UNIVERSITY OF CALIFORNIA
RIVERSIDE

Understanding Epigenetics: Molecular Mechanisms of siRNA Biogenesis and DNA Methylation

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Genetics, Genomics and Bioinformatics

by

Shaofang Li

December 2014

Dissertation Committee:
Dr. Xuemei Chen, Chairperson
Dr. Thomas Girke
Dr. Shouwei Ding

The Dissertation of Shaofang Li is approved:

_____

_____

_____

Committee Chairperson

University of California, Riverside

Acknowledgements

I am sincerely grateful to all of the people who provided help in numerous ways to make my dissertation possible.

First of all, I owe my deepest gratitude to my advisor, Dr. Xuemei Chen. I feel really lucky to have an advisor who's always there whenever you need help. In the five years of pursuing my Ph.D. degree in Dr. Chen's lab, Dr. Chen not only gave me consistent help and guidance but also served as a perfect model for what it means to be a scientist. From her, I learned how to think critically about experimental design, from a single experiment to a whole project. In addition, Dr. Chen's excellence as a female scientist really inspired me, and I am extremely grateful for all of her considerations and the help she provided to make my work possible.

Second, I'd like to thank Dr. Thomas Girke and the members of his lab. When I first began my Ph.D. studies, I had no background in bioinformatics, and working in Dr. Girke's lab for a short time helped me tremendously in mastering these valuable techniques.

I thank the members of my dissertation committee, Dr. Thomas Girke and Dr. Shou-wei Ding, and my qualifying exam committee, Dr. Thomas Girke, Dr. Shou-wei Ding, Dr. Stefano Lonardi, Dr. Shizhong Xu and Dr. Jun Li, for all of their effort, time and insightful suggestions.

I am grateful to all of the members of Dr. Chen's lab. They have accompanied me on my five-year scientific research journey, giving me both help and happiness. They are alongside me at times when I am frustrated with experimental results and at times when

the results are more exciting, in short, whenever I need help. Without their kind help and support, the past five years would not have been a happy process. In particular, I owe tremendous thanks to Dr. Lei Gao and Dr. Shengben Li.

The Genetics, Genomics and Bioinformatics graduate program provided both financial support and an ideal learning environment. The program seminars gave me a unique opportunity to present my work and receive critical input and the services provided by IIGB were of utmost importance for my research. I thank Deidra Kornfeld for her constant help over the past five years.

I am also grateful to Dr. Brian Gregory from the University of Pennsylvania for his assistance with the double-stranded RNA-seq libraries. Without his help, my project wouldn't have gone smoothly.

Last but not least, I would like to express my gratitude for all of my family members. I especially thank my loving husband, Dr. Huanbin Zhou: he is always there to accompany, encourage and support me. I feel so lucky to have my dearest two little cute sons (Frank Dingyi Zhou and Benjamin Shouyi Zhou), who understand why mommy is always busy. I also want to thank my parents for their support and help with my sons; without their constant support and understanding, it would be impossible for me to pursue scientific research.

Shaofang Li

Riverside, California

December 2014

ABSTRACT OF THE DISSERTATION

Understanding Epigenetics: Molecular Mechanisms of siRNA Biogenesis and DNA Methylation

by

Shaofang Li

Doctor of Philosophy, Graduate Program in Genetics, Genomics and Bioinformatics
University of California, Riverside, December 2014
Dr. Xuemei Chen, Chairperson

Although transposons constitute large portions of eukaryotic genomes, certain mechanisms have evolved to suppress the detrimental effects caused by the movement of transposons. In *Arabidopsis*, DNA methylation plays a vital role in suppressing transposon expression at the transcriptional level, and the underlying mechanisms have been thoroughly investigated. Numerous factors participating in the DNA methylation pathway have been reported, from the establishment of DNA methylation through RNA-directed DNA methylation (RdDM) to the maintenance of symmetrical DNA methylation by MET1/CMT3 and asymmetrical DNA methylation by RdDM and CMT2.

Despite this well-established framework, however, two important questions remain. The first concerns the mystery precursors to siRNAs that function as guidance signals for RdDM. Although it has been proposed that Pol IV transcribes methylated

DNA to produce primary transcripts at RdDM loci and that RDR2 converts these transcripts to dsRNAs to serve as siRNA precursors, no such siRNA precursor transcripts have been reported. In my Ph.D. studies, I was able to identify Pol IV/RDR2-dependent transcripts from tens of thousands of loci through genome-wide profiling of RNAs in genotypes with compromised siRNA precursor processing. On the one hand, Pol IV/RDR2-dependent transcripts differ from Pol II-dependent transcripts in the following ways: they correspond to both DNA strands instead of one strand, they have a 5' monophosphate instead of a 5' cap, they lack a polyA tail at the 3' end, and they do not have introns. On the other hand, both Pol IV/RDR2-transcribed regions and Pol II-transcribed regions are flanked by A/T-rich sequences depleted in nucleosomes. Computational analysis of siRNA abundance in various mutants also revealed differences in the regulation of siRNA biogenesis at two types of loci that undergo CHH methylation through two different DNA methyltransferases.

The second question is how the silencing effect of DNA methylation is controlled to prevent the stochastic silencing of genes or to allow the expression of genes that reside nearby transposons. In my Ph.D. studies, I identified SUVH1 as an anti-silencing factor through a forward genetic screen and showed that it promotes the expression of two transgenes and several endogenous genes. 5-Aza-2'-deoxycitidine (a DNA methylation inhibitor) treatment and methylation level analysis using McrBC-PCR and MethylC-seq subsequently showed that SUVH1 functions downstream of DNA methylation to promote the expression of genes harboring promoter DNA methylation. In addition, SUVH1 was found to maintain H3K4me3 levels. These findings from the functional studies of

SUVH1 shed light on the regulatory network acting at genes with various epigenetic marks.

Table of Contents

List of Figures

List of Tables

Chapter 1. The regulation of chromatin through DNA and histone modifications

The first step of epigenetic gene regulation occurs at the transcriptional level through chromatin. Understanding the modifications on DNA and histones, two basic components of chromatin, is critical for understanding the phenomenon of epigenetic gene regulation as a whole. Here, I summarize the current knowledge of DNA methylation, histone modification and the crosstalk between different epigenetic marks.

Introduction to epigenetic regulation

Gregor Mendel's studies of trait inheritance and the elucidation of the structure of DNA by James Watson and Francis Crick provided the foundation for our understanding of how traits are inherited from parents to children. However, DNA sequences alone cannot explain the fact that many different types of cells develop from embryonic stem cells all possessing the same genome; in fact, the term "epigenetics" was first used in the context of genetic studies of developmental processes (Bonasio et al. 2010).

At the end of 20th century, the discovery of RNAi by Craig Mello and his group (Fire et al. 1998) led to increased interest in the field of epigenetics. Nowadays, the word "epigenetics" sounds familiar to everybody. But what is epigenetics? To derive a consensus definition of epigenetics, the Banbury Conference Center and Cold Spring Harbor Laboratory hosted a special meeting in Dec. 2008. A 2009 report subsequently described an epigenetic trait as follows. "An epigenetic trait is a stably heritable phenotype resulting from changes in a chromosome without alterations in the DNA

sequence" (Berger et al. 2009). Later on, epigenetics was more broadly framed in Science by Danny Reinberg as a term referring to "the inheritance of variation (-genetics) above and beyond (epi-) changes in the DNA sequence" (Bonasio et al. 2010). A similar definition can be found on Wikipedia: "In biology, epigenetics is the study of cellular and physiological traits that are heritable by daughter cells and not caused by changes in the DNA sequence; Epigenetics describes the study of stable, long-term alterations in the transcriptional potential of a cell. " (http://en.wikipedia.org/wiki/Epigenetics).

In a more narrow sense, epigenetics refers only to inheritable changes occurring at the chromosome level (Berger et al. 2009). It has been proposed that three types of signals establish heritable epigenetic modifications: "Epigenator", "Epigenetics Initiator" and "Epigenetics Maintainer" (Berger et al. 2009). The Epigenator, representing the most upstream event, involves the sensing of an environmental change and signal transduction to the Epigenetics Initiator. The temperature change in the paramutation process is one example of the Epigenator signal type. The Epigenetics Initiator is the link between the Epigenator and Epigenetics Maintainer, and examples include long non-coding RNA, siRNAs and certain DNA-binding factors, insofar as they transduce the cell environmental signal to direct downstream epigenetic status establishment. The Epigenetics Maintainer maintains the chromatin status, permitting the status to be inherited by offspring. DNA methylation and histone modification are typical examples of the Epigenetics Maintainer (Berger et al. 2009).

The regulatory roles of DNA methylation

DNA methylation, the addition of a methyl group to a DNA nucleotide, is an important epigenetic modification that affects various biological processes. Three methylated DNA bases are known: 5-methylcytosine (m5C), 4-methylcytosine (m4C) and N^6-methyladenosine (m6A). In eukaryotes, DNA methylation usually refers to m5C, which is associated with the suppression of gene expression and transposon activity (Law and Jacobsen 2010); however, adenine methylation has also been reported in eukaryotes (Baniushin 2005). In prokaryotes, m6A is the major methylated base, with m5C and m4C occurring less frequently (Ratel et al. 2006; Fang et al. 2012). m6A is known to be essential for survival in several bacteria (Ratel et al. 2006) and is critical for numerous aspects of prokaryotic life, including the regulation of bacterial gene expression and virulence (Low et al. 2001) and DNA replication (Demarre and Chattoraj 2010). In a recent methylome profiling analysis of *Escherichia coli* K12 by whole-genome bisulfite sequencing, DNA cytosine methylation was found to be a regulator of stationary phase gene expression (Kahramanoglou et al. 2012).

Among the three methylated DNA bases, 5mC is the most well studied DNA methylation, and hereafter, DNA methylation will refer specifically to 5mC. DNA methylation is a conserved gene silencing mechanism critical for preserving genome integrity in many eukaryotes. A notable exception is the model organism *Caenorhabditis elegans*, which lacks genomic DNA methylation (Simpson et al. 1986). DNA methylation was also thought to be absent in yeast and *Drosophila melanogaster*, but low levels of DNA methylation have now been reported in these organisms (Lyko et al. 2000; Tang et

al. 2012; Capuano et al. 2014). In animals, DNA methylation predominantly occurs in the CG context, with non-CG methylation rarely detected. Non-CG methylation has recently been reported in oocytes, pluripotent embryonic stem cells and mature neurons (Xie et al. 2012b; Lister et al. 2013; Shirane et al. 2013; Wu and Zhang 2014), but its precise role remains to be discovered. DNA methylation is associated with several key developmental processes in animals, including genome imprinting, transposon suppression and X-chromosome inactivation (Feng et al. 2010). The vital roles of DNA methylation are further supported by the fact that loss of DNA methylation leads to embryonic lethality in animals (Law and Jacobsen 2010). DNA methylation and/or the incorrect transmission of DNA methylation patterns have been associated with aging (Horvath 2013) and several diseases, including cancer (Fukushige and Horii 2013) and atherosclerosis (Zaina and Lund 2013). In plants, DNA methylation commonly occurs in both CG and non-CG contexts, which are further characterized as symmetric (CG and CHG, where H = A, T or C) or asymmetric (CHH). The major function of DNA methylation is to control transposon activity to maintain genome integrity. In *Arabidopsis*, the repression of transposon activity involves a triple-layer pathway, with two layers of transcriptional gene silencing (TGS) achieved though DNA methylation and a third layer of posttranscriptional gene silencing (PTGS) (Bourc'his and Voinnet 2010).

DNA methylation maintenance

To maintain CG methylation during replication in mammals, DNA (cytosine-5)-methyltransferase 1 (DNMT1) is recruited to replication foci through interactions with

the proliferating cell nuclear antigen component of the replication machinery (Chuang et al. 1997) and a chromatin-associated protein, ubiquitin-like plant homeodomain and RING finger domain 1 (UHRF1), that specifically binds to hemimethylated CG dinucleotides through the SET and RING finger associated (SRA) domain (Bostick et al. 2007; Sharif et al. 2007; Arita et al. 2008). Similar CG methylation maintenance mechanisms are found in plants. METHYLTRANSFERASE 1 (MET1), a homolog of DNMT1, is responsible for all CG methylation, as evidenced by the genome-wide elimination of CG methylation in *met1* (Vongs et al. 1993; Stroud et al. 2013). In *Arabidopsis*, CG methylation also requires three VARIANT IN METHYLATION (VIM) proteins, which are SRA domain-containing homologs of UHRF1 (Woo et al. 2007). In the *vim1 vim2 vim3* triple mutant, the decrease in CG methylation resembles that observed in *met1* (Stroud et al. 2013). DECREASED DNA METHYLATION 1 (DDM1), a SWI2/SN2-like chromatin remodeler, controls CG methylation through its ATPase activity and nucleosome remodeling (Hirochika et al. 2000; Stroud et al. 2013).

The maintenance of CHG methylation in plants requires the plant-specific methyltransferase CHROMOMETHYLASE 3 (CMT3) (Lindroth et al. 2001). CMT3 binds H3K9me2-containing nucleosomes through both its bromo adjacent homology (BAH) and chromo domains (Du et al. 2012). H3K9 is methylated by KRYPTONITE/SU(VAR) 3-9 HOMOLOG 4 (KYP/SUVH4) (Jackson et al. 2002) and its homologs SUVH5 and SUVH6 (Ebbs and Bender 2006; Rajakumara et al. 2011; Stroud et al. 2014), which possess SRA domains that recognize CHH and CHG methylation (Ebbs and Bender 2006; Johnson et al. 2007). The reinforcing loop between

DNA methylation and histone methylation is evidenced by the high correlation of these marks on a genome-wide scale (Stroud et al. 2013; Stroud et al. 2014). The strong relationship between DNA and histone methylation is also observed in mammals, although most cases involve protein interactions between the DNA and histone methyltransferases (Cedar and Bergman 2009). Methylation in the asymmetric CHH context requires *de novo* methylation involving two distinct methyltransferases that will be introduced below.

The establishment of DNA methylation through *de novo* methylation

Genome-wide reprogramming of DNA methylation occurs in both plant and animal development. In mammalian development, DNA methylation is erased in the primordial germ cells and early embryo cells (Feng et al. 2010) then reestablished through the *de novo* methyltransferases DNA (cytosine-5)-methyltransferase 3A and 3B (DNMT3A and DNMT3B) (Okano et al. 1998; Okano et al. 1999). DNA (cytosine-5)-methyltransferase 3-like (DNMT3L), a DNMT3 homolog with no catalytic activities, is also essential for the establishment of DNA methylation alongside DNMT3A and DNMT3B (Hata et al. 2002). In *Arabidopsis*, DNA methylation is erased in the central cell in the female gametophyte and reestablished through the *de novo* methyltransferase DOMAINS REARRANGED METHYLTRANSERASE 2 (DRM2), a homolog of DNMT3 (Feng et al. 2010). DOMAINS REARRANGED METHYLTRANSERASE 3 (DRM3), a catalytically mutated DRM2 homolog, is also required for DNA methylation maintained by DRM2 (Henderson et al. 2010). Recently, the plant specific protein

CHROMOMETHYLASE 2 (CMT2) was characterized as another *de novo* methyltransferase acting through interactions with DDM1 and histone H1 (Zemach et al. 2013).

