatches for 3' portion. This analysis confirmed that majority of P4RNAs carry one or two 3'-misincorporated nucleotides. Therefore we chose to allow three mismatches for P4RNA mapping for all analysis.

Normalization of P4RNA and sRNA abundance

Abundances of P4RNA in each library were normalized to the total number of sequenced reads, and abundances of siRNAs were normalized to the sum of all miRNAs in each library. For example, in Figure 2b, in Col PATH library we obtained 161,241,523 reads in total and of which 63,739 were classified as sasRNA (27+ nt), while in dcl2/3/4 PATH library we obtained 89,226,121 reads in total and of which 4,283,063 were classified as sasRNA. Therefore the percentage of P4RNA in dcl2/3/4 compared to Col is calculated as percentage = (4,283,063/89,226,121) / (63,739/161,241,523) = 121.43 $\approx$ 120 fold.

BS-seq analysis and DMR calling

Analysis of the floral BS-seq libraries of wild-type and *ddm1* mutant (Creasey et al., 2014) were performed using BSMAP (Xi and Li, 2009), allowing only uniquely mapped reads and discarded sibling PCR products, with the tolerance of 2 mismatches per 50 bp. DMR calling was performed as previously described (Stroud et al., 2013), with a more stringent criteria: 1) sum of all sequenced cytosines in the 100bp bin need to be at least 100; 2) the difference in CG methylation at each bin needs to be at least 0.5. This filtering allows us to focus on regions that are not only highly methylated in wild-type but also lose methylation dramatically in the *ddm1* mutant. CHG DMRs were filtered with similar

criteria, with the minimal loss of CHG in *ddm1* compared to wild-type being 0.4, and the count of covered cytosines in each bin no less than 50.

Nucleotide composition and enrichment

P4RNAs a defined set of loci were measured for their 3'-end sequence composition. For those P4RNAs that contain the 3' misincorporation, different compositions of the reference sequence at the first mismatched nucleotide (mono-nucleotide), the last matched plus first mismatched (di-nucleotide), or the last two matched plus first mismatched (tri-nucleotide) were counted as the observed value. The mono-, di- or tri-nucleotide compositions of the DNA sequence (or K-mer) were calculated using Jellyfish (Marcais and Kingsford, 2011) and considered as expected value assuming P4RNAs end equally likely on any nucleotide. The observed value and expected value were both normalized to their own population, and the ratio of observed/expected is used to show the relative enrichment of certain mono-, di- or tri-nucleotide composition.

Accession codes

Sequencing data have been deposited at GEO (GSE61439). We used public data of floral *ddm1* and wild-type BS-seq (GSE52346), *ddm1/dcl3* sRNA-seq (GSE57191), and *dcl3* sRNA-seq (GSE62801).

Acknowledgments

We thank members of the Jacobsen lab for insightful discussions. We thank Molly Megraw for suggestions on TSS analysis, and Mahnaz Akhavan for technical assistance.

J.Z. thanks Dr. Chomdao Chommy and Seksan Sapsubbsakul for encouragement. High throughput sequencing was performed at the UCLA BSCRC BioSequencing Core Facility. This work was supported by NIH grant GM60398 to S.E.J and NSF award 1051576 to B.C.M. S.B. is supported by a postdoctoral fellowship of the Swiss National Science Foundation. J.G-B is a Human Frontiers Science Program fellow (LT000425/2012-L). J.Z. is a Life Science Research Foundation postdoctoral fellow, sponsored by the Gordon and Betty Moore Foundation. S.E.J. is an Investigator of the Howard Hughes Medical Institute.

Author Contributions

J.Z., S.B., S.F., I.A, I.H., X.C., and C.T. performed experiments. T.L., S.F., and J.G-B. provided materials. S.P., L.L., and X.C. participated in the genetic experiments. S.E.J. and B.C.M. oversaw the study. J.Z., H.W., and W.L. analyzed data. J.Z. and S.E.J. designed the study and wrote the manuscript.

Figure Legends

Fig. 1. Identification of siRNA-loci associated short RNAs (P4RNAs).

a. Procedure for the construction and analysis of PATH libraries. An example siRNA locus is shown at the bottom with IGV screenshots of P4RNAs and siRNAs matching to that region.

b. Abundances are highly correlated of PATH reads (27+ nt) and sRNA reads (18 to 26 nt) in wild-type Col from previously defined Pol IV siRNA loci (Law et al., 2013).

c. Size distribution plots of all PATH reads and sRNA reads in Col from Pol IV siRNA loci. A distinct peak at 30 to 40 nt can be seen in the PATH library; we named these "P4RNAs".

d. P4RNAs and siRNAs share the same strand bias at Pol IV siRNA loci. Only Pol IV siRNA loci matched by more than 100 P4RNAs in Col were selected to obtain a robust calculation of strandedness. The plus-strand ratio was calculated as the abundance of reads matching to the plus strand divided by the total number of reads at that locus.

Fig. 2. **Biogenesis and processing of P4RNAs.**

a. Both Pol IV siRNAs and P4RNAs are eliminated in *nrpd1, nrp(d/e)2, rdr2, dcl3/nrpd1,* and *dcl3/rdr2*. Small RNA abundances were normalized to the sum of all TAIR10-annotated miRNAs in each sRNA library, and then compared to Col;

P4RNA abundances were normalized to the total number of reads in each PATH library.

b. Pol IV siRNAs are reduced in *dcl2/3/4,* whereas P4RNAs accumulate substantially.

c. Abundances of P4RNA and siRNA at each locus in Col and *dcl2/3/4*.

d. At long TEs, P4RNAs resemble the distribution of siRNAs – enriched at promoters and termini of TEs, and also spread into TE bodies.

Fig. 3. **P4RNAs in *dcl* mutant combinations.**

a. Size of Pol IV siRNAs shifted from 24 nt in Col to 21/22 nt in *dcl3*-containing mutants, accompanied by the accumulation of P4RNAs. P4RNA abundances are normalized to that in Col.

b. RNA blot analysis of siRNAs and P4RNAs in *dcl* mutant backgrounds.

c. Size distribution of P4RNAs in *dcl3*-containing mutants.

Fig. 4. **P4RNAs feature Pol II like TSSs and favor 5'-Adenine.**

a. TSSs of Pol IV exhibit preference for C/T at the -1 position and A/G at the +1 position, resembling the "Y/R rule" of Pol II TSS, calculated using P4RNAs from *dcl2/3/4*.

b. Dinucleotide enrichment at -1/+1 of P4RNA, calculated using P4RNAs from *dcl2/3/4*.

c. Both siRNA and P4RNAs from Col have a strong preference for 5' adenine.

Fig. 5. **Pol IV transcription preferentially terminates at methylated cytosines with misincorporated nucleotides.**

a. Example of 3'-nontemplated nucleotides on P4RNAs. The black bar represents 50 bp in length.
b. Length of the 3'-end nontemplated nucleotides, defined by the first mismatched nucleotide to the last nucleotide. If a P4RNA has no mismatch, the length is zero.
c. Di-nucleotide enrichment at the first mismatched position at the 3' end.
d. Frequency of CG dinucleotide on reference sequence over the P4RNAs with misincorporation, "-1" marks the last perfectly matched position, and "+1" marks the first mismatched position. The count of each CG is designated to the position of the G, therefore the peak at "+1" represents a peak of CG at "-1/+1".

Fig. 6. **Pol IV transcription preferentially terminates at methylated cytosines with misincorporated nucleotides.**

a. CG hypomethylated DMRs in *ddm1* compared to a wild-type control.
b. 3'-misincorporated nucleotides of P4RNA at *ddm1* CG DMRs in *ddm1 dcl3* compared to *dcl3*.

Fig. 7. **"one precursor, one siRNA" model for the biogenesis of Pol IV dependent 24-nt siRNAs**.

a. Pol IV transcription is initiated at Pol II-like TSSs. A short RNA of ~30 to 40 nt (with 5'-adenine preference) is produced by Pol IV at heterochromatic regions; misincorporation then occurs at the cytosine position (red bar) and terminates Pol

IV transcription. This process typically yields a P4RNA with 5’ adenine and 3’ misincorporation. P4RNA then goes through processing that involves RDR2 synthesizing the complementary strand (gray), DCL3 cutting from the 5’ end (major) or 3’ end (minor) of the P4RNA, and loading of the P4RNA-derived siRNA strand into AGO4 as the guide-strand. Eventually one P4RNA precursor gives rise to one siRNA that is derived from either its 5’ or 3’ end. (“M” underneath the DNA template indicates DNA methylation at the heterochromatic region.)

b. When DCL3 cuts from the 5’ end of P4RNAs, the resulting siRNAs are more likely to carry a 5’ adenine and be perfectly matched. Indeed a reduction of 3’ mismatches is observed for siRNAs with a 5’ adenine.

c. When DCL3 cuts from the 3’ end of P4RNAs, the resulting siRNAs are more likely to carry the 3’ misincorporation and less likely the 5’ adenine. This is consistent with the observation that siRNAs with mismatch have lower percentage of 5’ adenine.

References

Batista, P.J., Ruby, J.G., Claycomb, J.M., Chiang, R., Fahlgren, N., Kasschau, K.D., Chaves, D.A., Gu, W., Vasale, J.J., Duan, S.*, et al.* (2008). PRG-1 and 21U-RNAs interact to form the piRNA complex required for fertility in C. elegans. Mol Cell *31*, 67-78.

Cokus, S.J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C.D., Pradhan, S., Nelson, S.F., Pellegrini, M., and Jacobsen, S.E. (2008). Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature *452*, 215-219.

Colome-Tatche, M., Cortijo, S., Wardenaar, R., Morgado, L., Lahouze, B., Sarazin, A., Etcheverry, M., Martin, A., Feng, S., Duvernois-Berthet, E.*, et al.* (2012). Features of the Arabidopsis recombination landscape resulting from the combined loss of sequence variation and DNA methylation. Proc Natl Acad Sci U S A *109*, 16240-16245.

Creasey, K.M., Zhai, J., Borges, F., Van Ex, F., Regulski, M., Meyers, B.C., and Martienssen, R.A. (2014). miRNAs trigger widespread epigenetically activated siRNAs from transposons in Arabidopsis. Nature *508*, 411-415.

Cumbie, J., Ivanchenko, M., and Megraw, M. (2015). nanoCAGE-XL and CapFilter: an Approach to Genome Wide Identification of High Confidence Transcription Start Sites. BMC Genomics *in press*.

