SUPPLEMENTAL METHODS

Chromatin Immunoprecipitation

ChIP was performed as described (Rowley *et al.*, 2011) with slight modifications. 3 grams of 2-3 week old seedling tissue was harvested and crosslinked with 0.5% formaldehyde for 2 min followed by 8 min vacuum infiltration. Glycine was added to 80 mM and vacuum reapplied for 1 min then 4 min. Crosslinked tissue was rinsed with water and frozen in liquid nitrogen. Nuclei were extracted by grinding frozen tissue into powder using a mortar and pestle, suspended in 25 ml of Honda Buffer (20 mM HEPES-KOH pH 7.4, 0.44 M sucrose, 1.25% Ficoll, 2.5% Dextran T40, 10 mM $MgCl_2$, 0.5% Triton X-100, 5 mM DTT, 1 mM PMSF, 1% plant protease inhibitors (Sigma)), filtered through two layers of Miracloth and centrifuged at 2000 x g for 15 min. Nuclear pellets were washed three times with 1 ml of Honda buffer, resuspended in Nuclei Lysis Buffer (50 mM Tris-HCl pH 8.0, 10 mM EDTA, 1% SDS, 1 mM PMSF, 1% Plant Protease Inhibitors from Sigma) and DNA was fragmented to the average size of 250 bp by 8 pulses of sonication each 10 seconds long with 1 minute pauses in between pulses using Fisher Scientific 100 Sonic Dismembrator at power setting 1. After centrifugation at 15,000 x g for 10 min, the supernatant was diluted 10-fold with 1.1% Triton X-100, 1.2 mM EDTA, 16.7 mM Tris-HCl pH 8.0, 167 mM NaCl. 50 µl Dynabeads Protein A (Invitrogen) and the appropriate antibody were added and samples were incubated for 8 h at 4 ˚C on a rotating mixer. Bead-antibody complexes were washed 5 times, 5 min each, with binding/washing buffer (150 mM NaCl, 20 mM Tris-HCl pH 8.0, 2 mM EDTA, 1% Triton X-100, 0.1% SDS, 1 mM PMSF and twice for 5 min each with TE. Samples for ChIP-seq were eluted with RIP elution buffer (100 mM Tris-HCl pH 8.0, 10 mM EDTA, 1% SDS) for 20 min at 65 ˚C and were digested with 20 µg of proteinase K (Invitrogen) overnight at 60 ˚C. An equal volume of phenol/chloroform/isoamyl alcohol pH 6.7 25:24:1 was added to extract DNA, followed by addition of an equal volume of chloroform/isoamyl alcohol 24:1 and subsequent precipitation by addition of 2 volumes 100% EtOH, 0.1 volume 3 M Sodium Acetate and 4 µl Glycoblue (Ambion). Precipitated samples were washed once with 70% EtOH and resuspended in 30 µl TE. Other ChIP samples were eluted using 100 µl of 10% (w/v) Chelex (Bio Rad) resin, in water, added to the beads and crosslinking was reversed at 99 ˚C for 10 min. Samples were digested with 20 µg of proteinase K (Invitrogen) for 2 h at 60 ˚C followed by heat-inactivation at 95 ˚C for 10 min. ChIP samples were amplified in triplicate in Applied Biosystems 7500 real time PCR machine and obtained data were analyzed using comparative Ct relative to inputs. All ChIP-real time PCR experiments were replicated in two or three independent biological repeats, which yielded very similar results.

ChIP-seq library construction

All ChIP-seq libraries (6 total; Col-0, *ago4,* and *nrpe1* ChIP and input samples) were prepared according to the Illumina ChIP-seq library preparation protocol.

High-throughput sequencing

All ChIP-seq or input libraries were sequenced using an Illumina Genome Analyzer IIx at the University of Michigan DNA Sequencing Core as per manufacturer's instructions for 80 nt single-end sequencing.

Pre-processing and mapping of sequencing reads

In essence, all reads were pre-processed and mapped to the *Arabidopsis* genome using a pipeline as previously described (Zheng *et al.*, 2010) with slight modifications. The detailed procedures are described below:

Trimming of 3'-adaptors. All raw reads were aligned to the Illumina Genomic DNA 3'-adaptor sequence using cross-match program from Phrap package (http://www.phrap.org/phredphrapconsed.html#block_phrap), and those with ≥10 nts of alignment at the 3'-end with ≤10% mismaches were subsequently trimmed at the insert/adaptor junctions. Reads without detectable 3'-adaptors were also kept un-changed for subsequent processing.