In plants, DRM2-mediated *de novo* methylation, or RNA-directed DNA methylation (RdDM), was first described in 1994 (Wassenegger et al. 1994). RdDM requires siRNAs as the guidance signal and core RNAi machinery for the recruitment of DRM2 to methylate cystosines at the corresponding sites in the genome. Using *Arabidopsis* as a model system, numerous RdDM factors have been identified, with roles in Pol IV-mediated siRNA biogenesis, DRM2-mediated DNA methylation and downstream chromatin alterations.

P4siRNA biogenesis

The initial step of RdDM is the generation of 24 nt siRNAs (Law and Jacobsen 2010). RNA polymerase IV (Pol IV), a plant-specific RNA polymerase, has been proposed to generate the primary transcripts for these 24 nt siRNAs (hereafter referred to as P4siRNAs) (Zhang et al. 2007; Mosher et al. 2008). The transcripts are then converted into double-stranded RNAs (dsRNAs) by RNA-DEPENDENT RNA POLYMERASE2 (RDR2) (Xie et al. 2004; Jia et al. 2009). As described in Chapter 2, genome-wide profiling data of Pol IV/RDR2-dependent transcripts provide more direct evidence that Pol IV and RDR2 are indeed responsible for the generation of P4siRNAs. Mass-spectrometric analysis of NRPD1 affinity purifications helped identify Pol IV complex proteins (Law et al. 2011), which include the following: subunit proteins specific to Pol

IV (NRPD1 and NRPD7A); proteins shared by Pol IV and RNA polymerase V (Pol V) (NRPD2/E2, NRPD3B/E3B, NRPD4/E4, NRPD5B/E5B and NRPD7B/E7B); proteins shared by RNA polymerase II (Pol II) and Pol IV (NRPB5/D5); and proteins shared by Pol II, Pol IV and Pol V (NRPB3/D3/E3A, NRPB6A/D6A/E6A, NRPB8B/D8B/E8B, NRPB9B/D9B/E9B, NRPB10/D10/E10, NRPB11/D11/E11 and NRPB12/D12/E1) (Law et al. 2011). In addition to the Pol IV subunit proteins, RdDM proteins were also identified, including RDR2, RNA-DIRECTED DNA METHYLATION 4 (DMS4), CLASSY 1 (CLSY1), CLASSY 2, CLASSY 3 and SAWADEE HOMEODOMAIN HOMOLOG 1 (SHH1) (Law et al. 2011).

Although the Pol IV complex contains several RdDM proteins, only RDR2 is as critical as Pol IV for P4siRNA biogenesis (Kasschau et al. 2007) (Chapter 2), and the interaction between Pol IV and RDR2 have been confirmed by co-immunoprecipitation experiments (Law et al. 2011; Haag et al. 2012). *In vitro* transcription analysis of Pol IV-RDR2 complex proteins using different mutant versions indicated that Pol IV activity does not require RDR2 and that RDR2 is not functional in the absence of Pol IV (Haag et al. 2012). Based on these findings, it was proposed that the activities of RDR2 and Pol IV are coupled for the synthesis of dsRNAs (Pikaard et al. 2012). The failure to detect P4siRNA precursors in the *rdr2* background prompted us to re-evaluate the role of RDR2 (Chapter 2). For example, RDR2 may be essential for Pol IV activity *in vivo*, either by directly affecting Pol IV activity or the recruitment of Pol IV to chromatin loci. Alternatively, RDR2 may simply affect the stability of Pol IV-dependent transcripts.

SHH1, another Pol IV complex component, was also identified in a forward genetic screen in the *ros1* mutant background with a silenced *RD29A-LUC* transgene (Liu et al. 2011a). In *shh1*, both DNA methylation and P4siRNA abundance are decreased (Law et al. 2011; Liu et al. 2011a) (Chapter 2), indicating the involvement of SHH1 in RdDM. Moreover, decreased Pol IV-occupation in *shh1* indicates a role of SHH1 in Pol IV recruitment. The crystal structure of the SHH1 SAWADEE domain suggests that this particular domain adopts a tandem Tudor domain-like fold and functions as a chromatin-binding module to read unmethylated H3K4 and methylated H3K9 on histone tails (Law et al. 2013).

The Pol IV complex protein RDM4 was identified though forward genetic screens using two reporter lines under RdDM regulation (He et al. 2009b; Kanno et al. 2010). In contrast to other RdDM proteins, the *rdm4* mutation leads to developmental phenotypes of short siliques and partial sterility (He et al. 2009b). *RDM4* encodes a protein conserved in yeast, *Drosophila* and human. The yeast homolog IWR1 is characterized as a transcription factor that interacts with Pol II. In *Arabidopsis*, RDM4 has been shown to interact with the largest subunit of Pol II (He et al. 2009b). In the *rdm4* mutant, P4siRNAs, Pol V-dependent scaffold transcripts and Pol II-dependent genes are all affected, indicating that RDM4 may be a transcription factor that interacts with several RNA polymerases (Kanno et al. 2010). CLASSY1, an SNF2 domain-containing protein, was also identified from a forward genetic screen using a silencing signal reporter line (Smith et al. 2007). The same studies indicated that CLASSY1 acts together with RDR2

and NRPD1 in P4siRNA biogenesis and the spread of the transgene silencing signal, which is consistent with the identification of CLASSY1 as a Pol IV-complex protein.

After the dsRNAs have been generated by the Pol IV/RDR2 complex, the ribonuclease III family protein DICER-LIKE 3 (DCL3) cleaves the dsRNAs to generate 24 nt siRNAs (Cho et al. 2008; Liu et al. 2009). In *Arabidopsis*, there are four DICER-LIKE (DCL) proteins, DCL1, DCL2, DCL3 and DCL4. DCL1 is a miRNA biogenesis factor that cleaves pri-miRNAs into pre-miRNAs and pre-miRNAs into mature 21 nt miRNAs (Xie et al. 2003; Chen 2009). DCL2, DCL3 and DCL4 are associated with different types of siRNAs. DCL2 generates 22 nt siRNAs from natural cis-acting antisense transcripts (Mlotshwa et al. 2008) and is required for the biogenesis of virus/fungal-induced siRNAs (Garcia-Ruiz et al. 2010; Weiberg et al. 2013). DCL4 is required for the biogenesis of 21 nt tasiRNAs and post-transcriptional silencing processes (Liu et al. 2007) and has also been reported to participate in the biogenesis of virus-derived siRNAs (Garcia-Ruiz et al. 2010; Wang et al. 2011). DCL3 generates 24 nt P4siRNAs from heterochromatic regions to direct *de novo* methylation (Xie et al. 2004). When DCL3 is absent, however, DCL2 and DCL4 act redundantly to produce P4siRNAs (Henderson et al. 2006) (Chapter 2). Once P4siRNAs have been produced, the methyltransferase HUA ENHANCER 1 (HEN1) adds a methyl group to the 3’ terminal nucleotides (Yu et al. 2010).

DRM2-mediated DNA methylation

After the generation of P4siRNA duplexes in the nucleus, the duplexes are exported into the cytoplasm and loaded onto ARGONAUTE 4 (AGO4) to form the RISC complex through the activity of HSP90 (Iki et al. 2010; Ye et al. 2012). AGO4 was first named based on its homology to AGO1 and was cloned from a forward genetic screen for mutants with suppressed silencing of the *Arabidopsis SUPERMAN* (*SUP*) gene (Zilberman et al. 2003). In *Arabidopsis*, there are ten AGO proteins, which are divided into three clades (Vaucheret 2008). The four AGO proteins in the AGO4 clade, AGO4, AGO6, AGO8 and AGO9, play partially redundant roles and preferentially bind small RNAs with a 5' adenosine (Mi et al. 2008; Mallory and Vaucheret 2010). AGO6 was isolated in a genetic screen for TGS factors using the *ros1* mutant and acts redundantly with AGO4 (Zheng et al. 2007). AGO9 has been shown to bind 24 nt small RNAs *in vitro* and is necessary for suppressing long terminal repeat retrotransposons in the ovule (Duran-Figueroa and Vielle-Calzada 2010). AGO8 is probably a pseudogene, considering its low expression and a splicing-induced frame-shift (Takeda et al. 2008). The binding of AGO4 to P4siRNAs leads to a conformational change that exposes the nuclear localization signal, which facilitates the redistribution of AGO4-P4siRNAs into the nucleus. AGO4/Pol V/P4siRNAs complex is assembled in Cajal bodies (Li et al. 2006a; Pontes et al. 2006) but also facilitated by the conserved AGO-binding GW/WG motif in the C-terminal domain of NRPE1, the largest Pol V subunit (El-Shami et al. 2007; Till and Ladurner 2007). The AGO4-P4siRNA complex is recruited to chromatin loci by Pol V-generated scaffold transcripts (Wierzbicki et al. 2009).

The plant-specific RNA polymerases Pol IV and Pol V are both composed of 12 subunits that are paralogous or identical to the subunits of Pol II (Ream et al. 2009). As indicated above, Pol V transcribes long non-coding RNAs, which serve as scaffold transcripts that facilitate heterochromatin formation and the silencing of overlapping and adjacent genes (Wierzbicki et al. 2008). These Pol V-generated transcripts have different RNA structures than Pol II- and Pol IV- generated transcripts. Pol II-generated RNAs typically encode functional proteins and have a 5' cap and a 3' polyA tail. Pol IV-generated RNAs function as precursors of P4siRNAs, have a 5' monophosphate and lack the 3' polyA tail (Chapter 2). Finally, Pol V-generated scaffold RNAs have a 5' cap and also lack the 3' polyA tail (Wierzbicki et al. 2008). The DDR protein complex (DRD1, DMS3 and RDM1) and SUVH2/9 are required for the recruitment of Pol V to the chromatin loci (Law et al. 2010; Zhong et al. 2012; Johnson et al. 2014; Liu et al. 2014).

The scaffold transcripts generated by Pol V at the chromatin loci can also recruit the *de novo* methyltransferase DRM2 (Law and Jacobsen 2010; Matzke and Mosher 2014). In *Arabidopsis*, there are three DRMs: DRM1, DRM2 and DRM3. Although both DRM1 and DRM2 are active methyltransferases, DRM2 is recognized as the major player because *drm2* and *drm1 drm2* have similar CHH methylation patterns (Cao and Jacobsen 2002). Because the catalytic motifs of DRM3 are rearranged, DRM3 is not an active methyltransferase, but it participates in the RdDM pathway by affecting DRM2 activity (Henderson et al. 2010). Recently, the crystal structure of the methyltransferase domain in *Nicotiana tabacum* DRM (NtDRM) revealed that NtDRM forms a homodimer critical for its catalytic activity (Zhong et al. 2014). In addition, *Arabidopsis* DRM2 has

been shown to occur in the same complex as AGO4 and preferentially methylates the DNA strand that acts as the template for Pol V and has greater P4siRNA abundance (Zhong et al. 2014).

Additional factors participating in RdDM

There are numerous RdDM factors in addition to those introduced above. KOW DOMAIN-CONTAINING TRANSCRIPTION FACTOR 1 (KTF1) was identified as an RdDM factor from both a forward genetic screen and from searching for AGO-interacting GW/WG motif-containing proteins (Bies-Etheve et al. 2009; He et al. 2009a). KTF1, a homolog of SPT5 elongation factor, binds to chromatin loci subject to TGS and functions as a facultative RNAP elongation factor (Rowley et al. 2011). This binding occurs downstream of Pol V and parallel to (i.e., independently of) AGO4 binding (He et al. 2009a). AtMORC1 and AtMORC6, members of the conserved Microchidia (MORC) adenosine triphosphatase (ATPase) family, were identified in a forward genetics screen for mutants with increased *SDC-GFP* expression (Moissiard et al. 2012). AtMORC1 and AtMORC6 are required for the heterochromatin condensation that leads to TGS through modest changes in DNA methylation (Moissiard et al. 2012). Using the GFP reporter system, the ability of AtMORC6 to form high order chromatin structure was found to influence RdDM and to be required for the efficient initiation or maintenance of DNA methylation at some loci (Brabbs et al. 2013).

INVOLVED IN DE NOVO 2 (IDN2), a homolog of SUPPRESSOR OF GENE SILENCING 3 (SGS3), was identified in three independent forward genetics screens.

The *IDN2* gene was first isolated from a screen of a collection of T-DNA insertion mutants using *FWA* and *Agrobacterium tumefaciens*-mediated transformation in 2009 (Ausin et al. 2009). It was later isolated in a *ros1* suppressor screen (Zheng et al. 2010b) and in a screen for RdDM mutations using a *ProNOS-NPTII* reporter construct (Finke et al. 2012). Mass spectrometric analysis of IDN2 purification products indicated that IDN2 forms a complex with two partially redundant paralogs, IDN2 PARALOG 1 (IDP1) and IDN2 PARALOG 2 (IDP2), and that IDN2 acts downstream of the RdDM pathway (Ausin et al. 2012b; Xie et al. 2012a; Zhang et al. 2012). In contrast to IDP1 and IDP2, the RNA recognition motif of the IDN2 XS domain permits the binding of dsRNAs by IDN2; additionally, the XH domain of IDN2 is required for its interaction with IDP1 and IDP2 (Ausin et al. 2012b; Zhang et al. 2012). The interaction of the IDN2-IDP complex with the SWI/SNF chromatin remodeling complex may stabilize the base-pairing between P4siRNAs and Pol V-generated scaffold transcripts and stabilize the nucleosome positions (Zhu et al. 2013; Matzke and Mosher 2014).

Recently, splicing factors were also discovered as RdDM pathway components (Huang and Zhu 2014). In a forward genetic screen using an *FWA* transgene as the reporter, ARGININE/SERINE-RICH 45 (SR45), a member of a highly conserved family of spliceosome proteins, was isolated from a late-flowering mutant (Ausin et al. 2012a). In another screen using *RD29A-LUC* and *35S-NPTII* constructs, two splicing factors were identified (Huang et al. 2013; Zhang et al. 2013). The pre-mRNA splicing factor RDM16, a component of the U4/U6 snRNP protein complex, is required for biogenesis of Pol V – dependent scaffold transcripts but not that of P4siRNAs (Huang et al. 2013). ZOP1, an

OCRE domain-containing protein, was identified as a splicing factor through its interactions with spliceosome and intron-retention components; the *zop1* mutant exhibits both reduced DNA methylation and lower P4siRNA abundance (Zhang et al. 2013). The RRP6-like splicing factor STA1 was isolated as a DNA methylation factor using methylation-sensitive Chop-PCR of the *AtSN1* locus from a pool of T-DNA insertion mutants (Zhang et al. 2014). Similar to the *zop1* mutant, both DNA methylation and P4siRNA abundance are decreased in *sta1*. Because P4siRNA precursors lack introns, the effect of splicing factors on P4siRNA abundance must be indirect probably through CHH DNA methylation (Chapter 2).