Gu, W., Lee, H.C., Chaves, D., Youngman, E.M., Pazour, G.J., Conte, D., Jr., and Mello, C.C. (2012). CapSeq and CIP-TAP identify Pol II start sites and reveal capped small RNAs as C. elegans piRNA precursors. Cell *151*, 1488-1500.

Haag, J.R., Ream, T.S., Marasco, M., Nicora, C.D., Norbeck, A.D., Pasa-Tolic, L., and Pikaard, C.S. (2012). In vitro transcription activities of Pol IV, Pol V, and RDR2 reveal coupling of Pol IV and RDR2 for dsRNA synthesis in plant RNA silencing. Mol Cell *48*, 811-818.

Han, B.W., Wang, W., Li, C., Weng, Z., and Zamore, P.D. (2015). Noncoding RNA. piRNA-guided transposon cleavage initiates Zucchini-dependent, phased piRNA production. Science *348*, 817-821.

Havecker, E.R., Wallbridge, L.M., Hardcastle, T.J., Bush, M.S., Kelly, K.A., Dunn, R.M., Schwach, F., Doonan, J.H., and Baulcombe, D.C. (2010). The Arabidopsis RNA-directed DNA methylation argonautes functionally diverge based on their expression and interaction with target loci. Plant Cell *22*, 321-334.

Henderson, I.R., Zhang, X., Lu, C., Johnson, L., Meyers, B.C., Green, P.J., and Jacobsen, S.E. (2006). Dissecting Arabidopsis thaliana DICER function in small RNA processing, gene silencing and DNA methylation patterning. Nature genetics *38*, 721-725.

Herr, A.J., Jensen, M.B., Dalmay, T., and Baulcombe, D.C. (2005). RNA polymerase IV directs silencing of endogenous DNA. Science *308*, 118-120.

Ji, L., Liu, X., Yan, J., Wang, W., Yumul, R.E., Kim, Y.J., Dinh, T.T., Liu, J., Cui, X., Zheng, B.*, et al.* (2011). ARGONAUTE10 and ARGONAUTE1 regulate the termination of floral stem cells through two microRNAs in Arabidopsis. PLoS Genet *7*, e1001358.

Johnson, L.M., Du, J., Hale, C.J., Bischof, S., Feng, S., Chodavarapu, R.K., Zhong, X., Marson, G., Pellegrini, M., Segal, D.J.*, et al.* (2014). SRA- and SET-domain-containing proteins link RNA polymerase V occupancy to DNA methylation. Nature *507*, 124-128.

Kanno, T., Huettel, B., Mette, M.F., Aufsatz, W., Jaligot, E., Daxinger, L., Kreil, D.P., Matzke, M., and Matzke, A.J. (2005). Atypical RNA polymerase subunits required for RNA-directed DNA methylation. Nature genetics *37*, 761-765.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol *10*, R25.

Law, J.A., Du, J., Hale, C.J., Feng, S., Krajewski, K., Palanca, A.M., Strahl, B.D., Patel, D.J., and Jacobsen, S.E. (2013). Polymerase IV occupancy at RNA-directed DNA methylation sites requires SHH1. Nature *498*, 385-389.

Law, J.A., and Jacobsen, S.E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nature reviews *11*, 204-220.

Law, J.A., Vashisht, A.A., Wohlschlegel, J.A., and Jacobsen, S.E. (2011). SHH1, a homeodomain protein required for DNA methylation, as well as RDR2, RDM4, and chromatin remodeling factors, associate with RNA polymerase IV. PLoS Genet *7*, e1002195.

Lee, H.C., Gu, W., Shirayama, M., Youngman, E., Conte, D., Jr., and Mello, C.C. (2012). C. elegans piRNAs mediate the genome-wide surveillance of germline transcripts. Cell *150*, 78-87.

Li, C.F., Henderson, I.R., Song, L., Fedoroff, N., Lagrange, T., and Jacobsen, S.E. (2008). Dynamic regulation of ARGONAUTE4 within multiple nuclear bodies in Arabidopsis thaliana. PLoS Genet *4*, e27.

Li, S., Vandivier, L.E., Tu, B., Gao, L., Won, S.Y., Li, S., Zheng, B., Gregory, B.D., and Chen, X. (2015). Detection of Pol IV/RDR2-dependent transcripts at the genomic scale in Arabidopsis reveals features and regulation of siRNA biogenesis. Genome Res *25*, 235-245.

Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H., and Ecker, J.R. (2008). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell *133*, 523-536.

Marcais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics *27*, 764-770.

Martin, C.T., and Coleman, J.E. (1989). T7 RNA polymerase does not interact with the 5'-phosphate of the initiating nucleotide. Biochemistry *28*, 2760-2762.

Matzke, M.A., and Mosher, R.A. (2014). RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. Nature reviews *15*, 394-408.

McCue, A.D., Panda, K., Nuthikattu, S., Choudury, S.G., Thomas, E.N., and Slotkin, R.K. (2015). ARGONAUTE 6 bridges transposable element mRNA-derived siRNAs to the establishment of DNA methylation. EMBO J *34*, 20-35.

Mi, S., Cai, T., Hu, Y., Chen, Y., Hodges, E., Ni, F., Wu, L., Li, S., Zhou, H., Long, C., *et al.* (2008). Sorting of small RNAs into Arabidopsis argonaute complexes is directed by the 5' terminal nucleotide. Cell *133*, 116-127.

Mohn, F., Handler, D., and Brennecke, J. (2015). Noncoding RNA. piRNA-guided slicing specifies transcripts for Zucchini-dependent, phased piRNA biogenesis. Science *348*, 812-817.

Nagano, H., Fukudome, A., Hiraguri, A., Moriyama, H., and Fukuhara, T. (2014). Distinct substrate specificities of Arabidopsis DCL3 and DCL4. Nucleic Acids Res *42*, 1845-1856.

Nechaev, S., Fargo, D.C., dos Santos, G., Liu, L., Gao, Y., and Adelman, K. (2010). Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. Science *327*, 335-338.

Onodera, Y., Haag, J.R., Ream, T., Costa Nunes, P., Pontes, O., and Pikaard, C.S. (2005). Plant nuclear RNA polymerase IV mediates siRNA and DNA methylation-dependent heterochromatin formation. Cell *120*, 613-622.

Pontier, D., Yahubyan, G., Vega, D., Bulski, A., Saez-Vasquez, J., Hakimi, M.A., Lerbs-Mache, S., Colot, V., and Lagrange, T. (2005). Reinforcement of silencing at transposons and highly repeated sequences requires the concerted action of two distinct RNA polymerases IV in Arabidopsis. Genes & development *19*, 2030-2040.

Ranjith-Kumar, C.T., Gutshall, L., Kim, M.J., Sarisky, R.T., and Kao, C.C. (2002). Requirements for de novo initiation of RNA synthesis by recombinant flaviviral RNA-dependent RNA polymerases. Journal of virology *76*, 12526-12536.

Ream, T.S., Haag, J.R., Wierzbicki, A.T., Nicora, C.D., Norbeck, A.D., Zhu, J.K., Hagen, G., Guilfoyle, T.J., Pasa-Tolic, L., and Pikaard, C.S. (2009). Subunit compositions of the RNA-silencing enzymes Pol IV and Pol V reveal their origins as specialized forms of RNA polymerase II. Mol Cell *33*, 192-203.

Rountree, M.R., and Selker, E.U. (1997). DNA methylation inhibits elongation but not initiation of transcription in Neurospora crassa. Genes & development *11*, 2383-2395.

Ruby, J.G., Jan, C., Player, C., Axtell, M.J., Lee, W., Nusbaum, C., Ge, H., and Bartel, D.P. (2006). Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans. Cell *127*, 1193-1207.

Shen, L., Shao, N., Liu, X., and Nestler, E. (2014). ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. BMC Genomics *15*, 284.

Stroud, H., Greenberg, M.V., Feng, S., Bernatavichute, Y.V., and Jacobsen, S.E. (2013). Comprehensive analysis of silencing mutants reveals complex regulation of the Arabidopsis methylome. Cell *152*, 352-364.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105-1111.

Xi, Y., and Li, W. (2009). BSMAP: whole genome bisulfite sequence MAPping program. BMC bioinformatics *10*, 232.

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

Yamamoto, Y.Y., Ichida, H., Matsui, M., Obokata, J., Sakurai, T., Satou, M., Seki, M., Shinozaki, K., and Abe, T. (2007). Identification of plant promoter constituents by analysis of local distribution of short sequences. BMC Genomics *8*, 67.

Ye, R., Wang, W., Iki, T., Liu, C., Wu, Y., Ishikawa, M., Zhou, X., and Qi, Y. (2012). Cytoplasmic assembly and selective nuclear import of Arabidopsis Argonaute4/siRNA complexes. Mol Cell *46*, 859-870.

Zheng, B., Wang, Z., Li, S., Yu, B., Liu, J.Y., and Chen, X. (2009). Intergenic transcription by RNA polymerase II coordinates Pol IV and Pol V in siRNA-directed transcriptional gene silencing in Arabidopsis. Genes & development *23*, 2850-2860.

Zhong, X., Du, J., Hale, C.J., Gallego-Bartolome, J., Feng, S., Vashisht, A.A., Chory, J., Wohlschlegel, J.A., Patel, D.J., and Jacobsen, S.E. (2014). Molecular mechanism of action of plant DRM de novo DNA methyltransferases. Cell *157*, 1050-1060.