Reducing to NR-tags. Both trimmed and untrimmed reads were reduced to non-redundant (NR) tags by collapsing reads with identical sequences; the goal of this step is to save processing time and space requirements. The clone-number for each NR-tag was also recorded and is then used for all subsequent analysis. We will use the term “read” and “NR-tag” interchangeably hereafter.

Mapping to *Arabidopsis* genome. The trimmed and untrimmed reads were mapped to *Arabidopsis thaliana* genome (TAIR9 assembly) independently using the Bowtie program (Langmead *et al.*, 2009), with parameters tuned to allow ≤6% of seed mismatches (using 34 nt seeds), ≤8% of total mismatches and all valid alignments are reported. A subsequent parsing step was implemented to enforce these restraints, as well as to require insert lengths of ≥15 nt or ≥30 nt for the trimmed and untrimmed reads, respectively. It is of importance that the actual “insert length” for untrimmed reads was determined according to their alignments, by implementing a one-dimensional dynamic programming algorithm that could identify the most possible insert-fragment length based on output from Bowtie. Finally, we filtered only “best-stratum” alignments that contain ≤4% more mismatches compared to the best-hits for any given read.

Summary of mapping and clone-number information. All mapped trimmed and untrimmed reads (NR-tags) were combined and their mapping and clone-number information was recorded. All of these data were loaded into a local MySQL database for subsequent fast queries.

Calling AGO4-bound peaks

To call AGO4-bound peaks (AGO4 binding regions) using our ChIP-seq data, input tables were prepared in which genome coordinates and weighted clone-number were included for all 6 libraries (ChIP and input for Col-0, *ago4*, and *nrpe1* plants). The weighted clone-number is defined as $W_i = round(C_i / L_i)$, where the W_i, C_i and L_i is the weighted clone-number, raw clone-number and number of mapped loci for a given NR-tag i. It is of note that by using weighted clone-numbers, we have the advantages of allowing non-uniquely mapping reads and the non-biased estimation of their clone-abundance. This step is necessary because AGO4 is thought to target heterochromatin and repetitive elements in *Arabidopsis*.

We then called different sets of peaks using the CSAR(Muiño *et al.*, 2011) R package, with non-default parameters set as: w = 250, considerStrand = "Sum", uniquelyMapped = FALSE, backg = 0, norm = 2e10. Taken together, these parameters extend all mapped reads up-to 250 nt (average size of the ChIP-seq fragments used to construct the sequencing libraries), merge them from both strands, normalize, and finally call AGO4-bound peaks. All calls required significant fold-enrichment between test and control with FDR < 0.05; the FDR was achieved by randomly permuting the mapped reads from test samples 10 times using the CSAR (Muiño *et al.*, 2011) package. As a result, 5 sets of peaks were called:

A = Col-0 ChIP vs. Col-0 input

B = *ago4* ChIP vs. *ago4* input

C = *nrpe1* ChIP vs. *nrpe1* input

D = Col-0 ChIP vs. *ago4* ChIP
E = *nrpe1* ChIP vs. *ago4*-ChIP
The A, B, C peak sets are traditional ChIP against input calls and D, E sets are “direct-comparison” peaks using the *ago4* null mutant sample that were included in this study to eliminate effect of non-specific binding of DNA to the AGO4 antibody. We then defined Pol V-dependent and Pol V-independent peaks using the following “peak-arithmetic”. Specifically, Pol V-dependent peaks are defined as $F – G$ and Pol V-independent peaks are $F \cap G$, where $F = (A – B) \cap D$ and $G = (C – B) \cap E$. This peak-arithmetic was designed to identify high-quality peaks enriched for both ChIP vs. input and WT vs. mutant comparisons and minimize the effects of non-specific interactions. All the peak-arithmetic was performed using BEDTools (Quinlan and Hall, 2010), with the overlapping proportion being no less than half (-f = 0.5) of the peaks being compared.