CMT2-mediated *de novo* DNA methylation

RdDM is a well-established *de novo* methylation pathway targeting loci spread throughout euchromatic regions (Chapter 2) (Wierzbicki et al. 2012). In *Arabidopsis*, the methyltransferase CMT2, a homolog of CMT3, is responsible for the maintenance of CHH methylation that is concentrated at pericentromeric regions (Chapter 2). Among the three chromomethylases in *Arabidopsis* (CMT1, CMT2 and CMT3), CMT1 appears to play a minimal role based on its low expression levels and truncated form in many *Arabidopsis* ecotypes (Henikoff and Comai 1998). CMT3 preferentially methylates CHG over CHH (Du et al. 2012), while CMT2 methylates both CHG and CHH sites with high activity *in vitro* (Stroud et al. 2014). The greater loss of CHG methylation in *cmt2 cmt3* than in *cmt3* is indicative of the redundant role of CMT2 in CHG methylation *in vivo*.

Additionally, the strong elimination of CHH methylation in *cmt2* indicates that CMT2 preferentially methylates CHH loci (Stroud et al. 2014).

The nucleosome remodeler DDM1 can facilitate CMT2 access at H1-containing heterochromatic regions (Zemach et al. 2013). Together, DDM1 and CMT2 tend to methylate CHH at long transposons at pericentromeric regions, while DRM2-mediated RdDM tends to target CHH sites of short transposons dispersed along the chromosome arms (Chapter 2) (Zemach et al. 2013). In other words, CMT2 and DRM2 can control all of the CHH methylation throughout the genome with almost no overlapping sites between them (Chapter 2) (Stroud et al. 2014). In *Arabidopsis*, P4siRNAs and CHH methylation are highly correlated and peak at the pericentromeric regions (Chapter 2). In a mutant with disrupted Pol IV function, the loss of P4siRNAs leads to CHH methylation only at DRM2-targeted loci in the euchromatic arms, where CHH methylation and P4siRNA abundance are high. Although P4siRNAs are produced at low levels at CMT2-targeted pericentromeric loci, the loss of P4siRNAs does not lead to the loss of CHH methylation, which indicates that P4siRNAs are not required in the CMT2/DDM1 methylation pathway (Chapter 2).

The active demethylation process

Active demethylation is important for developmental processes and prevents stochastic methylation at gene regions in both plants and animals. Global epigenetic reprogramming in primordial germ cells and in the early embryo involves active demethylation and the loss of histone modification. It has been proposed that active demethylation involves

oxidation or deamination and the base excision repair (BER) pathway (Kohli and Zhang 2013). Ten-eleven-translocation (TET), a 5-methylcytosine hydroxylase, can modify 5mC through oxidation to generate 5-hydroxymethylcytosine (5hmC) (Tahiliani et al. 2009). Moreover, TET can sequentially oxidize 5mC to generate 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) (Ito et al. 2011). AID (activation-induced deaminase) and APOBEC (apolipoprotein B mRNA-editing catalytic polypeptides) can deaminate 5hmC, but not 5mC, into 5hmU (Bhutani et al. 2010). Thymine DNA glycosylase (TDG) exhibits high glycosylase activity on 5hmU and 5caC rather than 5hmC (Cortellino et al. 2011), and the gap cleaved by TDG is replaced with unmethylated cytosine through the BER pathway to complete the active demethylation process (Gong and Zhu 2011).

The demethylation process has been studied more thoroughly in plants than in animals, and this study was initiated by the discovery of two glycosylases, DEMETER (DME) and REPRESSOR OF SILENCING 1 (ROS1). *DME* was identified from a mutant with parent-of-origin effects on seed viability and with seed abortion caused by impaired endosperm and embryo development (Choi et al. 2002). DME, a 5'-methylcytosine glycosylase, is primarily expressed in the central cell of the female gametophyte and activates the expression of maternal imprinted genes, such as *FWA*, *MEA*, *FIS2* and *MPC* (Zhu 2009). *ROS1*, also known as *DEMETER-LIKE 1* (*DML1*), was discovered in a forward genetic screen using a luciferase reporter gene driven by the RD29A promoter, which is sensitive to salt, drought, cold and abscisic acid (Gong et al. 2002). Unlike the restricted expression of DMR, ROS1 is widely expressed in all tissues, which suggests a more general role of ROS1 in the demethylation process. In *Arabidopsis*,

the DEMETER-LIKE (DML) proteins DML2 and DML3 are expressed in a wide-range of plant tissues and act redundantly with ROS1 in active demethylation (Lister et al. 2008; Ortega-Galisteo et al. 2008). Although DNA glycosylases can be classified as mono-functional or bi-functional, all of these four *Arabidopsis* proteins have both glycosylase activity (to hydrolyze the glycosylic bond between a base and deoxyribose) and lyase activity (to nick the DNA backbone at the abasic site) (Zhu 2009). With both of these enzymatic activities, the DML glycosylase family can cleave the *N*-glycosidic bond to release the methylated cytosine, thereby generating an abasic site, then break the phosphodiester linkage to generate a single nucleotide gap in the methylated DNA. This gap is subsequently repaired by as yet unidentified DNA polymerases and DNA ligases through the BER pathway (Law and Jacobsen 2010).

In addition to the glycosylase proteins, several other factors facilitate the active demethylation function of ROS1. ROS3, identified from the same screen as ROS1, contains an RNA recognition motif that binds small RNAs (Zheng et al. 2008). In light of the fact that ROS1 and ROS3 act in the same genetic pathway and co-localize throughout the nucleus, it was proposed that DNA demethylation by ROS1 is targeted to specific sequences by ROS3-bound RNAs (Zheng et al. 2008). *INCREASED DNA METHYLATION 1* (*IDM1*)*/ROS4*, a histone acetyltransferase, was identified in two independent screens, the screen in which ROS1 was identified and a screen using Chop-PCR to detect DNA methylation levels in a collection of T-DNA insertion mutants (Li et al. 2012; Qian et al. 2012). IDM1/ROS4 was shown to function in the ROS1 pathway through single-loci DNA methylation level analysis and the overlap of targeted loci in

genome-wide analysis of single and double mutants. IDM1/ROS4 acetylates H3 at chromatin sites without H3K4 di- or trimethylation to create a chromatin environment for ROS1 function (Qian et al. 2012). Like ROS4, ROS5 was identified in two independent screens and is also known as *IDM2* (Qian et al. 2014; Zhao et al. 2014). IDM2/ROS5, a protein in the small heat shock protein family with an α-crystallin domain, physically interacts with IDM1/ROS4 and partially colocalizes with IDM1/ROS4 in the nucleus. These properties indicate that IDM2/ROS5 participates in active DNA demethylation by regulating IDM1/ROS4 (Qian et al. 2014; Zhao et al. 2014).

ZDP was identified as a ROS1 pathway factor through its homology to polynucleotide kinase 3’ phosphatase (PNKP) in animals (Martinez-Macias et al. 2012). In the active demethylation process of ROS1, the cleavage of the phosphodiester backbone by β-elimination generates a single nucleotide gap flanked by 3’ phosphate and 5’ phosphate termini. Since all of the DNA polymerases require a 3’ hydroxyl terminus to initiate synthesis, the 3’ phosphate must be removed, which is performed by PNKP in animals (Jilani et al. 1999). In plants, the PNKP homolog ZDP can remove the blocking 3’ phosphate to permit the subsequent activity of the BER pathway. Additionally, ZDP interacts with ROS1 *in vitro*, and the two proteins colocalize in nucleoplasmic foci *in vivo*. Methylome analysis of *zdp* uncovered hundreds of hypermethylated endogenous loci, indicating that ZDP functions downstream of the ROS1 demethylation pathway (Martinez-Macias et al. 2012).

The regulatory roles of histone modifications

Histones are a family of highly alkaline proteins with positive charges that allow them to associate with negatively charged DNA to form nucleosomes, the core components of chromatin (Kornberg 1977). There are five major histone families: the core histones, H2A, H2B, H3 and H4, and the linker histones, H1 and H5 (Berger 2001; Fan and Roberts 2006). Although the basic histone octamer structure of nucleosomes always includes two copies of each of the four core histone proteins, the octamer can be modified to regulate gene function, either through different histone subunit variants or post-translational modifications of the histone subunits.

Histone variants may differ from a major histone by only a few amino acids and confer specific effects on nucleosome structure and function (http://www.nature.com/subjects/histone-variants). Histone variants contribute to a variety of chromatin functions, including transcriptional repression and activation, heterochromatic barriers, genome stability, DNA repair and chromatin segregation (Kamakaka and Biggins 2005). The majority of histone variants are H2A and H3 subtypes, while H2B has only a limited number of variants. No variants have been detected for H4. In *Arabidopsis*, H2A.Z has been correlated with DNA methylation and associated with responses to environmental and developmental stimuli (Coleman-Derr and Zilberman 2012). Another histone variant, H2A.W, is required for heterochromatin condensation and functions together with H3K9me2 and DNA methylation marks to control transposon expression (Yelagandula et al. 2014). The two main histone H3 variants (H3.1 and H3.3) have distinct locations and functions. In *Arabidopsis*, H3.1 is associated with silencing histone marks including H3K27, H3K9 and DNA methylation, while H3.3 is associated with actively transcribed

genes; the active histone marks include histone H3K4 methylation and H2B ubiquitynation (Stroud et al. 2012).

Since the pioneering studies by Vincent Allfrey on the possible roles of histone acetylation and methylation on gene expression regulation (Allfrey et al. 1964), the effects of post-translational modifications of histones are now more fully understood. At present, more than 200 distinct post-translational histone modifications have been identified, including acetylation, phosphorylation, methylation, ubiquitynation, sumoylation, deimination, β-N-acetylglucosamine, ADP ribosylation, histone tail clipping, histone proline isomerization and histone lysine crotonylation (Bannister and Kouzarides 2011; Tan et al. 2011).

Histone acetylation is the addition of an acetyl group to the ε-amino group of lysine side chains, with dynamic regulation by histone acetyltransferases (HATs) and histone deacetylases (HDACs). The addition of a negatively charged acetyl group can neutralize a positively charged lysine and loosen the chromatin structure to allow active transcription (Bannister and Kouzarides 2011). Histone phosphorylation is the addition of a negatively charged phosphate group onto serines, threonines and tyrosines preferentially within the N-terminal histone tail. As with histone acetylation, histone phosphorylation is a dynamic active mark regulated by two opposing enzymes: kinases, which phosphorylate histones, and phosphatases, which remove the phosphate group. Histone phosphorylation can either associate with condensed or de-condensed chromatin depending on the modification locus (Wei et al. 1998; Strahl and Allis 2000). Histone ubiquitynation and sumoylation are the addition of mono- or poly-ubiquitin proteins or small ubiquitin-like modifier (SUMO) proteins, respectively, and result in relatively

small molecular changes to amino acid side chains compared to other modifications. Histone ubiquitynation, achieved through the sequential action of E1-activating, E2-conjugating and E3-ligating enzymes, primarily occurs at H2A and H2B and participates in many regulatory processes within the nucleus, including transcription initiation and elongation, silencing and DNA repair (Sridhar et al. 2007; Weake and Workman 2008; Schmitz et al. 2009). Histone sumoylation has been detected on all core histone subunits and plays a role in transcriptional repression by antagonizing acetylation and ubiquitynation (Shiio and Eisenman 2003; Nathan et al. 2006).

Histone methylation primarily occurs on lysine and arginine. Histone arginine methylation is performed by a complex that includes protein arginine methyltransferase (PRMT), and while the histone lysine methylation requires a specific histone methyltransferase (HMT) containing a conserved SET domain. Unlike other histone modifications, the histone modification does not change the chemical structure of the histone, however, it can be recognized by proteins with Tudor, chromo, PWWP, MBT or PHD domains (Bannister and Kouzarides 2011). For example, the human Spindlin1, a protein with triple Ttudor-like Spin/Ssty repeats, can sense a cis-tail histone H3 methylation pattern involving trimethyllysine 4 (H3K4me3) and asymmetric dimethylarginine 8 (H3R8me2a) marks (Su et al. 2014).

Histone lysine methylation is one of the best studied epigenetic marks and occurs at several positions, including H3K4, H3K9, H3K14, H3K27, H3K36, H3K79 and H4K20; additionally, the number of methyl groups added can vary (Berger 2001; Roudier et al. 2011). Depending on the position and number of methyl groups added, histone

lysine methylation could be an active or repressive mark. For example, H3K4me3 is a conserved active mark observed in many organisms including *Tetrahymena*, yeast and *Arabidopsis* (Strahl et al. 1999; Santos-Rosa et al. 2002; Zhang et al. 2009). Studies in human cells indicate that H3K4me3 activates gene transcription at two mechanistic levels (Nishioka et al. 2002). First, H3K4me3 inhibits the association of the deacetylase NuRD complex with the H3 tail. Second, H3K4me3 specifically impairs Suv39h1-mediated H3K9me2, thereby thwarting heterochromatin formation. Other studies have shown that H3K4me3 can recruit chromatin remodeling factors, including chromodomain helicase DNA binding protein (CHD1) and bromodomain and PHD domain transcription factor (BPTF), to open chromatin (Flanagan et al. 2005; Li et al. 2006b) and prevent the binding of repressive complexes such as inhibitor of acetyltransferases (INHAT) (Schneider et al. 2004). In *Arabidopsis*, H3K4me3 marks are deposited by ATX1 and ATXR3 and are primarily located in promoters and 5' genic regions, in a manner mutually exclusive with DNA methylated regions (Zhang et al. 2009; Guo et al. 2010). H3K9 methylation is a well-studied repressive mark maintained by SUV39H or G9a (Rea et al. 2000; Tachibana et al. 2002), and numerous studies have demonstrated the silencing effects of H3K9 methylation. These effects have been particularly well explored using *Xenopus* oocytes (Stewart et al. 2005). H3K9 methylation can suppress gene expression through HETEROCHROMATIN PROTEIN 1 (HP1) recruitment or through a mechanism involving histone deacetylation. Because direct interaction between SUV39H1 and HP1 is necessary for HP1 recruitment in addition to H3K9 methylation, SUVH39H1-targeted H3K9 loci can recruit HP1, while G9a-targeted loci cannot. In plants, the most abundant

H3K9 methylation is H3K9me2, which is maintained by SUVH4/5/6 (Stroud et al. 2014). As previously introduced, H3K9me2 is important for the recruitment of CMT3, a major CHG methyltransferase for CHG methylation maintenance.