Figure 1

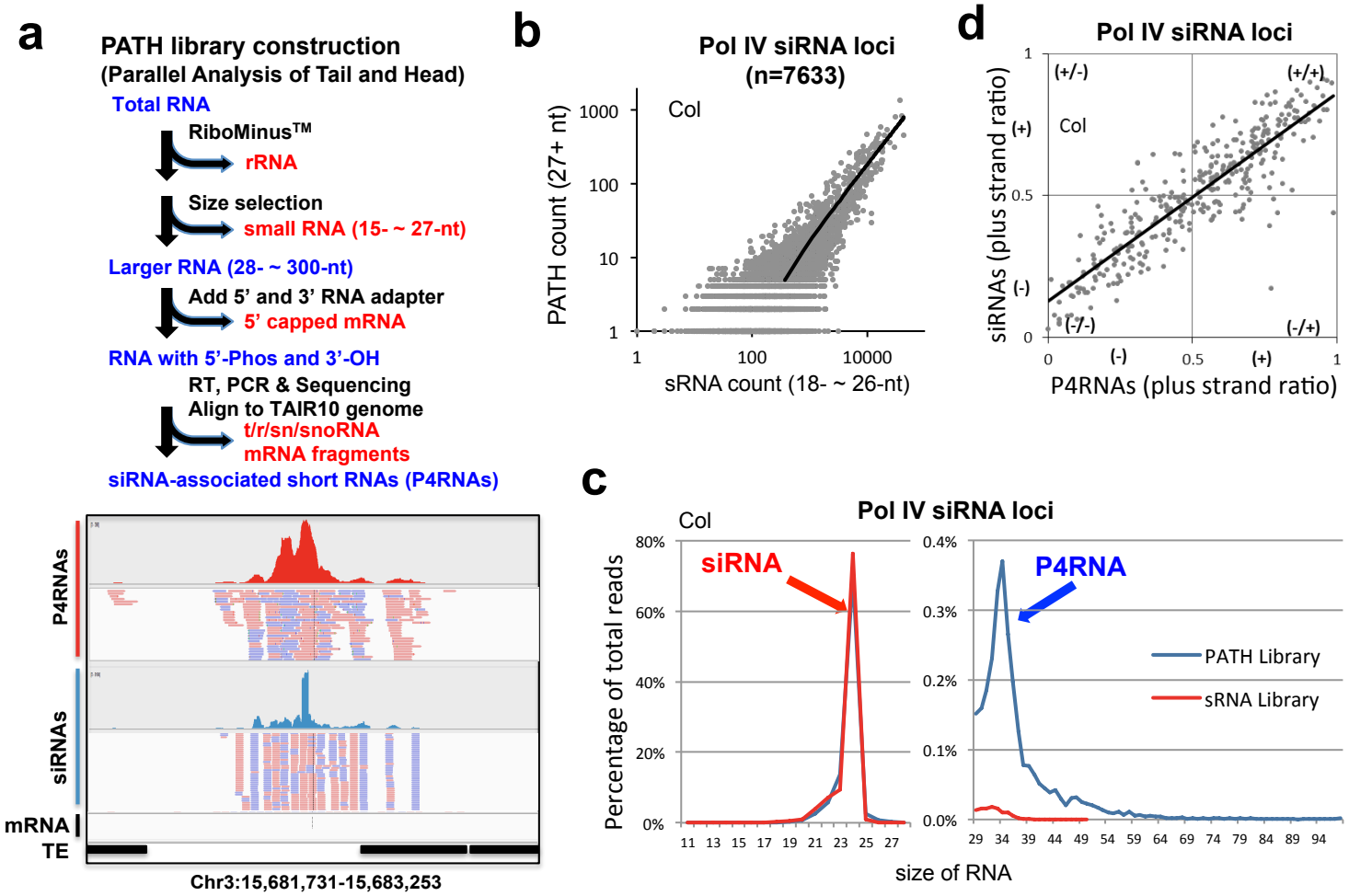


Figure 1. Identification of siRNA-loci associated short RNAs (P4RNAs) .

a. Procedure for the construction and analysis of PATH libraries. An example siRNA locus is shown at the bottom with IGV screenshots of P4RNAs and siRNAs matching to that region.

b. Abundances are highly correlated of PATH reads (27+ nt) and sRNA reads (18 to 26 nt) in wild-type Col from previously defined Pol IV siRNA loci (Law et al., 2013).

c. Size distribution plots of all PATH reads and sRNA reads in Col from Pol IV siRNA loci. A distinct peak at 30 to 40 nt can be seen in the PATH library; we named these "P4RNAs".

d. P4RNAs and siRNAs share the same strand bias at Pol IV siRNA loci. Only Pol IV siRNA loci matched by more than 100 P4RNAs in Col were selected to obtain a robust calculation of strandedness. The plus-strand ratio was calculated as the abundance of reads matching to the plus strand divided by the total number of reads at that locus.

Figure 2

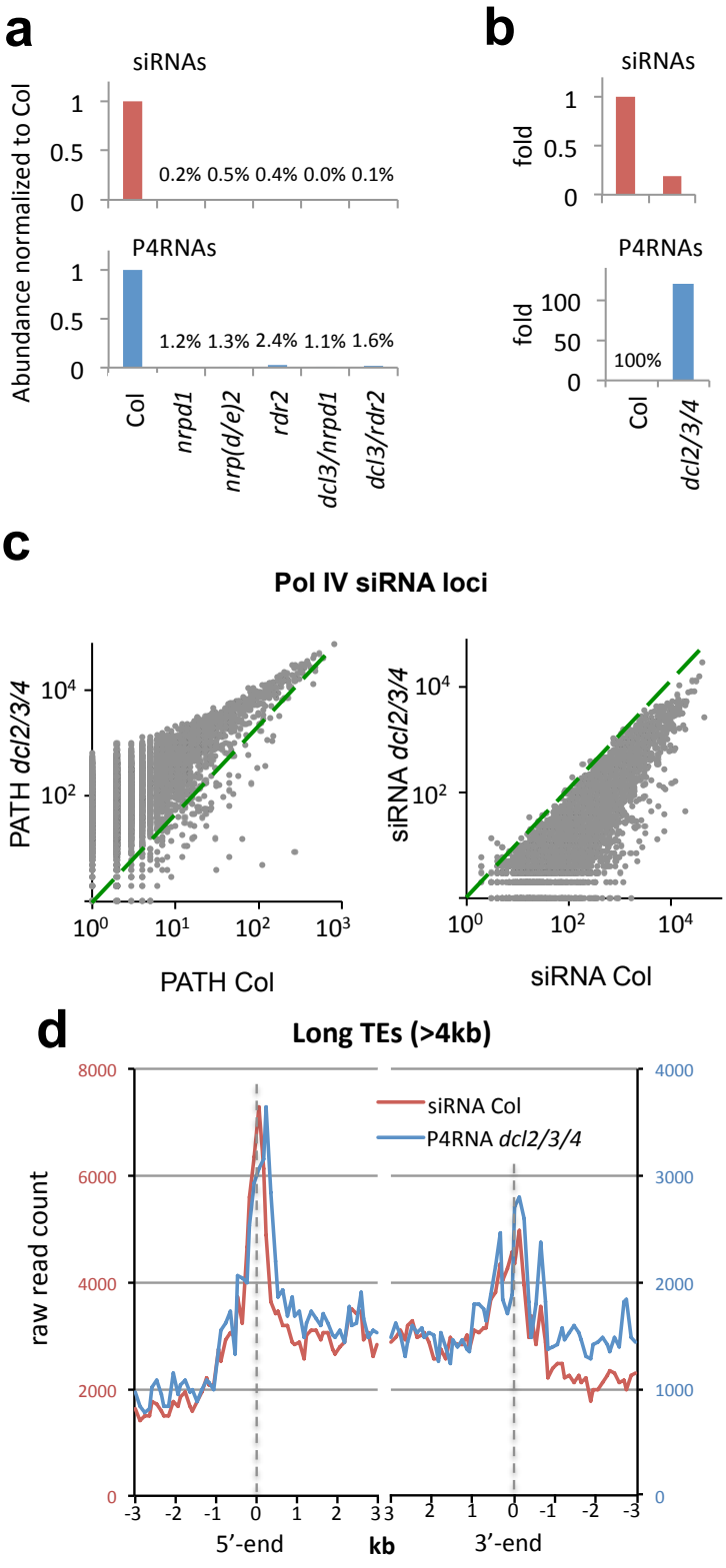


Figure 2. Biogenesis and processing of P4RNAs.

a. Both Pol IV siRNAs and P4RNAs are eliminated *in nrpd1, nrp(d/e)2, rdr2, dcl3/nrpd1* and *dcl3/rdr2*. Small RNA abundances were normalized to the sum of all TAIR10-annotated miRNAs in each sRNA library, and then compared to Col; P4RNA abundances were normalized to the total number of reads in each PATH library.

b. Pol IV siRNAs are reduced in *dcl2/3/4*, whereas P4RNAs accumulate substantially.

c. Correlation of P4RNA and siRNA abundance at each locus in Col and *dcl2/3/4*.

d. At long TEs, P4RNAs resemble the distribution of siRNAs – enriched at promoters and termini of TEs, and also spread into TE bodies.

Figure 3

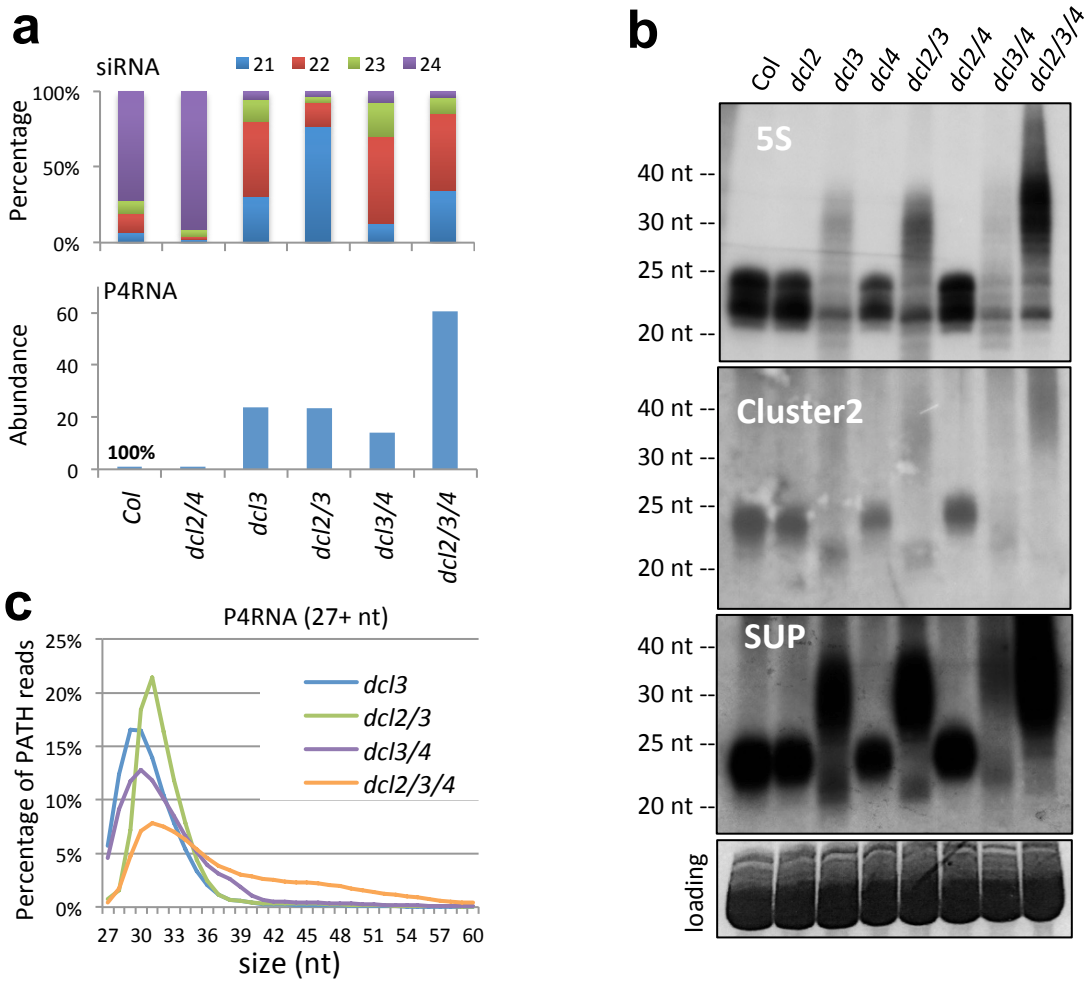


Figure 3. P4RNAs in *dcl* mutant combinations.

a. Size of Pol IV siRNAs shifted from 24 nt in Col to 21/22 nt in *dcl3*-containing mutants, accompanied by the accumulation of P4RNAs. P4RNA abundances are normalized to that in Col.

b. RNA blot analysis of siRNAs and P4RNAs in *dcl* mutant backgrounds.

c. Size distribution of P4RNAs in *dcl3*-containing mutants.