Filtering Pol V-dependent and Pol V-independent peaks

Our *ago4* mutant plants were originally identified (Zilberman *et al.*, 2003) in the Landsberg (Ler-1) ecotype of *Arabidopsis*, which was subsequently back-crossed to Col-0 plants for 3 successive times. As a consequence, the *ago4* plants could still contain some proportion of the genome that originates from Ler-1, and thus the calling procedure of Pol V-dependent and Pol V-independent peaks could include ecotype biases. Therefore, we further filtered out peaks that could originate specifically from Ler-1. In essence, any peak that either ①cannot be mapped to Ler-1 draft genome (Ler-1 unmappable) or ②can be better mapped to Ler-1 draft genome (Ler-1 better mapped) was recognized as potentially originating from Ler-1, and thus discarded from further analysis. More specifically, we first pulled out all raw reads as well as their qualities (in FASTQ format) in all called Pol V-dependent or Pol V-independent peaks, then re-mapped them to both the Col-0 genome and Ler-1 draft genome as described above. Only the *ago4* ChIP library was used as a proxy for this analysis. It is of note that we used the “standard” assembly of the 2011-08-25 release of the Ler-1 genome from the 1001 genomes project (http://1001genomes.org/), which was in draft status and still lacked a significant portion of the genome compared to the Col-0 genome sequence (TAIR9 assembly). We also re-mapped the reads to Col-0 genome while retaining read quality information, so that the mapping quality between the Col-0 and Ler-1 genomes for any given read could be better distinguished. All mapping criteria were kept identical as described above. By comparing the alignments for the Col-0 and Ler-1 genomes for each read, we defined “Ler-1 unmappable peaks” as those that contain <50% reads mapping to Ler-1 genome relative to Col-0, and “Ler-1 better mapped peaks” as those that contain more reads that can either exclusively or better map to Ler-1 compared to the Col-0 genome. A read is deemed as “better mapped” to a genome if the best hit to that genome contains fewer mismatches, and if it is a tie (same number of mismatches) they are further resolved by comparing the total quality scores over all mismatch sites.

Sampling of random peaks as negative controls

To generate negative control peaks (NC-peaks) for our analysis, we randomly sampled genomic regions from the Col-0 genome, with the same number and size-distribution as the filtered Pol V-dependent peaks, and this sampling was repeated 1000 times. All described analyses were based on these same sets of NC-peaks.

Partitioning the Pol V-completely dependent and Pol V-partially dependent peaks

To distinguish the AGO4 peaks that are completely-dependent from those that are partially-dependent on Pol V activity, we calculated the total number of reads within all Pol V-dependent peaks for all ChIP samples, then plotted the log-odds of enrichment for Col-0 ChIP vs. *ago4* ChIP against *nrpe1* ChIP vs. *ago4* ChIP on a scatter-plot (Fig. 1C). The Pol V-completely dependent peaks were defined as those with less than 2 fold enrichment when comparing *nrpe1* ChIP vs. *ago4* ChIP (|abs(log-odds)| < 1), whereas Pol V-partially dependent peaks were defined as all of the remaining peaks. This partitioning of peaks is based on the assumption that Pol V-completely dependent peaks should show no significant difference in clone-abundance between *nrpe1* and *ago4* samples. As expected, most Pol V-dependent peaks are completely dependent, and we didn’t separate these peaks in further analyses for convenience, since partially dependent peaks are an insignificant fraction of the total AGO4 peaks.

Classification and annotation of AGO4 peaks

All AGO4-peaks were classified according to their genomic coordinates compared to known genetic elements on the *Arabidopsis* genome using the GFF annotation file downloaded from TAIR9 FTP repository for varies kinds of elements, including protein-coding genes (exons and introns), rRNAs, tRNAs, miRNAs, snoRNAs, snRNAs, ncRNAs, pseudogenes, and transposable elements (TEs). We also defined gene promoters as the upstream 1 kb regions of the transcription-start sites (TSS) of protein-coding genes. Additionally, we also searched the *Arabidopsis* genome (TAIR9 assembly) for more repetitive elements using the RepeatMasker program (RepeatMasker-open-3.2.8) (http://www.repeatmasker.org) with the repeat libraries from RepBase (release14.06) (Jurka *et al.*, 2005). We used the RepeatMasker annotated repeats because the TEs annotated by TAIR don't have detailed class or family information. Other than the TEs, the RepeatMasker (RMSK) program could also identify repeat-rRNAs (RMSK-rRNAs) and tandem-repeats (RMSK-TRs).