Arabidopsis SET domain proteins

The SET domain was first recognized as a conserved domain in the following *Drosophila* proteins: Suppressor of variegation 3-9 (Su(var)3-9) (Tschiersch et al. 1994), Enhancer of zeste (E(z)) (Jones and Gelbart 1993) and TRITHORAX (TRX) (Stassen et al. 1995). All of the presently known histone lysine methyltransferases contain a SET domain harboring methyltransferase activity, with only one exception: DISRUPTOR OF TELOMERIC SILENCING 1 (DOT1, also called KMT4) and DOT1-LIKE (DOT1L) possess histone methyltransferase activity toward histone H3K79 but do not have SET domains (Nguyen and Zhang 2011).

The SET domain proteins in maize and *Arabidopsis*, a monocot and dicot, respectively, can be grouped into five classes based on phylogenetic analysis and domain organization (Springer et al. 2003). The SET domain proteins are also known as SET DOMAIN GROUP (SDG) proteins, and some of these proteins are disrupted by insertions 50 to 120 amino acids in length in the SET domain (Jenuwein and Allis 2001). In *Arabidopsis*, class I SET proteins include the following E(z) orthologs: EZA1/SDG10, CLF/SDG1 and MEA/SDG5. In addition to the SET domain, these proteins contain SANT domains (SWI3, ADA2, N-CoR and TFIIIB DNA-binding domains) with nonspecific DNA-binding activity and are polycomb-group (PcG) proteins responsible

for H3K27me3. Class II SET proteins include four ASSH proteins (ASSH1/SDG26, ASSH2/SDG8, ASSH3/SDG7 and ASSH4/SDG24) and three ASH1-RELATED proteins (ASHR1/SDG37, ASHR2/SDG39 and ASHR3/SDG4) and are associated with mono-, di- and trimethylation of H3K36 (Xu et al. 2008; Valencia-Morales Mdel et al. 2012). In addition to the conserved ASSOCIATED WITH SET (AWS) and SET domains, class II SET proteins contain several other domains, including PWWP (domain containing Pro-Trp-Trp-Pro motif), PHD (plant homeodomain), bromo and BAH domains. Class III SET proteins encode TRX orthologs with conserved PWWP, PHD and FYP/DAST domains. The five class III SET proteins in *Arabidopsis* are ATX1/SDG27, ATX2/SDG30, ATX3/SDG14, ATX4/SDG16 and ATX5/SDG29. ATX1 and ATX2 are required for H3K4 methylation (Zhang et al. 2009), while the functions of the other three ATX proteins have not yet been reported. Class IV SET domain proteins are ATX-related (ATXR) proteins with PHD and SET domains, and the seven *Arabidopsis* ATXR proteins exhibit methyltransferase activity on different lysines or have unknown functions. ATXR3/SDG2 is the major H3K4me3 methyltransferase (Berr et al. 2010; Guo et al. 2010); ATXR5/SDG15 and ATXR6/SDG34 function redundantly in controlling the level of H3K27me1 (Jacob et al. 2009); and ATXR7/SDG25 is required for H3K4 methylation at the *FLC* locus (Tamada et al. 2009). The functions of ATXR1/SDG35, ATXR2/SDG36 and ATXR4/SDG38 have not been characterized.

Arabidopsis class V SET proteins, representing the largest SET protein family, encode Su(var)3-9 orthologs and include ten SUVH and five SUVH-RELATED (SUVR) proteins. The SUVH proteins contain SRA, pre-SET, SET and post-SET domains and can

be divided into the four following clades: SUVH1, SUVH2, SUVH4 and SUVH5 (Naumann et al. 2005). SUVH4/SDG33, SUVH5/SDG9 and SUVH6/ SDG23, which belong to the SUVH4 and SUVH5 clades, are the major active H3K9me2 methyltransferases (Ebbs and Bender 2006; Stroud et al. 2014). The SUVH2 subgroup members SUVH2/SDG3 and SUVH9/SDG22 are RdDM factors required for Pol V occupancy at DNA-methylated regions (Johnson et al. 2014; Liu et al. 2014). No functions have been reported for any of the SUVH1 proteins, which include SUVH1/SDG23, SUVH3/SDG19, SUVH7/SDG17, SUVH8/SDG21 and SUVH10/SDG11 (Naumann et al. 2005). The five SUVR proteins lack SRA domains. SUVR5/SDG6 contains a zinc figure/C2H2 domain, and SUVR1/SDG13, SUVR2/SDG18 and SUVR4/SDG31 all contain a conserved WIYLD domain (Thorstensen et al. 2006; Caro et al. 2012). SUVR4 binds free ubiquitin through its WIYLD domain and converts H3K9me1 to H3K9me3 at transposons and pseudogenes (Veiseth et al. 2011). SUVR5 is reported to mediate H3K9me2 deposition independently of DNA methylation (Caro et al. 2012). Recently, SUVR2/SDG18 was found to participate in RdDM process by association with SNF2 chromatin remodeler through a forward-genetic screen (Han et al. 2014) and mass-spectrometry analysis of the immoprecipitation product of SUVR2/SDG18 (Groth et al. 2014). SUVR1/SDG13 is also in the same complex with SUVR2/SDG18 and plays non-redundant roles in gene silencing (Han et al. 2014). In contrast, no histone methyltransferase activities have been reported for SUVR3/SDG20.

The interplay among different epigenetic modifications

DNA, which carries genetic information, is one component of chromatin along with histones and non-histone proteins; thus, transcriptional epigenetic regulation is accomplished through the combined effects of DNA and histone modifications. A number of recent studies have shed light on this complex interaction, which encompasses the following: the effects of DNA methylation and histone modifications, the interplay and crosstalk among different modifications and the effect of other factors on DNA and histone modification.

Genome-wide data from *Arabidopsis* have shown that the levels of CG, CHG and CHH DNA methylation are 24%, 6.7% and 1.7%, respectively (Cokus et al. 2008). Comprehensive methylome profiling of *Arabidopsis* mutants with silencing defects have helped uncover the interdependence of these three types of DNA methylation (Stroud et al. 2013). In *met1*, the loss of CG methylation is accompanied by large decreases in CHG and CHH methylation. Additionally, MET1-dependent CHG loci largely overlap with CMT3- and SUVH4/5/6-dependent CHG loci. One potential mechanism is that SUVH4/5/6 are recruited to chromatin loci for histone methylation through their SRA domains and binding to methylated cytosines. Because CG methylation is the most highly methylated DNA methylation type, the loss of CG methylation may affect SUVH4/5/6-directed H3K9me2 methylation, consistent with the loss of H3K9me2 in *met1* (Deleris et al. 2012). Consequently, H3K9me2-dependent CMT3 recruitment required for C methylation in the CHG context would also be affected (Du et al. 2012). MET1-dependent CHH methylation loci largely overlap with DRM1/2-dependent CHH loci

rather than overlapping with CMT3- and SUVH4/5/6-dependent CHH loci, which suggests that MET1 regulates CHH through a different pathway. The recent findings that methyl-DNA binding proteins (SUVH2 and SUVH9) participate in the RdDM pathway mediated by DRM2 raise the possibility that MET1 affects CHH methylation at RdDM loci by affecting SUVH2 and SUVH9 function. In *cmt3* or *suvh4/5/6*, the CHG methylation level is largely decreased, however, the CHH methylation level at loci where the CHH methylation is maintained by DRM2 is only modestly reduced while that at loci where the CHH methylation is maintained by CMT2 is greatly reduced. The lower percentage of CHG methylation among the three sequence contexts may help explain the reduced impact on CHH methylation observed at DRM2-maintained loci. There are two possible explanations for the decreased CHH methylation observed at CMT2-dependent CHH loci. The first possibility concerns the redundant roles played by CMT2 and CMT3 in CHG methylation. When CMT3 is absent, increased/compensatory CMT2 function at CHG loci may compromise the role of CMT2 at CHH loci. On the other hand, CMT2 functions through the chromatin-remodeling protein DDM1 and linker histone H1 (Zemach et al. 2013), which suggests that chromatin structure is important for CMT2 function. In *cmt3* and *suvh4/5/6,* CHG methylation and H3K9me2 are strongly affected in heterochromatic regions, which may lead to chromatin remodeling and thus affect CHH methylation. Reflecting their relative levels in the genome, CG methylation (high) may affect CHG and CHH methylation (moderate and low, respectively), and CHG methylation may affect CHH methylation. Consistently, the impact in the reverse

direction (i.e., the impact of CHG methylation on CG methylation and the impact of CHH methylation on either CH or CHG methylation) is trivial (Stroud et al. 2013).

As previously described, CHH methylation requires the *de novo* methylation pathway and guidance signals, such as P4siRNAs, to methylate specific loci. Although P4siRNAs are virtually eliminated when Pol IV function is compromised, CHH methylation is affected at DRM2-targeted loci but not at CMT2-targeted loci (Chapter 2). The fact the DRM2-targeted loci are primarily located in the euchromatic arms where epigenetic marks are rare probably underlies the indispensable roles of P4siRNAs in directing RdDM at these loci (Chapter 2). The concentration of CMT2-targeted loci at heterochromatic regions and the dependency on DDM1 and H1 suggest that higher order chromatin structures may function as guidance signals at these loci.

The impact of DNA methylation on chromatin structures has been well established through numerous studies (Keshet et al. 1986; Ballestar and Esteller 2002; Martinowich et al. 2003; Gilbert et al. 2007), but the underlying mechanisms are not well understood. At present, there are two conserved domains known to recognize DNA methylation: the SET and RING-associated (SRA) domain (Rajakumara et al. 2011) and the METHYL-CpG-BINDING domain (MBD) (Fournier et al. 2012). The investigation of the latter began with the discovery of MeCP2 (Lewis et al. 1992), the first protein found to bind methylated cytosines. MBD proteins are usually associated with other chromatin-associated domains (e.g., the bromo, SET and PHD finger domains) that promote histone deacetylase and histone methyltransferase activity (Jones et al. 1998; Nan et al. 1998; Zhang et al. 1999; Fuks et al. 2003a; Bogdanovic and Veenstra 2009). For example, MeCP2 may function as a bridge between DNA methylation and H3K9

methylation (Fuks et al. 2003b), and MBD1 interacts with the SUVH39H1 and HP1 heterochromatin complex to achieve transcriptional silencing (Fujita et al. 2003). Of the 13 MBD-containing proteins in *Arabidopsis*, only AtMBD5, AtMBD6 and AtMBD7 have been confirmed to bind methyl CpG and to colocalize in the highly methylated chromocenters (Zemach and Grafi 2007). The underlying mechanism of MBD in *Arabidopsis* is unclear, but interactions with DDM1 and HDAC have been detected (Zemach et al. 2005; Zemach and Grafi 2007).

SRA domain-containing proteins also play a role in connecting DNA methylation with other epigenetic marks. In mammals, ubiquitin-like with PHD and RING finger domains 1 (UHRF1) binds methylated CpG through the RING-associated SRA domain to recruit DNA methyltransferase DNMT1 to maintain DNA methylation during DNA replication (Rajakumara et al. 2011). In *Arabidopsis*, the VARIANT IN METHYLATION (VIM)/ORTHRUS (ORTH) family includes homologs of UHRF1 that play similar roles (Kim et al. 2014). SUV39H1 and its homologs, with both SRA and SET domains, can methylate H3K9 by binding methylated cytosines. (The *Arabidopsis* SUVH protein family was introduced in detail in the section describing the SET domain proteins.) Ultimately, all of these DNA methylation-associated proteins link DNA methylation to repressive chromatin marks (namely, DNA methylation, H3K9me2 and histone deacetylation).

The diversity of possible histone modifications hints at the complexity of the crosstalk among different modifications. The occurrence of different modifications on the same amino acid, e.g., acetylation, methylation and ubiquitynation on lysine, raises the possibility of competitive antagonism. Additionally, one type of modification may stimulate another. For example, findings

in yeast indicate that H2B monoubiquitination by Rad6/Bre1 can regulate H3K4 methylation by COMPASS and H3K79 methylation by Dot1 through the control of Cps35, which is required for the activity of the COMPASS complex and proper H3K79 methylation (Lee et al. 2007). In HeLa cells, H2K34 ubiquitynation mediated by the RING finger protein MSL2 in the MOF complex is important for global H3K4me3 and H3K79me2 through trans-tail crosstalk (Wu et al. 2011). In an opposing manner, a given modification may be abolished by another modification. For example, the chromodomain of Eaf3, a subunit in the active deacetylase complex, recognizes Set2-methylated histone H3K36 and initiates Rpd3 deacetylase activity (Lee and Shilatifard 2007).

Histone modifications may also affect other epigenetic marks such as DNA methylation. As described above, the *de novo* methyltransferase DNMT3 requires the inactive paralog DNMT3L to stimulate its enzymatic activity. The DNA-binding affinity of DNMT3A is blocked by the interaction of the ATRX–DNMT3–DNMT3L (ADD) domain with the catalytic domain (CD), while the binding of H3K4me0 may disrupt the ADD-CD interaction and stimulate the enzymatic activity of DNMT3 (Guo et al. 2014). In *Arabidopsis*, a mild reduction of RdDM has been observed in several H3K4 demethylase mutants, including *jumonji 14* (*jmj14*), *lysine-specific demethylase 1-like 1* (*ldl1*) and *ldl2* (Greenberg et al. 2013).

Perspective

Regulatory mechanisms are critical for the developmental processes of organisms and their ability to respond to environmental stimuli. Each regulatory factor must be coordinated to turn genes on or off at the proper time and location. At the epigenetic level,

DNA and histone modifications and the resulting chromatin structure changes are fundamental regulatory mechanisms. The well-studied regulation of *FLOWERING LOCUS C* (*FLC*) is a good example. In the early embryo, *FLC* is expressed at high levels due to the function of several conserved complexes and modifications, including the RNA polymerase-associated factor 1 complex (Paf1C); H2B ubiquitination; H3K4me2/me3 through ATX1, ATX2 and ATXR7; H3K36 methylation through EFS/SDG8; and the chromatin-remodeling complex WR1/SRCAP and FRIDIDA (FRI) with coiled-coil domains (Crevillen and Dean 2011; Song et al. 2013). During vernalization, *FLC* transitions from active expression to a silenced state through a series of steps. After two weeks in the cold, *FLC* transcription is greatly reduced, the *FLC* gene loop is disrupted, and the expression of an *FLC* antisense transcript called COOLAIR is increased. After three weeks, expression of another long non-coding RNA, named COLDAIR, helps recruit the histone methyltransferase subunit of Polycomb repressive complex 2 (PRC2) to deposit H3K27me3 at the *FLC* locus (Song et al. 2013). The cold stimulus also induces a conserved interaction between PRC2 and PHD-containing proteins, including the constitutively expressed VERNALIZATION5 (VRN5/VIL1) and VERNALIZATION5/VIN3-LIKE 1 (VEL1) proteins and the cold-induced VERNALIZATION INSENSITIVE 3 (VIN3) protein (Song et al. 2013). When the plants are returned to warm conditions, the interaction between PRC2 and PHD-containing proteins spreads throughout the entire *FLC* locus, and this is accompanied by an increase in H3K27me3. The maintenance of *FLC* silencing also requires the binding of LIKE HETEROCHROMATIN PROTEIN1 (LHP1) to H3K27me3 through its chromodomain

and other factors such as VERNALIZATION1 (VRN1), which has two plant-specific B3 domains (Song et al. 2013). The regulation of *FLC* demonstrates how the precise control of gene regulation may involve numerous proteins, epigenetic modifications and DNA sequence elements. Thus, our expanding knowledge of epigenetic modifications helps improve our understanding of the complexity and scope of gene regulatory networks.