Figure 4

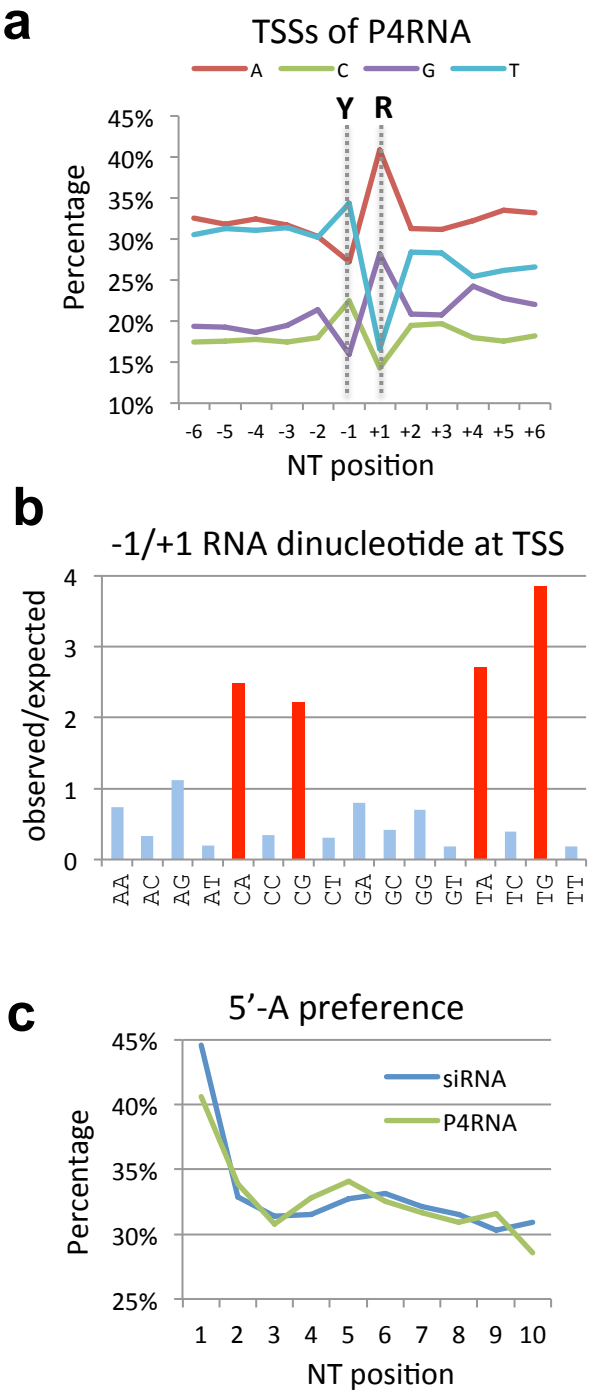


Figure 4. P4RNAs feature Pol II like TSSs and favor 5'-Adenine.

a. TSSs of Pol IV exhibit preference for C/T at the -1 position and A/G at the +1 position, resembling the "Y/R rule" of Pol II TSS, calculated using P4RNAs from *dcl2/3/4*.

b. Dinucleotide enrichment at -1/+1 of P4RNA, calculated using P4RNAs from *dcl2/3/4*.

c. Both siRNA and P4RNAs from Col have a strong preference for 5' adenine.

Figure 5

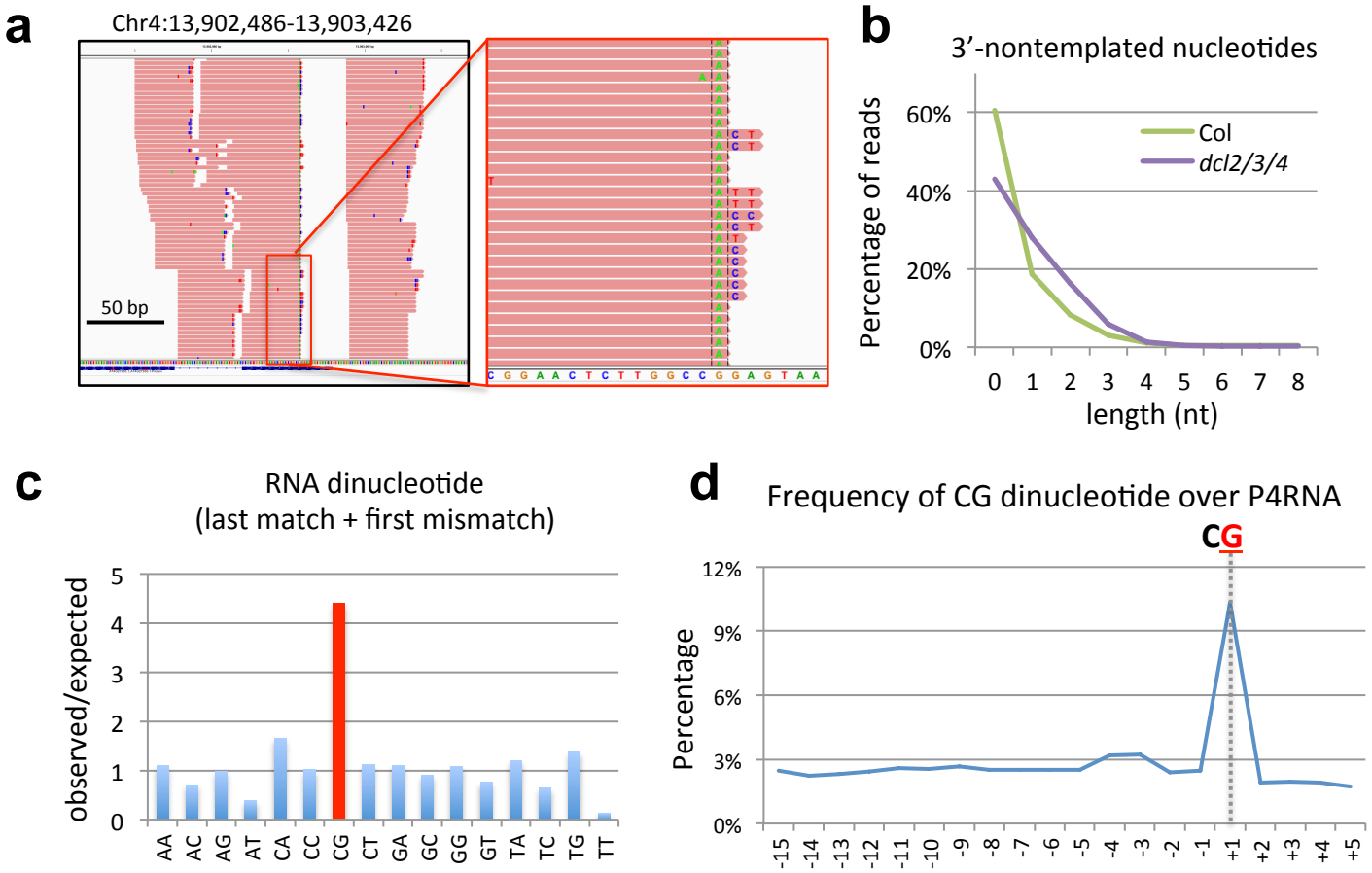


Figure 5. Pol IV transcription preferentially terminates at methylated cytosines with misincorporated nucleotides.

a. Example of 3'-nontemplated nucleotides on P4RNAs. The black bar represents 50 bp in length.

b. Length of the 3'-end nontemplated nucleotides, defined by the first mismatched nucleotide to the last nucleotide. If a P4RNA has no mismatch, the length is zero.

c. Di-nucleotide enrichment at the first mismatched position at the 3' end.

d. Frequency of CG dinucleotide on reference sequence over the P4RNAs with misincorporation, "-1" marks the last perfectly matched position, and "+1" marks the first mismatched position. The count of each CG is designated to the position of the G, therefore the peak at "+1" represents a peak of CG at "-1/+1".

Figure 6

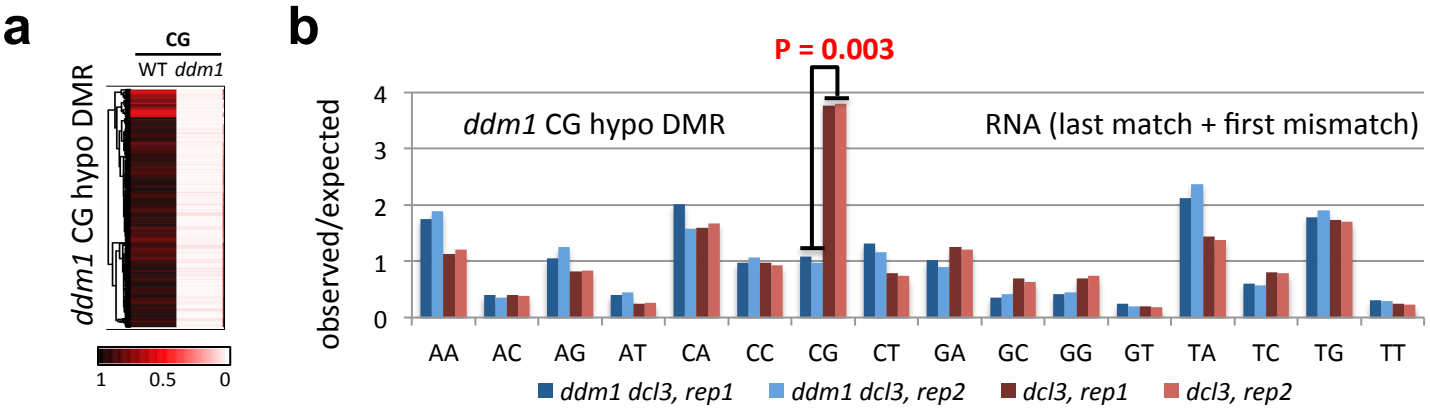


Figure 6. Misincorporation of P4RNAs at CG sites is suppressed in *ddm1/dcl3*.

a. CG hypomethylated DMRs in *ddm1* compared to a wild-type control.

b. 3'-misincorporated nucleotides of P4RNA at *ddm1* CG DMRs in *ddm1 dcl3* compared to *dcl3*.