To fast classify AGO4 peaks, we implemented a Java program that indexes various kinds of elements of the whole genome with bits. To produce a detailed annotation of the identified AGO4 peaks, all above genetic elements were loaded into the MySQL database and searched for overlapping ones for every AGO4 peak. As a control, all the NC-peaks were also classified and annotated as described for the AGO4 peaks, and the p-values of enrichment or depletion of specific categories were estimated using a bootstrapping method based on the 1000 sets of NC-peaks.

Characterizing smRNA profiles near AGO4 peaks

We downloaded published smRNA-IP and total smRNA datasets (Wang *et al.*, 2011) for both AGO4 and AGO1 from *Arabidopsis* seedlings (accession: GSE28591) for our analysis. Raw reads were dumped from the NCBI SRA, processed, and mapped to the *Arabidopsis* genome as described for our ChIP-seq libraries. Then smRNA-IP or total smRNA reads were searched within the AGO4 peaks as well as their flanking regions (2 kb of both upstream and downstream), and the base-wise coverage for every peak was determined for 24 nt and 21 nt reads, respectively. Finally, the coverage values were normalized by the total mapped reads of each specific sized library, and averaged across all AGO4 peaks.

Characterizing cytosine methylation in AGO4 peaks

To characterize the cytosine methylation (mC) in AGO4 peaks, we used the published single-nucleotide mC datasets (Lister *et al.*, 2008) provided by Dr. Ryan Lister. The original mC site coordinates were based on the TAIR8 assembly, so we transformed them into TAIR9 coordinates using the Perl script provided by TAIR. The mC sites were searched within all AGO4 peaks as well as NC-peaks, and the mC density was calculated and compared between AGO4 peaks and NC-peaks for CG, CHG and CHH methylation types or altogether. We also used the recently published single-nucleotide mC datasets (Wierzbicki *et al*., 2012) to directly compare the mC density between Col-0 and *nrpe1* mutant plants. It is of note that in this comparison a corresponding Col-0 wild type dataset was used.

Characterizing the class and family of transposable elements in AGO4 peaks

To get the class and family summaries for the TEs in AGO4 peaks, we extracted all unique overlapping TEs and grouped them into different classes or families based on the RepeatMasker annotation information (described above). TEs annotated by TAIR were not included in this analysis.

Displaying the chromosome-distribution of AGO4 peaks

All AGO4-peak coordinates were plotted against their sizes for all 5 chromosomes; the reference gene-density and TE-density were calculated by dividing the chromosome into 100 kb bins. Only protein-coding genes were used for calculating the gene-density; both the TAIR annotated and RepeatMasker annotated TEs were used for calculating the TE-density.

Characterizing AGO4 binding profiles around TSS

The log fold-change profile of ChIP-seq reads between Col-0 and *ago4* samples was generated using the CEAS program (Shin *et al.*, 2009) with relative positions to the TSS of all protein-coding genes.

Characterizing nucleosome profiles around AGO4 peaks

To characterize the nucleosome profiles, we used published MNase-seq datasets (Chodavarapu *et al.*, 2010) from NCBI GEO (accessions GSE21673, GSM543296), and merged raw reads from all 6 replicate runs. The MNase-seq reads were mapped to the *Arabidopsis* genome using Bowtie using the same parameters as described for our ChIP-seq libraries, with the exception that we only kept uniquely mapping reads. The log fold-change profile of MNase-seq reads between Col-0 and *ago4* samples were generated using the CEAS program (Shin *et al.*, 2009) with relative positions to the TSS of all protein-coding genes. Well-positioned nucleosomes were then called as previously described (Kaplan *et al.*, 2009), and the nucleosome-density profiles were determined near all AGO4-peaks or for only promoter overlapping peaks, respectively. It is of note that all mapped MNase-seq reads were extended to 147 nt before calling the well-positioned nucleosomes, which is the known average nucleosome size for eukaryotic genomes.

Identification of enriched biological processes in the genes whose promoters are bound by AGO4

To identify significantly enriched biological processes for AGO4-bound promoters, the corresponding gene IDs (TAIR AGI) of these promoters were extracted and subjected to the GOEAST online Batch-Genes analysis tool (Zheng and Wang, 2008) with an FDR < 0.05, and other parameters set as default.