References

Allfrey VG, Faulkner R, Mirsky AE. 1964. Acetylation and Methylation of Histones and Their Possible Role in the Regulation of Rna Synthesis. *Proceedings of the National Academy of Sciences of the United States of America* **51**: 786-794.

Arita K, Ariyoshi M, Tochio H, Nakamura Y, Shirakawa M. 2008. Recognition of hemi-methylated DNA by the SRA protein UHRF1 by a base-flipping mechanism. *Nature* **455**(7214): 818-U812.

Ausin I, Greenberg MV, Li CF, Jacobsen SE. 2012a. The splicing factor SR45 affects the RNA-directed DNA methylation pathway in Arabidopsis. *Epigenetics* **7**(1): 29-33.

Ausin I, Greenberg MVC, Simanshu DK, Hale CJ, Vashisht AA, Simon SA, Lee TF, Feng SH, Espanola SD, Meyers BC et al. 2012b. INVOLVED IN DE NOVO 2-containing complex involved in RNA-directed DNA methylation in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America* **109**(22): 8374-8381.

Ausin I, Mockler TC, Chory J, Jacobsen SE. 2009. IDN1 and IDN2 are required for de novo DNA methylation in Arabidopsis thaliana. *Nature structural & molecular biology* **16**(12): 1325-1327.

Ballestar E, Esteller M. 2002. The impact of chromatin in human cancer: linking DNA methylation to gene silencing. *Carcinogenesis* **23**(7): 1103-1109.

Baniushin BF. 2005. [Methylation of adenine residues in DNA of eukaryotes]. *Molekuliarnaia biologiia* **39**(4): 557-566.

Bannister AJ, Kouzarides T. 2011. Regulation of chromatin by histone modifications. *Cell research* **21**(3): 381-395.

Berger SL. 2001. An embarrassment of niches: the many covalent modifications of histones in transcriptional regulation. *Oncogene* **20**(24): 3007-3013.

Berger SL, Kouzarides T, Shiekhattar R, Shilatifard A. 2009. An operational definition of epigenetics. *Genes & development* **23**(7): 781-783.

Berr A, McCallum EJ, Menard R, Meyer D, Fuchs J, Dong A, Shen WH. 2010. Arabidopsis SET DOMAIN GROUP2 is required for H3K4 trimethylation and is crucial for both sporophyte and gametophyte development. *The Plant cell* **22**(10): 3232-3248.

Bhutani N, Brady JJ, Damian M, Sacco A, Corbel SY, Blau HM. 2010. Reprogramming towards pluripotency requires AID-dependent DNA demethylation. *Nature* **463**(7284): 1042-U1057.

Bies-Etheve N, Pontier D, Lahmy S, Picart C, Vega D, Cooke R, Lagrange T. 2009. RNA-directed DNA methylation requires an AGO4-interacting member of the SPT5 elongation factor family. *EMBO reports* **10**(6): 649-654.

Bogdanovic O, Veenstra GJ. 2009. DNA methylation and methyl-CpG binding proteins: developmental requirements and function. *Chromosoma* **118**(5): 549-565.

Bonasio R, Tu S, Reinberg D. 2010. Molecular signals of epigenetic states. *Science* **330**(6004): 612-616.

Bostick M, Kim JK, Esteve PO, Clark A, Pradhan S, Jacobsen SE. 2007. UHRF1 plays a role in maintaining DNA methylation in mammalian cells. *Science* **317**(5845): 1760-1764.

Bourc'his D, Voinnet O. 2010. A small-RNA perspective on gametogenesis, fertilization, and early zygotic development. *Science* **330**(6004): 617-622.

Brabbs TR, He Z, Hogg K, Kamenski A, Li Y, Paszkiewicz KH, Moore KA, O'Toole P, Graham IA, Jones L. 2013. The stochastic silencing phenotype of Arabidopsis morc6 mutants reveals a role in efficient RNA-directed DNA methylation. *The Plant journal : for cell and molecular biology* **75**(5): 836-846.

Cao X, Jacobsen SE. 2002. Locus-specific control of asymmetric and CpNpG methylation by the DRM and CMT3 methyltransferase genes. *Proceedings of the National Academy of Sciences of the United States of America* **99 Suppl 4**: 16491-16498.

Capuano F, Mulleder M, Kok R, Blom HJ, Ralser M. 2014. Cytosine DNA Methylation Is Found in Drosophila melanogaster but Absent in Saccharomyces cerevisiae, Schizosaccharomyces pombe, and Other Yeast Species. *Anal Chem* **86**(8): 3697-3702.

Caro E, Stroud H, Greenberg MV, Bernatavichute YV, Feng S, Groth M, Vashisht AA, Wohlschlegel J, Jacobsen SE. 2012. The SET-domain protein SUVR5 mediates H3K9me2 deposition and silencing at stimulus response genes in a DNA methylation-independent manner. *PLoS genetics* **8**(10): e1002995.

Cedar H, Bergman Y. 2009. Linking DNA methylation and histone modification: patterns and paradigms. *Nature reviews Genetics* **10**(5): 295-304.

Chen XM. 2009. Small RNAs and Their Roles in Plant Development. *Annual review of cell and developmental biology* **25**: 21-44.

Cho SH, Addo-Quaye C, Coruh C, Arif MA, Ma Z, Frank W, Axtell MJ. 2008. Physcomitrella patens DCL3 is required for 22-24 nt siRNA accumulation, suppression of retrotransposon-derived transcripts, and normal development. *PLoS genetics* **4**(12): e1000314.

Choi YH, Gehring M, Johnson L, Hannon M, Harada JJ, Goldberg RB, Jacobsen SE, Fischer RL. 2002. DEMETER, a DNA glycosylase domain protein, is required for endosperm gene imprinting and seed viability in Arabidopsis. *Cell* **110**(1): 33-42.

Chuang LSH, Ian HI, Koh TW, Ng HH, Xu GL, Li BFL. 1997. Human DNA (cytosine-5) methyltransferase PCNA complex as a target for p21(WAF1). *Science* **277**(5334): 1996-2000.

Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE. 2008. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* **452**(7184): 215-219.

Coleman-Derr D, Zilberman D. 2012. Deposition of histone variant H2A.Z within gene bodies regulates responsive genes. *PLoS genetics* **8**(10): e1002988.

Cortellino S, Xu JF, Sannai M, Moore R, Caretti E, Cigliano A, Le Coz M, Devarajan K, Wessels A, Soprano D et al. 2011. Thymine DNA Glycosylase Is Essential for Active DNA Demethylation by Linked Deamination-Base Excision Repair. *Cell* **146**(1): 67-79.

Crevillen P, Dean C. 2011. Regulation of the floral repressor gene FLC: the complexity of transcription in a chromatin context. *Current opinion in plant biology* **14**(1): 38-44.

Deleris A, Stroud H, Bernatavichute Y, Johnson E, Klein G, Schubert D, Jacobsen SE. 2012. Loss of the DNA methyltransferase MET1 Induces H3K9 hypermethylation at PcG target genes and redistribution of H3K27 trimethylation to transposons in Arabidopsis thaliana. *PLoS genetics* **8**(11): e1003062.

Demarre G, Chattoraj DK. 2010. DNA adenine methylation is required to replicate both Vibrio cholerae chromosomes once per cell cycle. *PLoS genetics* **6**(5): e1000939.

Dominissini D, Moshitch-Moshkovitz S, Schwartz S, Salmon-Divon M, Ungar L, Osenberg S, Cesarkas K, Jacob-Hirsch J, Amariglio N, Kupiec M et al. 2012.

Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature* **485**(7397): 201-206.

Du J, Zhong X, Bernatavichute YV, Stroud H, Feng S, Caro E, Vashisht AA, Terragni J, Chin HG, Tu A et al. 2012. Dual binding of chromomethylase domains to H3K9me2-containing nucleosomes directs DNA methylation in plants. *Cell* **151**(1): 167-180.

Duran-Figueroa N, Vielle-Calzada JP. 2010. ARGONAUTE9-dependent silencing of transposable elements in pericentromeric regions of Arabidopsis. *Plant signaling & behavior* **5**(11): 1476-1479.

Ebbs ML, Bender J. 2006. Locus-specific control of DNA methylation by the Arabidopsis SUVH5 histone methyltransferase. *The Plant cell* **18**(5): 1166-1176.

El-Shami M, Pontier D, Lahmy S, Braun L, Picart C, Vega D, Hakimi MA, Jacobsen SE, Cooke R, Lagrange T. 2007. Reiterated WG/GW motifs form functionally and evolutionarily conserved ARGONAUTE-binding platforms in RNAi-related components. *Genes & development* **21**(20): 2539-2544.

Fan L, Roberts VA. 2006. Complex of linker histone H5 with the nucleosome and its implications for chromatin packing. *Proceedings of the National Academy of Sciences of the United States of America* **103**(22): 8384-8389.

Fang G, Munera D, Friedman DI, Mandlik A, Chao MC, Banerjee O, Feng ZX, Losic B, Mahajan MC, Jabado OJ et al. 2012. Genome-wide mapping of methylated adenine residues in pathogenic Escherichia coli using single-molecule real-time sequencing. *Nat Biotechnol* **30**(12): 1232-+.

Feng S, Jacobsen SE, Reik W. 2010. Epigenetic reprogramming in plant and animal development. *Science* **330**(6004): 622-627.

Finke A, Kuhlmann M, Mette MF. 2012. IDN2 has a role downstream of siRNA formation in RNA-directed DNA methylation. *Epigenetics* **7**(8): 950-960.

Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC. 1998. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature* **391**(6669): 806-811.

Flanagan JF, Mi LZ, Chruszcz M, Cymborowski M, Clines KL, Kim YC, Minor W, Rastinejad F, Khorasanizadeh S. 2005. Double chromodomains cooperate to recognize the methylated histone H3 tail. *Nature* **438**(7071): 1181-1185.

Fournier A, Sasai N, Nakao M, Defossez PA. 2012. The role of methyl-binding proteins in chromatin organization and epigenome maintenance. *Briefings in functional genomics* **11**(3): 251-264.

Fu Y, Dominissini D, Rechavi G, He C. 2014. Gene expression regulation mediated through reversible m(6)A RNA methylation. *Nature reviews Genetics* **15**(5): 293-306.

Fujita N, Watanabe S, Ichimura T, Tsuruzoe S, Shinkai Y, Tachibana M, Chiba T, Nakao M. 2003. Methyl-CpG binding domain 1 (MBD1) interacts with the Suv39h1-HP1 heterochromatic complex for DNA methylation-based transcriptional repression. *The Journal of biological chemistry* **278**(26): 24132-24138.

Fuks F, Hurd PJ, Wolf D, Nan X, Bird AP, Kouzarides T. 2003a. The methyl-CpG-binding protein MeCP2 links DNA methylation to histone methylation. *The Journal of biological chemistry* **278**(6): 4035-4040.

Fuks F, Hurd PJ, Wolf D, Nan X, Bird AP, Kouzarides T. 2003b. The Methyl-CpG-binding Protein MeCP2 Links DNA Methylation to Histone Methylation. *Journal of Biological Chemistry* **278**(6): 4035-4040.

Fukushige S, Horii A. 2013. DNA Methylation in Cancer: A Gene Silencing Mechanism and the Clinical Potential of Its Biomarkers. *Tohoku J Exp Med* **229**(3): 173-185.

Garcia-Ruiz H, Takeda A, Chapman EJ, Sullivan CM, Fahlgren N, Brempelis KJ, Carrington JC. 2010. Arabidopsis RNA-dependent RNA polymerases and dicer-like proteins in antiviral defense and small interfering RNA biogenesis during Turnip Mosaic Virus infection. *The Plant cell* **22**(2): 481-496.

Gilbert N, Thomson I, Boyle S, Allan J, Ramsahoye B, Bickmore WA. 2007. DNA methylation affects nuclear organization, histone modifications, and linker histone binding but not chromatin compaction. *The Journal of cell biology* **177**(3): 401-411.

Gong Z, Morales-Ruiz T, Ariza RR, Roldan-Arjona T, David L, Zhu JK. 2002. ROS1, a repressor of transcriptional gene silencing in Arabidopsis, encodes a DNA glycosylase/lyase. *Cell* **111**(6): 803-814.

Gong Z, Zhu JK. 2011. Active DNA demethylation by oxidation and repair. *Cell research* **21**(12): 1649-1651.

Greenberg MV, Deleris A, Hale CJ, Liu A, Feng S, Jacobsen SE. 2013. Interplay between active chromatin marks and RNA-directed DNA methylation in Arabidopsis thaliana. *PLoS genetics* **9**(11): e1003946.

Groth M, Stroud H, Feng S, Greenberg MV, Vashisht AA, Wohlschlegel JA, Jacobsen SE, Ausin I. 2014. SNF2 chromatin remodeler-family proteins FRG1 and -2 are required for RNA-directed DNA methylation. *Proceedings of the National Academy of Sciences of the United States of America*.

Guo L, Yu YC, Law JA, Zhang XY. 2010. SET DOMAIN GROUP2 is the major histone H3 lysie 4 trimethyltransferase in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America* **107**(43): 18557-18562.

Guo X, Wang L, Li J, Ding Z, Xiao J, Yin X, He S, Shi P, Dong L, Li G et al. 2014. Structural insight into autoinhibition and histone H3-induced activation of DNMT3A. *Nature*.

Haag JR, Ream TS, Marasco M, Nicora CD, Norbeck AD, Pasa-Tolic L, Pikaard CS. 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. *Molecular cell* **48**(5): 811-818.

Halfmann R, Lindquist S. 2010. Epigenetics in the extreme: prions and the inheritance of environmentally acquired traits. *Science* **330**(6004): 629-632.

Han YF, Dou K, Ma ZY, Zhang SW, Huang HW, Li L, Cai T, Chen S, Zhu JK, He XJ. 2014. SUVR2 is involved in transcriptional gene silencing by associating with SNF2-related chromatin-remodeling proteins in Arabidopsis. *Cell research* **24**(12): 1445-1465.

Hata K, Okano M, Lei H, Li E. 2002. Dnmt3L cooperates with the Dnmt3 family of de novo DNA methyltransferases to establish maternal imprints in mice. *Development* **129**(8): 1983-1993.