Figure 7

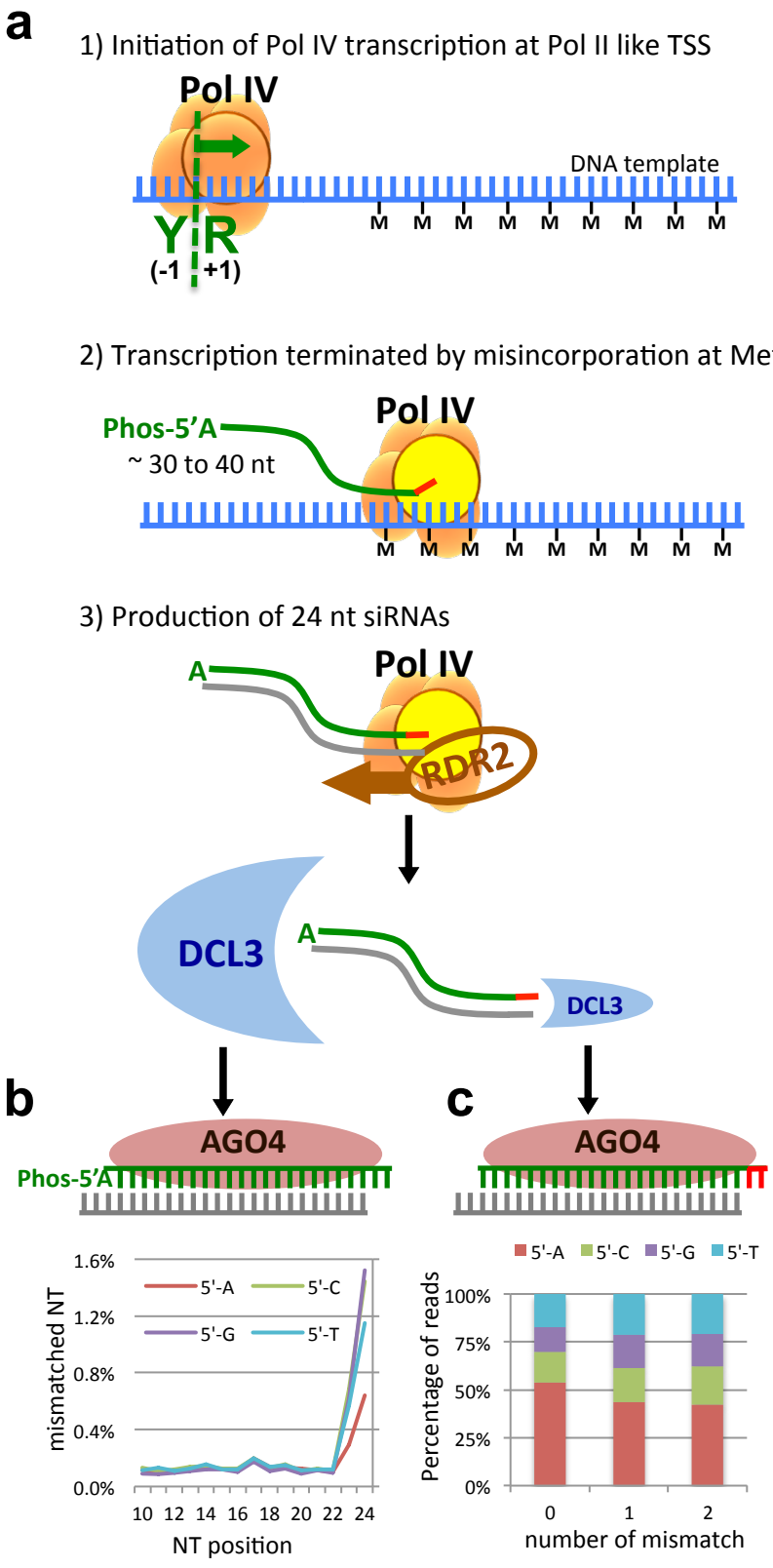


Figure 7. "one precursor, one siRNA" model for the biogenesis of Pol IV dependent 24-nt siRNAs.

a. Pol IV transcription is initiated at Pol II-like TSSs. A short RNA of ~30 to 40 nt (with 5'-adenine preference) is produced by Pol IV at heterochromatic regions; misincorporation then occurs at the cytosine position (red bar) and terminates Pol IV transcription. This process typically yields a P4RNA with 5' adenine and 3' misincorporation. P4RNA then goes through processing that involves RDR2 synthesizing the complementary strand (gray), DCL3 cutting from the 5' end (major) or 3' end (minor) of the P4RNA, and loading of the P4RNA-derived siRNA strand into AGO4 as the guide-strand. Eventually one P4RNA precursor gives rise to one siRNA that is derived from either its 5' or 3' end. ("M" underneath the DNA template indicates DNA methylation at the heterochromatic region.)

b. When DCL3 cuts from the 5' end of P4RNAs, the resulting siRNAs are more likely to carry a 5' adenine and be perfectly matched. Indeed a reduction of 3' mismatches is observed for siRNAs with a 5' adenine.

c. When DCL3 cuts from the 3' end of P4RNAs, the resulting siRNAs are more likely to carry the 3' misincorporation and less likely the 5' adenine. This is consistent with the observation that siRNAs with mismatch have lower percentage of 5' adenine.

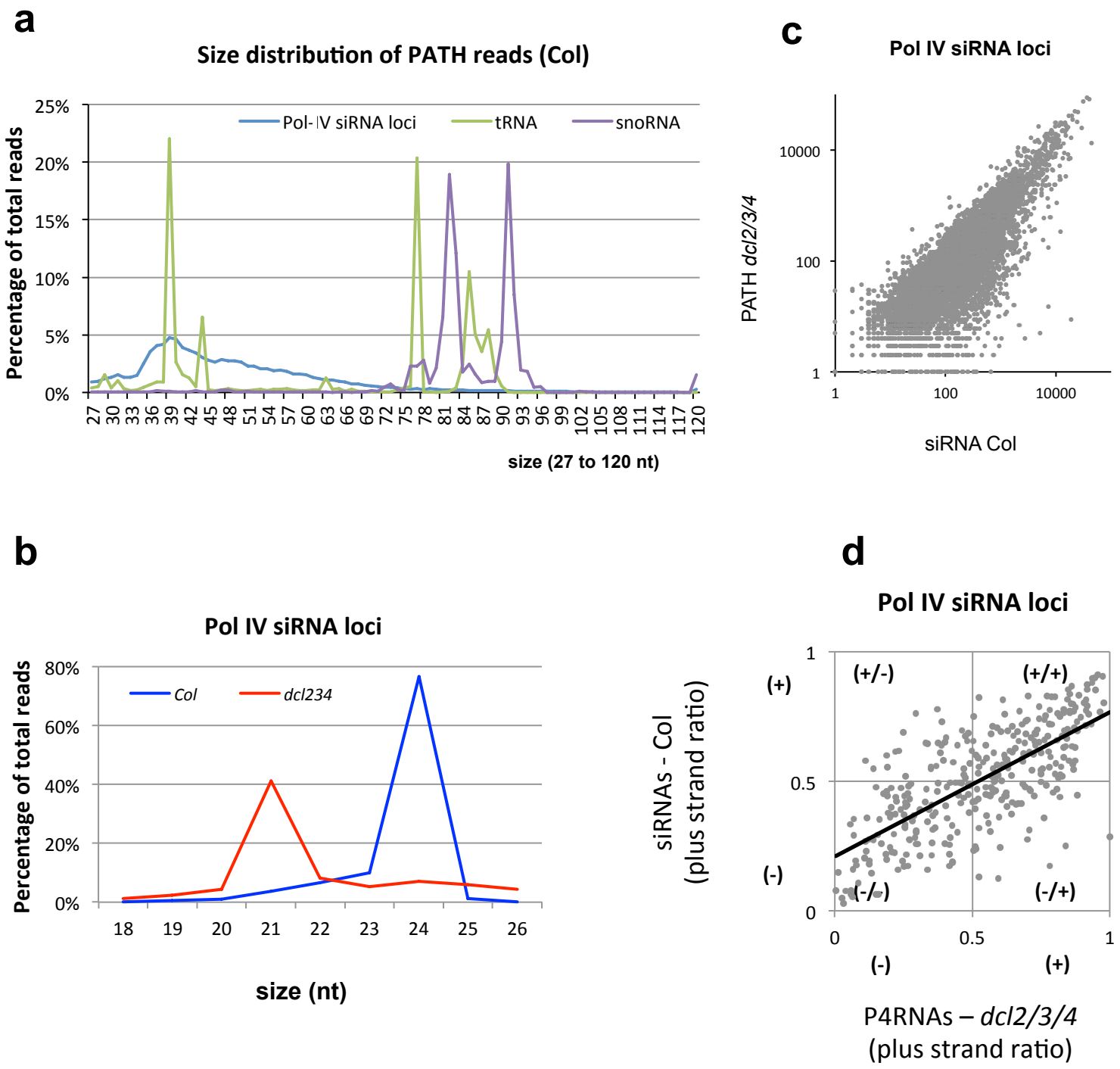


Figure S1. Size distribution of PATH reads and siRNAs

a. Size distributions of PATH reads aligned to Pol IV siRNA loci, tRNA loci, and snoRNA loci were shown as the percentage of total read count of that class of loci.
b. In Col the siRNAs from Pol IV siRNA loci are predominantly 24-nt in size, while in the *dcl2/3/4* mutant the peak of siRNAs is shifted to 21-nt.
c. Correlation of P4RNAs in *dcl2/3/4* with siRNAs in Col at Pol IV siRNA loci.
d. P4RNAs and siRNAs share the same strand bias. Shown is the same set of Pol IV siRNA loci that were shown in Figure 1d (bins with no less than 100 P4RNA reads in Col). The plus-strand ratio was calculated as the abundance of reads matching to the plus strand divided by the total number of reads at that locus.