Identification of overlaps between AGO4 binding and regions of differential DNA methylation

DMRs identified by Dowen *et al.* (2012) were overlapped with AGO4 peaks using PeakAnalyzer (Salmon-Divon et al., 2010). p-values were derived from 1000 random permutations.

Detection of Pol V-dependent transcripts

Total RNA was isolated from 100 mg 2.5 weeks old seedlings (Col0, *nrpe1*, *ago4*) using the Plant RNeasy Mini kit (Qiagen) including on-column DNase treatment. To remove any potential residual DNA, 1 unit of Turbo DNase (Ambion) was added to 1 µg of total RNA and heat-inactivated after incubation at 25°C for 15 minutes. For cDNA synthesis, 500 ng of the DNase-treated RNA were converted to cDNA using the Random Primer Mix (NEB) and Superscript III Reverse Transcriptase (Invitrogen) following manufacturers' instructions. To detect potential contaminations by genomic DNA, we also prepared control samples lacking reverse transcriptase. Subsequent teal time PCR reactions were performed using Platinum Taq (Invitrogen) on a Bio-Rad CFX Connect Real-Time System.

SUPPLEMENTAL REFERENCES

Chodavarapu, R.K., Feng, S., Bernatavichute, Y.V., *et al.*, (2010) Relationship between nucleosome positioning and DNA methylation. *Nature*, 466(7304), 388–392.

Dowen, R.H., Pelizzola, M., Schmitz, R.J., Lister, R., Dowen, J.M., Nery, J.R., Dixon, J.E. and Ecker, J.R., (2012) Widespread dynamic DNA methylation in response to biotic stress. *Proceedings of the National Academy of Sciences of the United States of America*, 109(32), E2183-E2191.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O. and Walichiewicz, J., (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research*, 110(1-4), 462–467.

Kaplan, N., Moore, I.K., Fondufe-Mittendorf, Y., *et al.*, (2009) The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*, 458(7236), 362–366.

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 Biology*, 10(3), R25.

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(3), 523–536.

Muino, J.M., Kaufmann, K., Ham, R.C. van, Angenent, G.C. and Krajewski, P., (2011) ChIP-seq Analysis in R (CSAR): An R package for the statistical detection of protein-bound genomic regions. *Plant Methods*, 7, 11.

Quinlan, A.R. and Hall, I.M., (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6), 841–842.

Rowley, M.J., Avrutsky, M.I., Sifuentes, C.J., Pereira, L. and Wierzbicki, A.T., (2011) Independent chromatin binding of ARGONAUTE4 and SPT5L/KTF1 mediates transcriptional gene silencing. *PLoS Genetics*, 7(6), e1002120.

Shin, H., Liu, T., Manrai, A.K. and Liu, X.S., (2009) CEAS: cis-regulatory element annotation system. *Bioinformatics*, 25(19), 2605–2606.

Salmon-Divon, M., Dvinge, H., Tammoja, K., Bertone, P. (2010) PeakAnalyzer: genome-wide annotation of chromatin binding and modification loci. *BMC Bioinformatics*, 11, 415.

Wang, H., Zhang, X., Liu, J., *et al.*, (2011) Deep sequencing of small RNAs specifically associated with Arabidopsis AGO1 and AGO4 uncovers new AGO functions. *The Plant Journal: For Cell and Molecular Biology*, 67(2), 292–304.

Wierzbicki, A.T., Cocklin, R., Mayampurath, A., Lister, R., Rowley, M.J., Gregory, B.D., Ecker, J.R., Tang, H., Pikaard, C.S., (2012) Spatial and functional relationships among Pol V-associated loci, Pol IV-dependent siRNAs, and cytosine methylation in the *Arabidopsis* epigenome. *Genes & Development,* 26(16), 1825-1836.

Zheng, Q., Ryvkin, P., Li, F., Dragomir, I., Valladares, O., Yang, J., Cao, K., Wang, L.-S. and Gregory, B.D., (2010) Genome-wide double-stranded RNA sequencing reveals the functional significance of base-paired RNAs in Arabidopsis. *PLoS Genetics*, 6(9), e1001141.

Zheng, Q. and Wang, X.-J., (2008) GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. *Nucleic Acids Research*, 36, W358–363.

Zilberman, D., Cao, X. and Jacobsen, S.E., (2003) ARGONAUTE4 control of locus-specific siRNA accumulation and DNA and histone methylation. *Science*, 299(5607), 716–719.