He XJ, Hsu YF, Zhu S, Wierzbicki AT, Pontes O, Pikaard CS, Liu HL, Wang CS, Jin H, Zhu JK. 2009a. An effector of RNA-directed DNA methylation in arabidopsis is an ARGONAUTE 4- and RNA-binding protein. *Cell* **137**(3): 498-508.

He XJ, Hsu YF, Zhu SH, Liu HL, Pontes O, Zhu JH, Cui XP, Wang CS, Zhu JK. 2009b. A conserved transcriptional regulator is required for RNA-directed DNA methylation and plant development. *Genes & development* **23**(23): 2717-2722.

Henderson IR, Deleris A, Wong W, Zhong X, Chin HG, Horwitz GA, Kelly KA, Pradhan S, Jacobsen SE. 2010. The de novo cytosine methyltransferase DRM2 requires intact UBA domains and a catalytically mutated paralog DRM3 during RNA-directed DNA methylation in Arabidopsis thaliana. *PLoS genetics* **6**(10): e1001182.

Henderson IR, Zhang X, Lu C, Johnson L, Meyers BC, Green PJ, Jacobsen SE. 2006. Dissecting Arabidopsis thaliana DICER function in small RNA processing, gene silencing and DNA methylation patterning. *Nature genetics* **38**(6): 721-725.

Henikoff S, Comai L. 1998. A DNA methyltransferase homolog with a chromodomain exists in multiple polymorphic forms in Arabidopsis. *Genetics* **149**(1): 307-318.

Hirochika H, Okamoto H, Kakutani T. 2000. Silencing of retrotransposons in arabidopsis and reactivation by the ddm1 mutation. *The Plant cell* **12**(3): 357-368.

Horvath S. 2013. DNA methylation age of human tissues and cell types. *Genome biology* **14**(10): R115.

Huang CF, Miki D, Tang K, Zhou HR, Zheng Z, Chen W, Ma ZY, Yang L, Zhang H, Liu R et al. 2013. A Pre-mRNA-splicing factor is required for RNA-directed DNA methylation in Arabidopsis. *PLoS genetics* **9**(9): e1003779.

Huang CF, Zhu JK. 2014. RNA Splicing Factors and RNA-Directed DNA Methylation. *Biology* **3**(2): 243-254.

Iki T, Yoshikawa M, Nishikiori M, Jaudal MC, Matsumoto-Yokoyama E, Mitsuhara I, Meshi T, Ishikawa M. 2010. In vitro assembly of plant RNA-induced silencing complexes facilitated by molecular chaperone HSP90. *Molecular cell* **39**(2): 282-291.

Ito S, Shen L, Dai Q, Wu SC, Collins LB, Swenberg JA, He C, Zhang Y. 2011. Tet Proteins Can Convert 5-Methylcytosine to 5-Formylcytosine and 5-Carboxylcytosine. *Science* **333**(6047): 1300-1303.

Jackson JP, Lindroth AM, Cao X, Jacobsen SE. 2002. Control of CpNpG DNA methylation by the KRYPTONITE histone H3 methyltransferase. *Nature* **416**(6880): 556-560.

Jacob Y, Feng S, LeBlanc CA, Bernatavichute YV, Stroud H, Cokus S, Johnson LM, Pellegrini M, Jacobsen SE, Michaels SD. 2009. ATXR5 and ATXR6 are H3K27 monomethyltransferases required for chromatin structure and gene silencing. *Nature structural & molecular biology* **16**(7): 763-768.

Jenuwein T, Allis CD. 2001. Translating the histone code. *Science* **293**(5532): 1074-1080.

Jia Y, Lisch DR, Ohtsu K, Scanlon MJ, Nettleton D, Schnable PS. 2009. Loss of RNA-dependent RNA polymerase 2 (RDR2) function causes widespread and

unexpected changes in the expression of transposons, genes, and 24-nt small RNAs. *PLoS genetics* **5**(11): e1000737.

Jilani A, Ramotar D, Slack C, Ong C, Yang XM, Scherer SW, Lasko DD. 1999. Molecular cloning of the human gene, PNKP, encoding a polynucleotide kinase 3'-phosphatase and evidence for its role in repair of DNA strand breaks caused by oxidative damage. *The Journal of biological chemistry* **274**(34): 24176-24186.

Johnson LM, Bostick M, Zhang X, Kraft E, Henderson I, Callis J, Jacobsen SE. 2007. The SRA methyl-cytosine-binding domain links DNA and histone methylation. *Current biology : CB* **17**(4): 379-384.

Johnson LM, Du J, Hale CJ, Bischof S, Feng S, Chodavarapu RK, Zhong X, Marson G, Pellegrini M, Segal DJ et al. 2014. SRA- and SET-domain-containing proteins link RNA polymerase V occupancy to DNA methylation. *Nature* **507**(7490): 124-128.

Jones PL, Jan Veenstra GC, Wade PA, Vermaak D, Kass SU, Landsberger N, Strouboulis J, Wolffe AP. 1998. Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. *Nat Genet* **19**(2): 187-191.

Jones RS, Gelbart WM. 1993. The Drosophila Polycomb-Group Gene Enhancer of Zeste Contains a Region with Sequence Similarity to Trithorax. *Molecular and Cellular Biology* **13**(10): 6357-6366.

Kahramanoglou C, Prieto AI, Khedkar S, Haase B, Gupta A, Benes V, Fraser GM, Luscombe NM, Seshasayee ASN. 2012. Genomics of DNA cytosine methylation in Escherichia coli reveals its role in stationary phase transcription. *Nat Commun* **3**.

Kamakaka RT, Biggins S. 2005. Histone variants: deviants? *Genes & development* **19**(3): 295-310.

Kanno T, Bucher E, Daxinger L, Huettel B, Kreil DP, Breinig F, Lind M, Schmitt MJ, Simon SA, Gurazada SGR et al. 2010. RNA-directed DNA methylation and plant development require an IWR1-type transcription factor. *EMBO reports* **11**(1): 65-71.

Kasschau KD, Fahlgren N, Chapman EJ, Sullivan CM, Cumbie JS, Givan SA, Carrington JC. 2007. Genome-wide profiling and analysis of Arabidopsis siRNAs. *PLoS biology* **5**(3): e57.

Keshet I, Liemanhurwitz J, Cedar H. 1986. DNA Methylation Affects the Formation of Active Chromatin. *Cell* **44**(4): 535-543.

Kim J, Kim JH, Richards EJ, Chung KM, Woo HR. 2014. Arabidopsis VIM proteins regulate epigenetic silencing by modulating DNA methylation and histone modification in cooperation with MET1. *Molecular plant* **7**(9): 1470-1485.

Kohli RM, Zhang Y. 2013. TET enzymes, TDG and the dynamics of DNA demethylation. *Nature* **502**(7472): 472-479.

Kornberg RD. 1977. Structure of chromatin. *Annual review of biochemistry* **46**: 931-954.

Law JA, Ausin I, Johnson LM, Vashisht AA, Zhu JK, Wohlschlegel JA, Jacobsen SE. 2010. A protein complex required for polymerase V transcripts and RNA-directed DNA methylation in Arabidopsis. *Current biology : CB* **20**(10): 951-956.

Law JA, Du J, Hale CJ, Feng S, Krajewski K, Palanca AM, Strahl BD, Patel DJ, Jacobsen SE. 2013. Polymerase IV occupancy at RNA-directed DNA methylation sites requires SHH1. *Nature* **498**(7454): 385-389.

Law JA, Jacobsen SE. 2010. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature reviews Genetics* **11**(3): 204-220.

Law JA, Vashisht AA, Wohlschlegel JA, Jacobsen SE. 2011. SHH1, a homeodomain protein required for DNA methylation, as well as RDR2, RDM4, and chromatin remodeling factors, associate with RNA polymerase IV. *PLoS genetics* **7**(7): e1002195.

Lee JS, Shilatifard A. 2007. A site to remember: H3K36 methylation a mark for histone deacetylation. *Mutation research* **618**(1-2): 130-134.

Lee JS, Shukla A, Schneider J, Swanson SK, Washburn MP, Florens L, Bhaumik SR, Shilatifard A. 2007. Histone crosstalk between H2B monoubiquitination and H3 methylation mediated by COMPASS. *Cell* **131**(6): 1084-1096.

Lewis JD, Meehan RR, Henzel WJ, Maurerfogy I, Jeppesen P, Klein F, Bird A. 1992. Purification, Sequence, and Cellular-Localization of a Novel Chromosomal Protein That Binds to Methylated DNA. *Cell* **69**(6): 905-914.

Li CF, Pontes O, El-Shami M, Henderson IR, Bernatavichute YV, Chan SW, Lagrange T, Pikaard CS, Jacobsen SE. 2006a. An ARGONAUTE4-containing nuclear processing center colocalized with Cajal bodies in Arabidopsis thaliana. *Cell* **126**(1): 93-106.

Li H, Ilin S, Wang W, Duncan EM, Wysocka J, Allis CD, Patel DJ. 2006b. Molecular basis for site-specific read-out of histone H3K4me3 by the BPTF PHD finger of NURF. *Nature* **442**(7098): 91-95.

Li S, Liu L, Zhuang X, Yu Y, Liu X, Cui X, Ji L, Pan Z, Cao X, Mo B et al. 2013. MicroRNAs inhibit the translation of target mRNAs on the endoplasmic reticulum in Arabidopsis. *Cell* **153**(3): 562-574.

Li X, Qian W, Zhao Y, Wang C, Shen J, Zhu JK, Gong Z. 2012. Antisilencing role of the RNA-directed DNA methylation pathway and a histone acetyltransferase in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America* **109**(28): 11425-11430.

Lindroth AM, Cao X, Jackson JP, Zilberman D, McCallum CM, Henikoff S, Jacobsen SE. 2001. Requirement of CHROMOMETHYLASE3 for maintenance of CpXpG methylation. *Science* **292**(5524): 2077-2080.

Lister R, Mukamel EA, Nery JR, Urich M, Puddifoot CA, Johnson ND, Lucero J, Huang Y, Dwork AJ, Schultz MD et al. 2013. Global Epigenomic Reconfiguration During Mammalian Brain Development. *Science* **341**(6146): 629-+.

Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR. 2008. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* **133**(3): 523-536.

Liu B, Chen Z, Song X, Liu C, Cui X, Zhao X, Fang J, Xu W, Zhang H, Wang X et al. 2007. Oryza sativa dicer-like4 reveals a key role for small interfering RNA silencing in plant development. *The Plant cell* **19**(9): 2705-2718.

Liu J, Bai G, Zhang CJ, Chen W, Zhou JX, Zhang SW, Chen Q, Deng X, He XJ, Zhu JK. 2011. An atypical component of RNA-directed DNA methylation machinery has both DNA methylation-dependent and -independent roles in locus-specific transcriptional gene silencing. *Cell research* **21**(12): 1691-1700.

Liu Q, Feng Y, Zhu Z. 2009. Dicer-like (DCL) proteins in plants. *Funct Integr Genomics* **9**(3): 277-286.

Liu ZW, Shao CR, Zhang CJ, Zhou JX, Zhang SW, Li L, Chen S, Huang HW, Cai T, He XJ. 2014. The SET domain proteins SUVH2 and SUVH9 are required for Pol V occupancy at RNA-directed DNA methylation loci. *PLoS genetics* **10**(1): e1003948.

Low DA, Weyand NJ, Mahan MJ. 2001. Roles of DNA adenine methylation in regulating bacterial gene expression and virulence. *Infect Immun* **69**(12): 7197-7204.

Lyko F, Ramsahoye BH, Jaenisch R. 2000. Development - DNA methylation in Drosophila melanogaster. *Nature* **408**(6812): 538-540.

Mallory A, Vaucheret H. 2010. Form, function, and regulation of ARGONAUTE proteins. *The Plant cell* **22**(12): 3879-3889.

Martinez-Macias MI, Qian W, Miki D, Pontes O, Liu Y, Tang K, Liu R, Morales-Ruiz T, Ariza RR, Roldan-Arjona T et al. 2012. A DNA 3' phosphatase functions in active DNA demethylation in Arabidopsis. *Molecular cell* **45**(3): 357-370.

Martinowich K, Hattori D, Wu H, Fouse S, He F, Hu Y, Fan G, Sun YE. 2003. DNA methylation-related chromatin remodeling in activity-dependent BDNF gene regulation. *Science* **302**(5646): 890-893.

Matzke MA, Mosher RA. 2014. RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. *Nature reviews Genetics* **15**(6): 394-408.

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**(1): 116-127.

Mlotshwa S, Pruss GJ, Peragine A, Endres MW, Li J, Chen X, Poethig RS, Bowman LH, Vance V. 2008. DICER-LIKE2 plays a primary role in transitive silencing of transgenes in Arabidopsis. *PloS one* **3**(3): e1755.

Moissiard G, Cokus SJ, Cary J, Feng S, Billi AC, Stroud H, Husmann D, Zhan Y, Lajoie BR, McCord RP et al. 2012. MORC family ATPases required for heterochromatin condensation and gene silencing. *Science* **336**(6087): 1448-1451.

Mosher RA, Schwach F, Studholme D, Baulcombe DC. 2008. PolIVb influences RNA-directed DNA methylation independently of its role in siRNA biogenesis. *Proceedings of the National Academy of Sciences of the United States of America* **105**(8): 3145-3150.

Nan X, Ng H-H, Johnson CA, Laherty CD, Turner BM, Eisenman RN, Bird A. 1998. Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature* **393**(6683): 386-389.

Nathan D, Ingvarsdottir K, Sterner DE, Bylebyl GR, Dokmanovic M, Dorsey JA, Whelan KA, Krsmanovic M, Lane WS, Meluh PB et al. 2006. Histone sumoylation is a negative regulator in Saccharomyces cerevisiae and shows dynamic interplay with positive-acting histone modifications. *Genes & development* **20**(8): 966-976.

Naumann K, Fischer A, Hofmann I, Krauss V, Phalke S, Irmler K, Hause G, Aurich AC, Dorn R, Jenuwein T et al. 2005. Pivotal role of AtSUVH2 in heterochromatic histone methylation and gene silencing in Arabidopsis. *The EMBO journal* **24**(7): 1418-1429.

Nguyen AT, Zhang Y. 2011. The diverse functions of Dot1 and H3K79 methylation. *Genes & development* **25**(13): 1345-1358.

Nishioka K, Chuikov S, Sarma K, Erdjument-Bromage H, Allis CD, Tempst P, Reinberg D. 2002. Set9, a novel histone H3 methyltransferase that facilitates transcription by precluding histone tail modifications required for heterochromatin formation. *Genes & development* **16**(4): 479-489.

Okano M, Bell DW, Haber DA, Li E. 1999. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* **99**(3): 247-257.

Okano M, Xie S, Li E. 1998. Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. *Nature genetics* **19**(3): 219-220.

Ortega-Galisteo AP, Morales-Ruiz T, Ariza RR, Roldan-Arjona T. 2008. Arabidopsis DEMETER-LIKE proteins DML2 and DML3 are required for appropriate distribution of DNA methylation marks. *Plant molecular biology* **67**(6): 671-681.