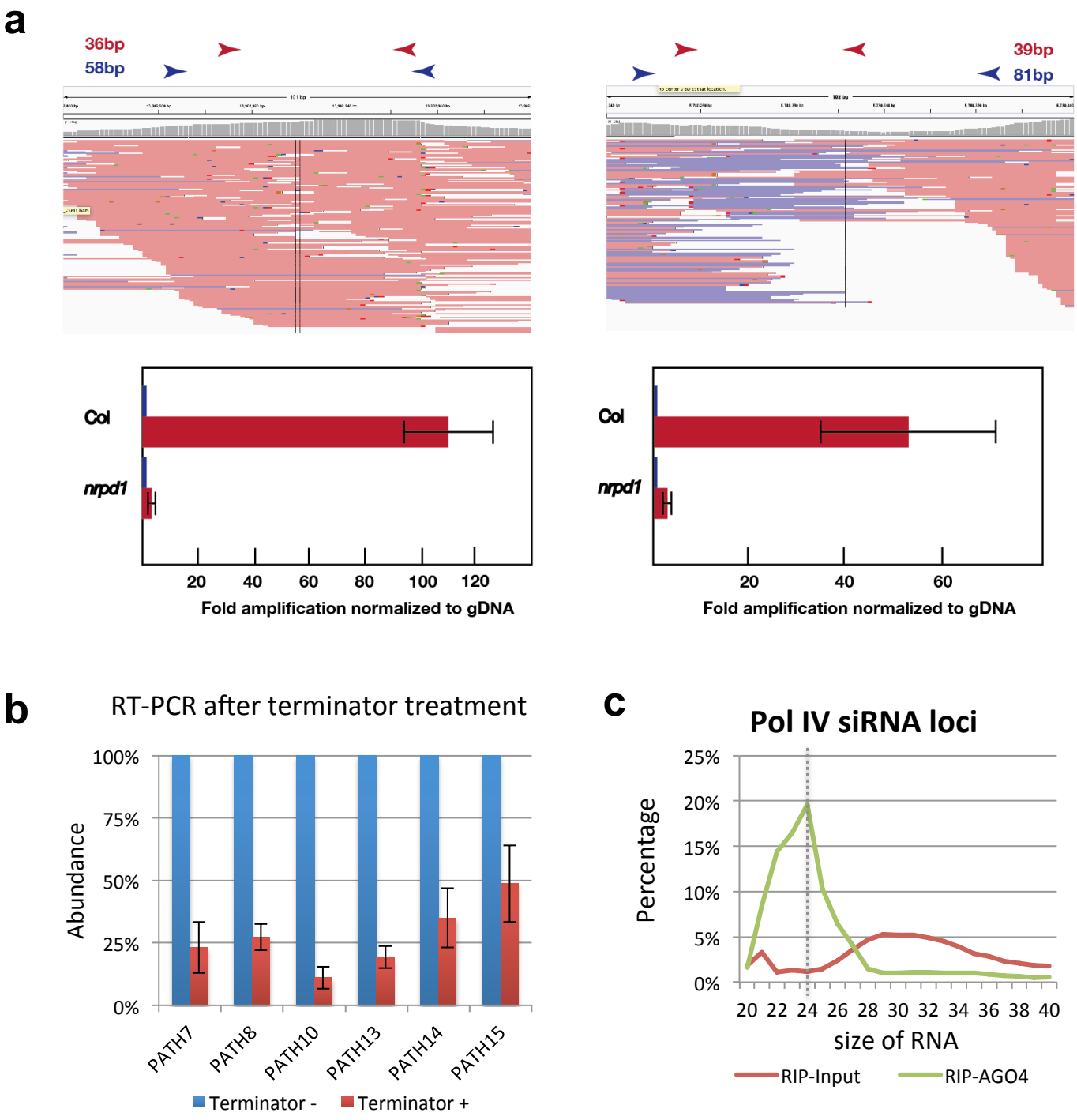


Figure S2. 5' modification and size of Pol IV transcripts.

a. Graphical representation of two Pol IV target loci. Red and blue arrows represent primers. Fold change amplification measured by quantitative PCR. Red bars represent the fold change in amplification of the small region (red primers) verses bigger region (blue primers) normalized against amplification of the same primers on genomic DNA. Error bars represent standard error across four replicate. Primer information can be found in Table S3.

b. Terminator exonuclease specifically digest RNAs with 5'-monophosphate. Six P4RNA loci were examined by RT-PCR using templates with and without terminator treatment. A large portion of P4RNAs have a 5' monophosphate, evidenced by their degradation via treatment with Terminator.

c. Size distribution of RNAs from AGO4 bound compared to input. Experiment was done in the *dcl2/3/4* background.

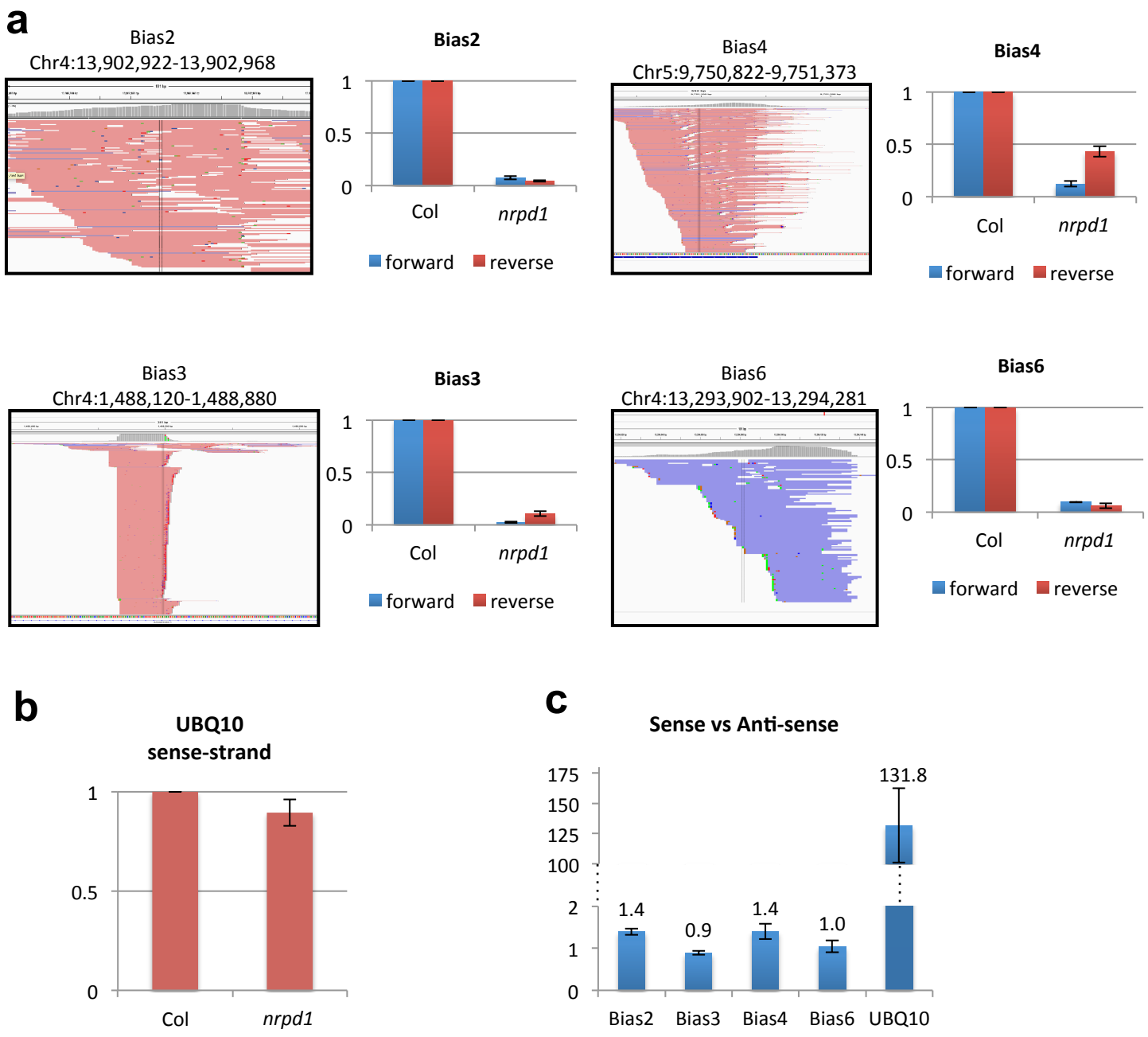
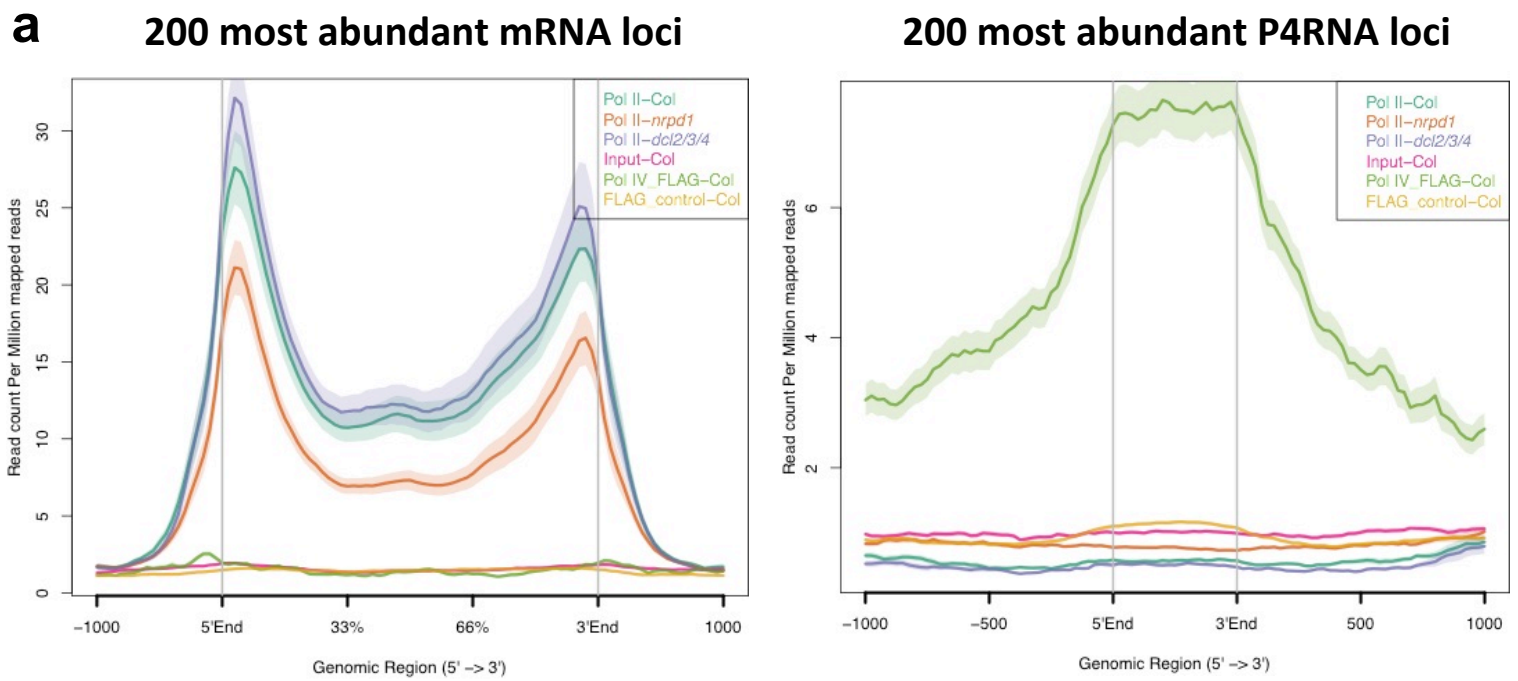
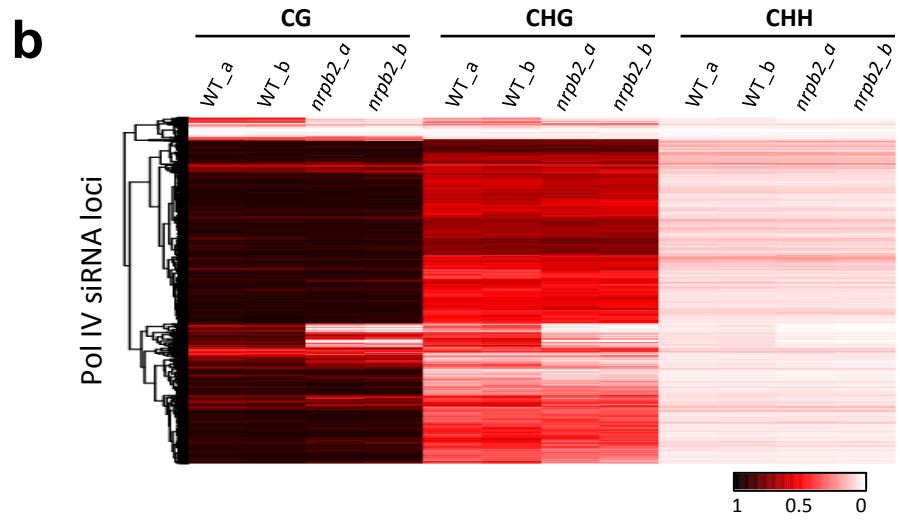