Pikaard CS, Haag JR, Pontes OM, Blevins T, Cocklin R. 2012. A transcription fork model for Pol IV and Pol V-dependent RNA-directed DNA methylation. *Cold Spring Harbor symposia on quantitative biology* **77**: 205-212.

Pontes O, Li CF, Costa Nunes P, Haag J, Ream T, Vitins A, Jacobsen SE, Pikaard CS. 2006. The Arabidopsis chromatin-modifying nuclear siRNA pathway involves a nucleolar RNA processing center. *Cell* **126**(1): 79-92.

Qian W, Miki D, Lei M, Zhu X, Zhang H, Liu Y, Li Y, Lang Z, Wang J, Tang K et al. 2014. Regulation of active DNA demethylation by an alpha-crystallin domain protein in Arabidopsis. *Molecular cell* **55**(3): 361-371.

Qian W, Miki D, Zhang H, Liu Y, Zhang X, Tang K, Kan Y, La H, Li X, Li S et al. 2012. A histone acetyltransferase regulates active DNA demethylation in Arabidopsis. *Science* **336**(6087): 1445-1448.

Rajakumara E, Law JA, Simanshu DK, Voigt P, Johnson LM, Reinberg D, Patel DJ, Jacobsen SE. 2011. A dual flip-out mechanism for 5mC recognition by the Arabidopsis SUVH5 SRA domain and its impact on DNA methylation and H3K9 dimethylation in vivo. *Genes & development* **25**(2): 137-152.

Ratel D, Ravanat JL, Berger F, Wion D. 2006. N6-methyladenine: the other methylated base of DNA. *BioEssays : news and reviews in molecular, cellular and developmental biology* **28**(3): 309-315.

Rea S, Eisenhaber F, O'Carroll D, Strahl BD, Sun ZW, Schmid M, Opravil S, Mechtler K, Ponting CP, Allis CD et al. 2000. Regulation of chromatin structure by site-specific histone H3 methyltransferases. *Nature* **406**(6796): 593-599.

Ream TS, Haag JR, Wierzbicki AT, Nicora CD, Norbeck AD, Zhu JK, Hagen G, Guilfoyle TJ, Pasa-Tolic L, Pikaard CS. 2009. Subunit compositions of the RNA-silencing enzymes Pol IV and Pol V reveal their origins as specialized forms of RNA polymerase II. *Molecular cell* **33**(2): 192-203.

Roudier F, Ahmed I, Berard C, Sarazin A, Mary-Huard T, Cortijo S, Bouyer D, Caillieux E, Duvernois-Berthet E, Al-Shikhley L et al. 2011. Integrative epigenomic mapping defines four main chromatin states in Arabidopsis. *The EMBO journal* **30**(10): 1928-1938.

Rowley MJ, Avrutsky MI, Sifuentes CJ, Pereira L, Wierzbicki AT. 2011. Independent chromatin binding of ARGONAUTE4 and SPT5L/KTF1 mediates transcriptional gene silencing. *PLoS genetics* **7**(6): e1002120.

Santos-Rosa H, Schneider R, Bannister AJ, Sherriff J, Bernstein BE, Emre NC, Schreiber SL, Mellor J, Kouzarides T. 2002. Active genes are tri-methylated at K4 of histone H3. *Nature* **419**(6905): 407-411.

Schmitz RJ, Tamada Y, Doyle MR, Zhang X, Amasino RM. 2009. Histone H2B deubiquitination is required for transcriptional activation of FLOWERING LOCUS C and for proper control of flowering in Arabidopsis. *Plant physiology* **149**(2): 1196-1204.

Schneider R, Bannister AJ, Weise C, Kouzarides T. 2004. Direct binding of INHAT to H3 tails disrupted by modifications. *The Journal of biological chemistry* **279**(23): 23859-23862.

Sharif J, Muto M, Takebayashi SI, Suetake I, Iwamatsu A, Endo TA, Shinga J, Mizutani-Koseki Y, Toyoda T, Okamura K et al. 2007. The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA. *Nature* **450**(7171): 908-U925.

Shiio Y, Eisenman RN. 2003. Histone sumoylation is associated with transcriptional repression. *Proceedings of the National Academy of Sciences of the United States of America* **100**(23): 13225-13230.

Shirane K, Toh H, Kobayashi H, Miura F, Chiba H, Ito T, Kono T, Sasaki H. 2013. Mouse Oocyte Methylomes at Base Resolution Reveal Genome-Wide Accumulation of Non-CpG Methylation and Role of DNA Methyltransferases. *PLoS genetics* **9**(4).

Simpson VJ, Johnson TE, Hammen RF. 1986. Caenorhabditis-Elegans DNA Does Not Contain 5-Methylcytosine at Any Time during Development or Aging. *Nucleic acids research* **14**(16): 6711-6719.

Smith LM, Pontes O, Searle I, Yelina N, Yousafzai FK, Herr AJ, Pikaard CS, Baulcombe DC. 2007. An SNF2 protein associated with nuclear RNA silencing and the spread of a silencing signal between cells in Arabidopsis. *The Plant cell* **19**(5): 1507-1521.

Song J, Irwin J, Dean C. 2013. Remembering the prolonged cold of winter. *Current biology : CB* **23**(17): R807-811.

Springer NM, Napoli CA, Selinger DA, Pandey R, Cone KC, Chandler VL, Kaeppler HF, Kaeppler SM. 2003. Comparative analysis of SET domain proteins in maize and Arabidopsis reveals multiple duplications preceding the divergence of monocots and dicots. *Plant physiology* **132**(2): 907-925.

Sridhar VV, Kapoor A, Zhang K, Zhu J, Zhou T, Hasegawa PM, Bressan RA, Zhu JK. 2007. Control of DNA methylation and heterochromatic silencing by histone H2B deubiquitination. *Nature* **447**(7145): 735-738.

Stassen MJ, Bailey D, Nelson S, Chinwalla V, Harte PJ. 1995. The Drosophila trithorax proteins contain a novel variant of the nuclear receptor type DNA binding domain and an ancient conserved motif found in other chromosomal proteins. *Mechanisms of development* **52**(2-3): 209-223.

Stewart MD, Li JW, Wong JM. 2005. Relationship between histone H3 lysine 9 methylation, transcription repression, and heterochromatin protein 1 recruitment. *Molecular and Cellular Biology* **25**(7): 2525-2538.

Strahl BD, Allis CD. 2000. The language of covalent histone modifications. *Nature* **403**(6765): 41-45.

Strahl BD, Ohba R, Cook RG, Allis CD. 1999. Methylation of histone H3 at lysine 4 is highly conserved and correlates with transcriptionally active nuclei in Tetrahymena. *Proceedings of the National Academy of Sciences of the United States of America* **96**(26): 14967-14972.

Stroud H, Do T, Du J, Zhong X, Feng S, Johnson L, Patel DJ, Jacobsen SE. 2014. Non-CG methylation patterns shape the epigenetic landscape in Arabidopsis. *Nature structural & molecular biology* **21**(1): 64-72.

Stroud H, Greenberg MV, Feng S, Bernatavichute YV, Jacobsen SE. 2013. Comprehensive analysis of silencing mutants reveals complex regulation of the Arabidopsis methylome. *Cell* **152**(1-2): 352-364.

Stroud H, Otero S, Desvoyes B, Ramirez-Parra E, Jacobsen SE, Gutierrez C. 2012. Genome-wide analysis of histone H3.1 and H3.3 variants in Arabidopsis thaliana. *Proceedings of the National Academy of Sciences of the United States of America* **109**(14): 5370-5375.

Su X, Zhu G, Ding X, Lee SY, Dou Y, Zhu B, Wu W, Li H. 2014. Molecular basis underlying histone H3 lysine-arginine methylation pattern readout by Spin/Ssty repeats of Spindlin1. *Genes & development* **28**(6): 622-636.

Tachibana M, Sugimoto K, Nozaki M, Ueda J, Ohta T, Ohki M, Fukuda M, Takeda N, Niida H, Kato H et al. 2002. G9a histone methyltransferase plays a dominant role in euchromatic histone H3 lysine 9 methylation and is essential for early embryogenesis. *Genes & development* **16**(14): 1779-1791.

Tahiliani M, Koh KP, Shen Y, Pastor WA, Bandukwala H, Brudno Y, Agarwal S, Iyer LM, Liu DR, Aravind L et al. 2009. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324**(5929): 930-935.

Takeda A, Iwasaki S, Watanabe T, Utsumi M, Watanabe Y. 2008. The mechanism selecting the guide strand from small RNA duplexes is different among Argonaute proteins. *Plant and Cell Physiology* **49**(4): 493-500.

Tamada Y, Yun JY, Woo SC, Amasino RM. 2009. ARABIDOPSIS TRITHORAX-RELATED7 is required for methylation of lysine 4 of histone H3 and for transcriptional activation of FLOWERING LOCUS C. *The Plant cell* **21**(10): 3257-3269.

Tan MJ, Luo H, Lee S, Jin FL, Yang JS, Montellier E, Buchou T, Cheng ZY, Rousseaux S, Rajagopal N et al. 2011. Identification of 67 Histone Marks and Histone Lysine Crotonylation as a New Type of Histone Modification. *Cell* **146**(6): 1015-1027.

Tang Y, Gao XD, Wang YS, Yuan BF, Feng YQ. 2012. Widespread Existence of Cytosine Methylation in Yeast DNA Measured by Gas Chromatography/Mass Spectrometry. *Anal Chem* **84**(16): 7249-7255.

Thorstensen T, Fischer A, Sandvik SV, Johnsen SS, Grini PE, Reuter G, Aalen RB. 2006. The Arabidopsis SUVR4 protein is a nucleolar histone methyltransferase with preference for monomethylated H3K9. *Nucleic acids research* **34**(19): 5461-5470.

Till S, Ladurner AG. 2007. RNA Pol IV plays catch with Argonaute 4. *Cell* **131**(4): 643-645.

Tschiersch B, Hofmann A, Krauss V, Dorn R, Korge G, Reuter G. 1994. The Protein Encoded by the Drosophila Position-Effect Variegation Suppressor Gene Su(Var)3-9 Combines Domains of Antagonistic Regulators of Homeotic Gene Complexes. *Embo Journal* **13**(16): 3822-3831.

Valencia-Morales Mdel P, Camas-Reyes JA, Cabrera-Ponce JL, Alvarez-Venegas R. 2012. The Arabidopsis thaliana SET-domain-containing protein ASHH1/SDG26 interacts with itself and with distinct histone lysine methyltransferases. *Journal of plant research* **125**(5): 679-692.

Vaucheret H. 2008. Plant ARGONAUTES. *Trends in plant science* **13**(7): 350-358.

Veiseth SV, Rahman MA, Yap KL, Fischer A, Egge-Jacobsen W, Reuter G, Zhou MM, Aalen RB, Thorstensen T. 2011. The SUVR4 histone lysine methyltransferase binds ubiquitin and converts H3K9me1 to H3K9me3 on transposon chromatin in Arabidopsis. *PLoS genetics* **7**(3): e1001325.

Vongs A, Kakutani T, Martienssen RA, Richards EJ. 1993. Arabidopsis thaliana DNA methylation mutants. *Science* **260**(5116): 1926-1928.

Wang XB, Jovel J, Udomporn P, Wang Y, Wu Q, Li WX, Gasciolli V, Vaucheret H, Ding SW. 2011. The 21-nucleotide, but not 22-nucleotide, viral secondary small interfering RNAs direct potent antiviral defense by two cooperative argonautes in Arabidopsis thaliana. *The Plant cell* **23**(4): 1625-1638.

Wang Y, Li Y, Toth JI, Petroski MD, Zhang Z, Zhao JC. 2014. N6-methyladenosine modification destabilizes developmental regulators in embryonic stem cells. *Nature cell biology* **16**(2): 191-198.

Wassenegger M, Heimes S, Riedel L, Sanger HL. 1994. RNA-directed de novo methylation of genomic sequences in plants. *Cell* **76**(3): 567-576.

Weake VM, Workman JL. 2008. Histone ubiquitination: triggering gene activity. *Molecular cell* **29**(6): 653-663.

Wei CM, Gershowitz A, Moss B. 1975. Methylated nucleotides block 5' terminus of HeLa cell messenger RNA. *Cell* **4**(4): 379-386.

Wei Y, Mizzen CA, Cook RG, Gorovsky MA, Allis CD. 1998. Phosphorylation of histone H3 at serine 10 is correlated with chromosome condensation during

mitosis and meiosis in Tetrahymena. *Proceedings of the National Academy of Sciences of the United States of America* **95**(13): 7480-7484.

Weiberg A, Wang M, Lin FM, Zhao H, Zhang Z, Kaloshian I, Huang HD, Jin H. 2013. Fungal small RNAs suppress plant immunity by hijacking host RNA interference pathways. *Science* **342**(6154): 118-123.

Wierzbicki AT, Cocklin R, Mayampurath A, Lister R, Rowley MJ, Gregory BD, Ecker JR, Tang H, Pikaard CS. 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.

Wierzbicki AT, Haag JR, Pikaard CS. 2008. Noncoding transcription by RNA polymerase Pol IVb/Pol V mediates transcriptional silencing of overlapping and adjacent genes. *Cell* **135**(4): 635-648.

Wierzbicki AT, Ream TS, Haag JR, Pikaard CS. 2009. RNA polymerase V transcription guides ARGONAUTE4 to chromatin. *Nature genetics* **41**(5): 630-634.

Woo HR, Pontes O, Pikaard CS, Richards EJ. 2007. VIM1, a methylcytosine-binding protein required for centromeric heterochromatinization. *Genes & development* **21**(3): 267-277.

Wu H, Zhang Y. 2014. Reversing DNA methylation: mechanisms, genomics, and biological functions. *Cell* **156**(1-2): 45-68.

Wu LP, Zee BM, Wang YM, Garcia BA, Dou YL. 2011. The RING Finger Protein MSL2 in the MOF Complex Is an E3 Ubiquitin Ligase for H2B K34 and Is Involved in Crosstalk with H3 K4 and K79 Methylation. *Molecular cell* **43**(1): 132-144.

Xie M, Ren G, Zhang C, Yu B. 2012a. The DNA- and RNA-binding protein FACTOR of DNA METHYLATION 1 requires XH domain-mediated complex formation for its function in RNA-directed DNA methylation. *The Plant journal : for cell and molecular biology* **72**(3): 491-500.

Xie W, Barr CL, Kim A, Yue F, Lee AY, Eubanks J, Dempster EL, Ren B. 2012b. Base-Resolution Analyses of Sequence and Parent-of-Origin Dependent DNA Methylation in the Mouse Genome. *Cell* **148**(4): 816-831.

Xie Z, Johansen LK, Gustafson AM, Kasschau KD, Lellis AD, Zilberman D, Jacobsen SE, Carrington JC. 2004. Genetic and functional diversification of small RNA pathways in plants. *PLoS biology* **2**(5): E104.