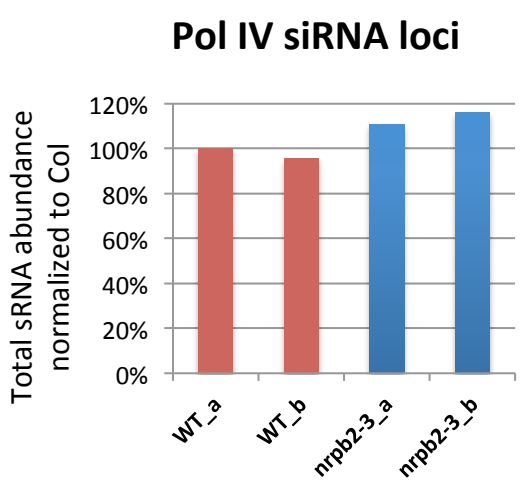
Figure S3. Strand-biased P4RNA loci

a. Four loci that exhibit strand-biased accumulation of P4RNAs in the *dcl2/3/4* PATH library were examined by strand-specific RT-PCR. Red bars are reads mapped to Watson strand and blue bars map to Crick strand. Amplifications from both sense and anti-sense strands are Pol IV-dependent.
b. Relative abundance of UBQ10 in Col and *nrpd1.* UBQ10 is not Pol IV-dependent and served as a control.
c. The relative ratio of sense verses anti-sense amplification is calculated assuming equal reverse transcription efficiency for sense and anti-sense strand at each locus. UBQ10 is used as a control to illustrate clear strand preference.

Primer information can be found in Table S3. Standard errors were calculated from four biological replicates.







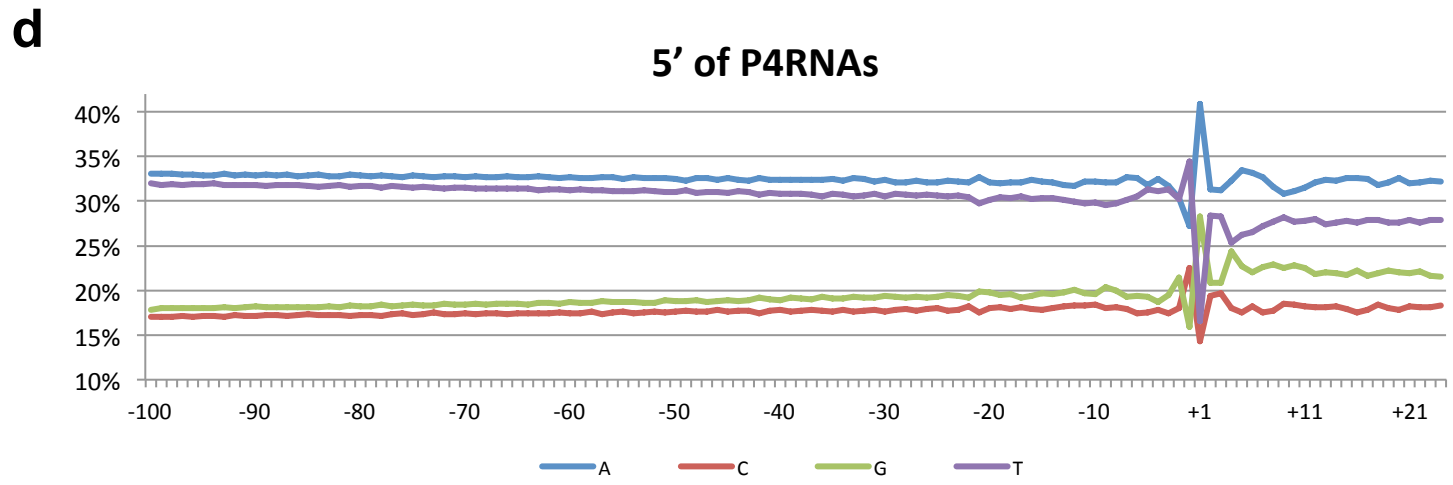


Figure S4. Pol II and Pol IV occupy distinct genomic loci.

a. ChIP-seq of Pol II and Pol IV in various genetic backgrounds at either the top expressing mRNA or top expressing P4RNA loci.

b. DNA methylation at Pol IV siRNA loci in the weak Pol II mutant (*nrpb2-3*) revealed by BS-seq. Two biological replicates _a _b were performed.

c. Abundance of 24-nt siRNAs at Pol IV siRNA loci, first normalized to the total number of mapped reads in each library then compared to Col (WT_a). Two biological replicates _a _b were performed.

d. Nucleotide composition at up to 100 nt upstream of P4RNA TSSs.

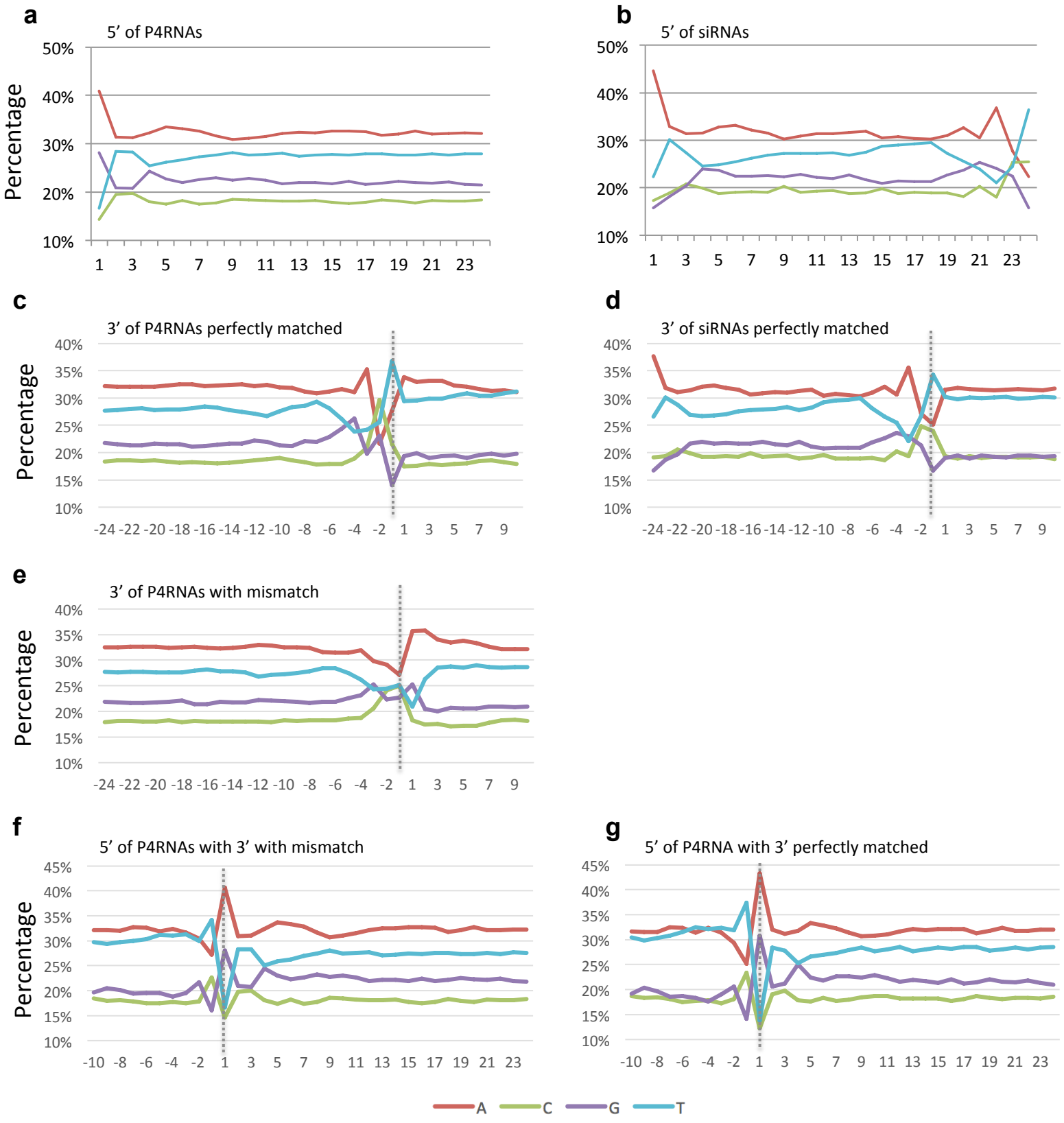


Figure S5. 5' and 3'-end sequence composition of P4RNA and siRNA.

The 5'-end sequence composition of the first 24 nt of P4RNA (a) and the entire 24-nt siRNA (b). The 3'-end sequence composition of perfectly matched P4RNA (c) and 24-nt siRNA (d). The 3'-end sequence composition of mismatched P4RNA (e). P4RNAs with 3' non-templated nucleotides (f) have similar 5' end composition as those without (g). For 3' analysis, "-1" position, indicated by the dashed line, is the last nucleotide of RNA that is perfectly matched, "1" position is the next nucleotide of in genomic DNA sequence; For 5' analysis, "1" position, indicated by the dashed line, is the first nucleotide of RNA.

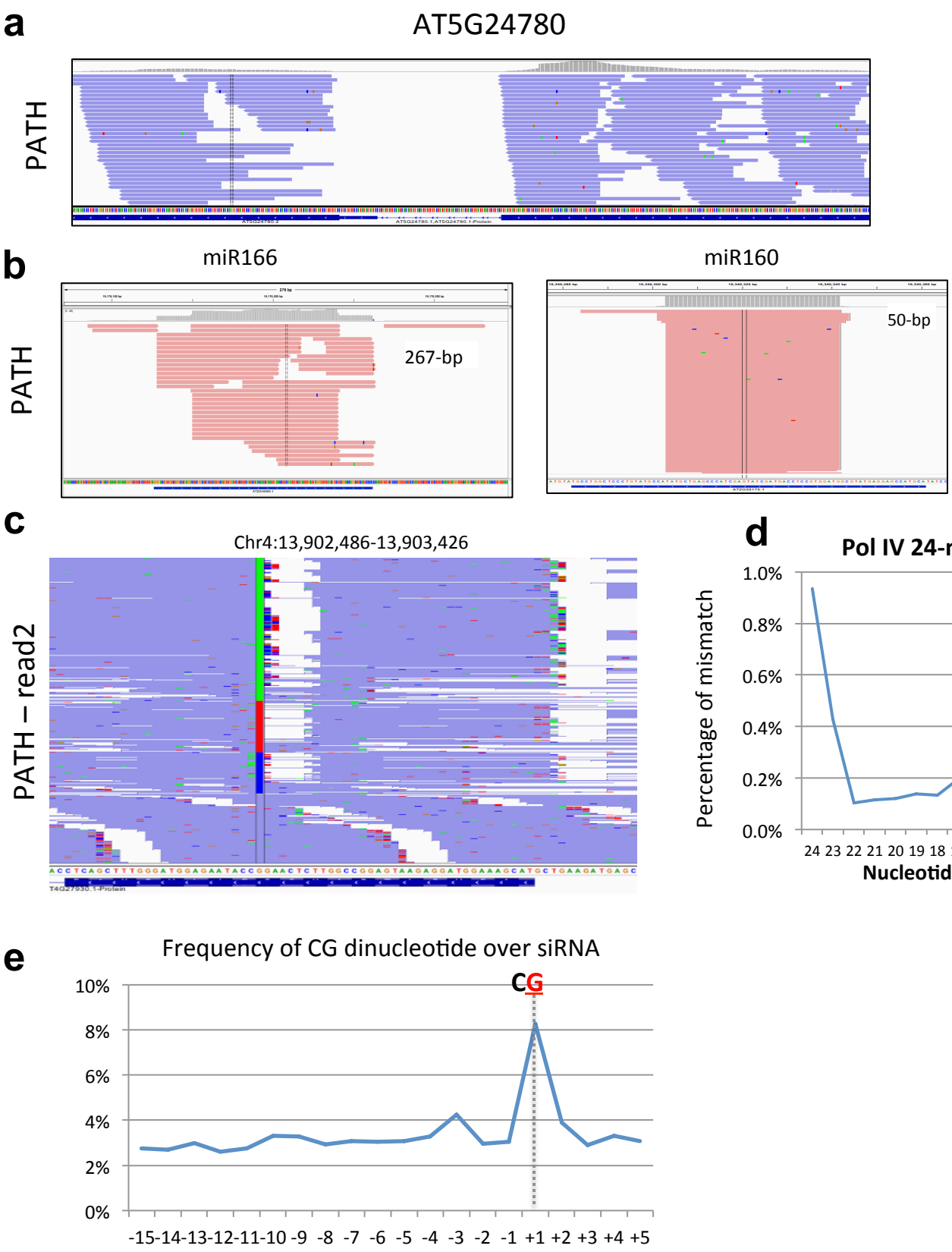


Figure S6. PATH reads from Pol II transcribed regions do not enrich for 3' mismatched nucleotides.

a. Example of an mRNA region.
b. Examples of miRNA loci.
c. Read2 of the paired-end sequencing from PATH library. The beginning nucleotide of read2 corresponds to the 3' end nucleotide of RNA.
d. Percentage of mismatched nucleotides on 3'-ends of 24-nt siRNAs. Position 24th represents the last position of the 24 nt siRNA.
e. Frequency of CG dinucleotide on reference sequence over the siRNA with misincorporation, "-1" marks the last perfectly matched position, and "+1" marks the first mismatched position. The count of each CG is designated to the position of the G, therefore the peak at "+1" represents a peak of CG at "-1/+1".

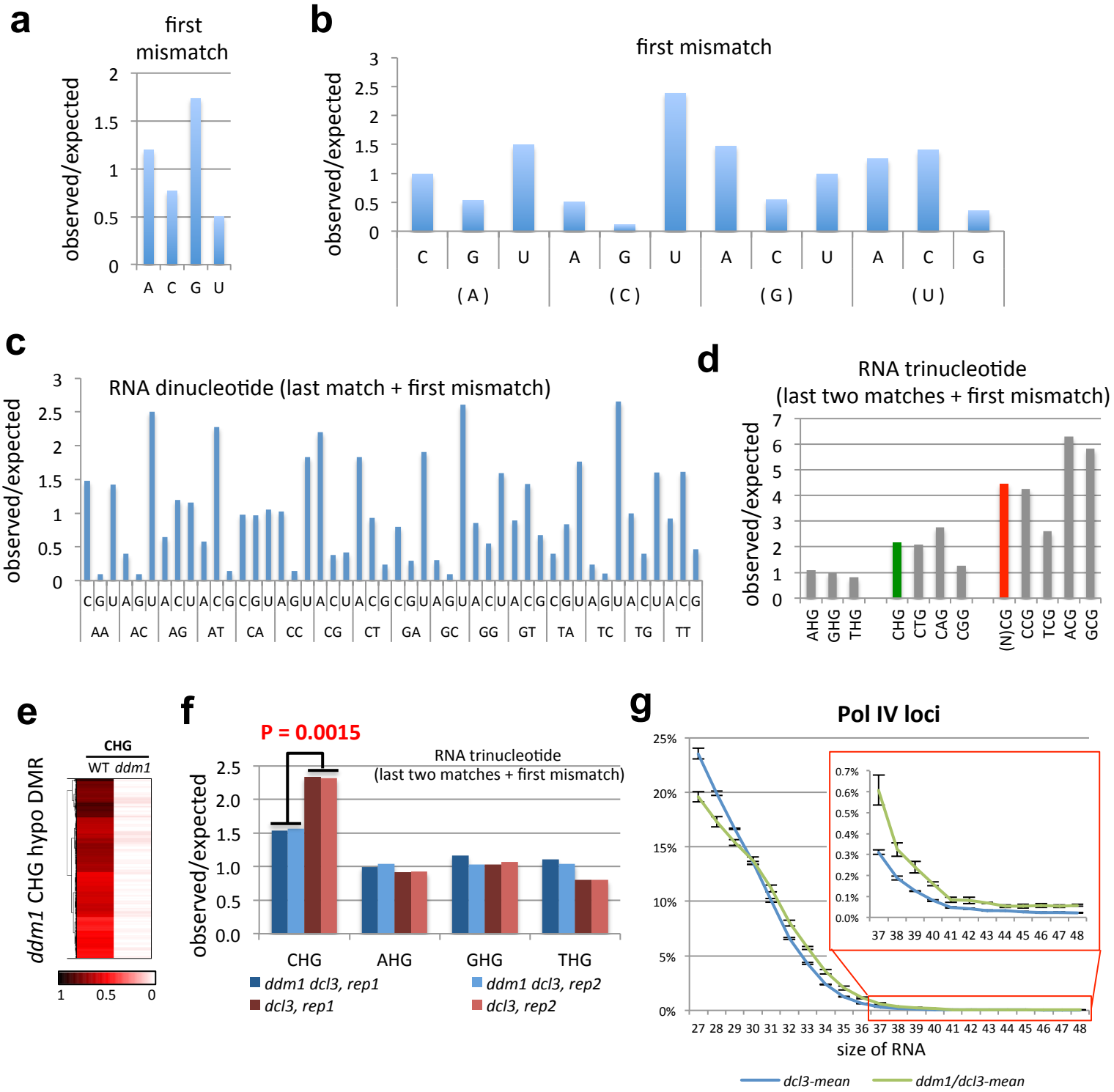


Figure S7. 3'-end mismatch nucleotide composition.

a. Mono-nucleotide enrichment at the first mismatched position.
b. Relative enrichment for different type of mismatches – nucleotide enclosed in parentheses is the supposed perfectly matched nucleotide, different nucleotide above represents different mismatches.
c. Relatively enrichment for different type of mismatches, corresponding to the last nucleotide of the di-nucleotide (last perfectly-matched + first matched).
d. Tri-nucleotide enrichment at the first mismatched position.
e. CHG hypomethylated DMRs in *ddm1* compared to a wild-type Col control.
f. 3'-misincorporated nucleotides of P4RNA at *ddm1* CHG DMRs in *ddm1 dcl3* compared to *dcl3*.
g. P4RNAs are slightly longer in *ddm1/dcl3* compared to *dcl3*. Mean and standard error were calculated from three biological replicates of each genotype.