Xie Z, Kasschau KD, Carrington JC. 2003. Negative feedback regulation of Dicer-Like1 in Arabidopsis by microRNA-guided mRNA degradation. *Current biology : CB* **13**(9): 784-789.

Xu L, Zhao Z, Dong A, Soubigou-Taconnat L, Renou JP, Steinmetz A, Shen WH. 2008. Di- and tri- but not monomethylation on histone H3 lysine 36 marks active transcription of genes involved in flowering time regulation and other processes in Arabidopsis thaliana. *Mol Cell Biol* **28**(4): 1348-1360.

Ye R, Wang W, Iki T, Liu C, Wu Y, Ishikawa M, Zhou X, Qi Y. 2012. Cytoplasmic assembly and selective nuclear import of Arabidopsis Argonaute4/siRNA complexes. *Molecular cell* **46**(6): 859-870.

Yelagandula R, Stroud H, Holec S, Zhou K, Feng S, Zhong X, Muthurajan UM, Nie X, Kawashima T, Groth M et al. 2014. The histone variant H2A.W defines heterochromatin and promotes chromatin condensation in Arabidopsis. *Cell* **158**(1): 98-109.

Yu B, Bi L, Zhai J, Agarwal M, Li S, Wu Q, Ding SW, Meyers BC, Vaucheret H, Chen X. 2010. siRNAs compete with miRNAs for methylation by HEN1 in Arabidopsis. *Nucleic acids research* **38**(17): 5844-5850.

Zaina S, Lund G. 2013. Atherosclerosis: cell biology and lipoproteins--panoramic views of DNA methylation landscapes of atherosclerosis. *Current opinion in lipidology* **24**(4): 369-370.

Zemach A, Grafi G. 2007. Methyl-CpG-binding domain proteins in plants: interpreters of DNA methylation. *Trends in plant science* **12**(2): 80-85.

Zemach A, Kim MY, Hsieh PH, Coleman-Derr D, Eshed-Williams L, Thao K, Harmer SL, Zilberman D. 2013. The Arabidopsis nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin. *Cell* **153**(1): 193-205.

Zemach A, Li Y, Wayburn B, Ben-Meir H, Kiss V, Avivi Y, Kalchenko V, Jacobsen SE, Grafi G. 2005. DDM1 binds Arabidopsis methyl-CpG binding domain proteins and affects their subnuclear localization. *The Plant cell* **17**(5): 1549-1558.

Zhang CJ, Ning YQ, Zhang SW, Chen Q, Shao CR, Guo YW, Zhou JX, Li L, Chen S, He XJ. 2012. IDN2 and its paralogs form a complex required for RNA-directed DNA methylation. *PLoS genetics* **8**(5): e1002693.

Zhang CJ, Zhou JX, Liu J, Ma ZY, Zhang SW, Dou K, Huang HW, Cai T, Liu R, Zhu JK et al. 2013. The splicing machinery promotes RNA-directed DNA methylation

and transcriptional silencing in Arabidopsis. *The EMBO journal* **32**(8): 1128-1140.

Zhang H, Tang K, Qian W, Duan CG, Wang B, Zhang H, Wang P, Zhu X, Lang Z, Yang Y et al. 2014. An Rrp6-like protein positively regulates noncoding RNA levels and DNA methylation in Arabidopsis. *Molecular cell* **54**(3): 418-430.

Zhang X, Bernatavichute YV, Cokus S, Pellegrini M, Jacobsen SE. 2009. Genome-wide analysis of mono-, di- and trimethylation of histone H3 lysine 4 in Arabidopsis thaliana. *Genome biology* **10**(6): R62.

Zhang X, Henderson IR, Lu C, Green PJ, Jacobsen SE. 2007. Role of RNA polymerase IV in plant small RNA metabolism. *Proceedings of the National Academy of Sciences of the United States of America* **104**(11): 4536-4541.

Zhang Y, Ng H-H, Erdjument-Bromage H, Tempst P, Bird A, Reinberg D. 1999. Analysis of the NuRD subunits reveals a histone deacetylase core complex and a connection with DNA methylation. *Genes & Development* **13**(15): 1924-1935.

Zhao Y, Xie S, Li X, Wang C, Chen Z, Lai J, Gong Z. 2014. REPRESSOR OF SILENCING5 Encodes a Member of the Small Heat Shock Protein Family and Is Required for DNA Demethylation in Arabidopsis. *The Plant cell* **26**(6): 2660-2675.

Zheng X, Pontes O, Zhu J, Miki D, Zhang F, Li WX, Iida K, Kapoor A, Pikaard CS, Zhu JK. 2008. ROS3 is an RNA-binding protein required for DNA demethylation in Arabidopsis. *Nature* **455**(7217): 1259-1262.

Zheng X, Zhu J, Kapoor A, Zhu JK. 2007. Role of Arabidopsis AGO6 in siRNA accumulation, DNA methylation and transcriptional gene silencing. *The EMBO journal* **26**(6): 1691-1701.

Zheng Z, Xing Y, He XJ, Li W, Hu Y, Yadav SK, Oh J, Zhu JK. 2010. An SGS3-like protein functions in RNA-directed DNA methylation and transcriptional gene silencing in Arabidopsis. *The Plant journal : for cell and molecular biology* **62**(1): 92-99.

Zhong X, Du J, Hale CJ, Gallego-Bartolome J, Feng S, Vashisht AA, Chory J, Wohlschlegel JA, Patel DJ, Jacobsen SE. 2014. Molecular mechanism of action of plant DRM de novo DNA methyltransferases. *Cell* **157**(5): 1050-1060.

Zhong X, Hale CJ, Law JA, Johnson LM, Feng S, Tu A, Jacobsen SE. 2012. DDR complex facilitates global association of RNA polymerase V to promoters and

evolutionarily young transposons. *Nature structural & molecular biology* **19**(9): 870-875.

Zhu JK. 2009. Active DNA demethylation mediated by DNA glycosylases. *Annual review of genetics* **43**: 143-166.

Zhu Y, Rowley MJ, Bohmdorfer G, Wierzbicki AT. 2013. A SWI/SNF chromatin-remodeling complex acts in noncoding RNA-mediated transcriptional silencing. *Molecular cell* **49**(2): 298-309.

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

Chapter 2. Detection of Pol IV/RDR2-dependent transcripts at the genomic scale in *Arabidopsis* reveals features and regulation of siRNA biogenesis

Abstract

24 nucleotide small interfering (siRNAs) are central players in RNA-directed DNA methylation (RdDM), a process that establishes and maintains DNA methylation at transposable elements to ensure genome stability in plants. The plant-specific RNA polymerase IV (Pol IV) is required for siRNA biogenesis and is thought to transcribe RdDM loci to produce primary transcripts that are converted to double-stranded RNAs (dsRNAs) by RDR2 to serve as siRNA precursors. Yet, no such siRNA precursor transcripts have ever been reported. Here, through genome-wide profiling of RNAs in genotypes that compromise the processing of siRNA precursors, we were able to identify Pol IV/RDR2-dependent transcripts from tens of thousands of loci. We show that Pol IV/RDR2-dependent transcripts correspond to both DNA strands, while the RNA polymerase II (Pol II)-dependent transcripts produced upon de-repression of the loci are derived primarily from one strand. We also show that Pol IV/RDR2-dependent transcripts have a 5' monophosphate, lack a polyA tail at the 3' end, and contain no introns; these features distinguish them from Pol II-dependent transcripts. Like Pol II-transcribed genic regions, Pol IV-transcribed regions are flanked by A/T-rich sequences depleted in nucleosomes, which highlights similarities in Pol II- and Pol IV-mediated transcription. Computational analysis of siRNA abundance from various mutants reveals differences in the regulation of siRNA biogenesis at two types of loci that undergo CHH methylation

via two different DNA methyltransferases. These findings begin to reveal features of Pol IV/RDR2-mediated transcription at the heart of genome stability in plants.

Introduction

In plants and mammals, DNA methylation influences gene expression and represses transposable elements (TEs) to ensure genome stability. DNA methylation occurs at CG, CHG and CHH (H represents A, C, or G) sequence contexts in plants (Law and Jacobsen 2010). In *Arabidopsis*, the methyltransferases DRM2 and CMT2 establish DNA methylation in all sequence contexts and maintain asymmetric CHH methylation (Cao and Jacobsen 2002; Zemach et al. 2013). The maintenance of symmetric CG and CHG methylation is mediated by MET1 and CMT3, respectively (Stroud et al. 2013).

In *Arabidopsis*, RNA-dependent DNA Methylation (RdDM) mediated by DRM2 deposits DNA methylation at TEs to cause their transcriptional silencing (Wierzbicki et al. 2008). 24 nucleotide (nt) siRNAs serve as the sequence determinants that guide DRM2 to RdDM target loci (Mosher et al. 2008). The plant-specific RNA polymerase IV (Pol IV) is thought to transcribe the RdDM loci to produce single-stranded RNAs (ssRNAs), which are converted to double-stranded RNAs (dsRNAs) by RNA-DEPENDENT RNA POLYMERASE2 (RDR2) (Xie et al. 2004; Jia et al. 2009). DICER-LIKE3 (DCL3) cleaves the dsRNAs to generate 24 nt siRNAs (Cho et al. 2008; Liu et al. 2009), which associate with AGO4 (Qi et al. 2006). Another plant-specific RNA polymerase, Pol V, produces nascent non-coding transcripts that recruit siRNA-containing AGO4 to RdDM loci (Wierzbicki et al. 2009) with the assistance of both the

SUVH2/9 proteins (Johnson et al. 2014) and the DDR complex (composed of DRD1, DMS3 and RDM1) (Zhong et al. 2012; Johnson et al. 2014). This association aids the recruitment of DRM2 leading to cytosine methylation (Law et al. 2010; Zhong et al. 2014). In the genome, loci that produce siRNAs are highly correlated with those that harbor CHH methylation (Lister et al. 2008). Loss of siRNAs in mutants of *NRPD1* encoding the largest subunit of Pol IV or *RDR2* results in decreased CHH methylation at numerous loci, usually those residing in euchromatic chromosomal arms and requiring DRM2 for methylation (Wierzbicki et al. 2012). Another pathway mediated by CMT2 together with DDM1 and histone H1 also maintains CHH methylation, but mainly acts at pericentromeric regions (Zemach et al. 2013; Stroud et al. 2014). Together, DRM2 and CMT2 are responsible for nearly all CHH methylation in the genome, and the DRM2-targeted and CMT2-targeted sites are non-overlapping. Although siRNAs are generated at both DRM2-targeted and CMT2-targeted sites, siRNAs are not required for the maintenance of CHH methylation at CMT2-targeted sites (Zemach et al. 2013; Stroud et al. 2014).

Many factors that participate in siRNA biogenesis are known. Some, such as Pol IV and RDR2 are essential, while others such as DCL3, CLASSY1, and SHH1 play a more limited role (Henderson et al. 2006; Smith et al. 2007; Law et al. 2011; Law et al. 2013). In the absence of DCL3, which generates 24 nt siRNAs, DCL2 and DCL4 produce endogenous siRNAs of 22 nt and 21 nt, respectively (Allen et al. 2005; Wang et al. 2011). Although Pol IV is purported to produce siRNA precursors, Pol IV-dependent transcripts have never been reported. One difficulty in the detection of Pol IV-dependent

transcripts is that they are probably short-lived, as they are likely quickly cleaved by DCL proteins upon their conversion into dsRNAs. The second difficulty lies in the fact that siRNA loci are silenced in wild type and de-repressed in Pol IV mutants (Herr et al. 2005; Pontier et al. 2005). This prevents the identification of Pol IV-dependent transcripts by searching for RNAs that are diminished in Pol IV mutants. The lack of knowledge of the Pol IV-dependent transcripts impedes a mechanistic understanding of siRNA biogenesis.

We reasoned that comparing RNAs between *NRPD1* and *nrpd1* genotypes in a *dcl2 dcl3 dcl4* triple mutant background would circumvent the difficulties in detecting Pol IV-dependent transcripts. In the *dcl2 dcl3 dcl4* background, Pol IV-dependent transcripts should be stabilized due to reduced processing by the DCLs. In addition, in the *dcl2 dcl3 dcl4* mutant background, RdDM loci are already de-repressed (Xie et al. 2004; Henderson et al. 2006) such that loss of function in *NRPD1* would not cause any further de-repression. Therefore, we sought to identify Pol IV-dependent transcripts by RNA sequencing (RNA-seq) and Pol IV/RDR2-dependent transcripts by dsRNA-seq in *dcl2 dcl3 dcl4* and *dcl2 dcl3 dcl4 nrpd1*. This effort led to the identification of Pol IV/RDR2-dependent transcripts from tens of thousands of genomic loci. Further molecular and bioinformatics analyses revealed features of Pol IV/RDR2-dependent transcripts as well as the genetic and epigenetic requirements for Pol IV transcription.

Results

Genome-wide discovery of Pol IV-dependent transcripts as siRNA precursors

To detect Pol IV-dependent transcripts, we compared the transcriptome of the *dcl2-1 dcl3-1 dcl4-2 nrpd1-3* quadruple mutant with that of the *dcl2-1 dcl3-1 dcl4-2* triple mutant (hereafter referred to as *dcl234*) through RNA sequencing (RNA-seq). Three biological replicates were conducted for each genotype using inflorescences containing unopened flower buds. To derive Pol IV-dependent siRNA loci from the same tissue types, small RNA sequencing (sRNA-seq) was performed with wild-type (WT) and *nrpd1-3* inflorescences. RNA-seq revealed 698 regions showing statistically significant reduction in transcript levels in *dcl234 nrpd1* relative to *dcl234* (Figures 2.1A-C and 2S.1). 47,442 Pol IV-dependent siRNA (hereafter referred to as P4siRNA) regions were identified from sRNA-seq (Figures 2.1A-C). 635 of the 698 regions that generated Pol IV-dependent transcripts (hereafter referred as P4RNAs) overlapped with the P4siRNA regions (Figure 2.1C), suggesting that the 635 regions are potential siRNA precursor regions. 22 of these regions (Table 2.1) were randomly selected for detection of P4RNAs by RT-PCR. P4RNAs were detected at all these loci in *dcl234*; these transcripts were either non-detectable or were reduced in abundance in *dcl234 nrpd1* (Figures 2.1D and 2S.2A). Therefore, our RNA-seq efforts resulted in the identification of hundreds of regions generating P4RNAs.

The 635 regions shown to produce P4RNAs above only constituted 1.3% of the 47,442 P4siRNA regions. We found that 98% of the P4siRNA regions had little read

coverage in the RNA-seq libraries. In *dcl234*, approximately 90% of the reads in the RNA-seq libraries were from genic regions, and less than 5% of the reads were from P4siRNA loci (Figure 2S.2B). Enrichment for P4RNAs in the total RNA population was necessary for the discovery of more P4RNAs.

P4RNAs are thought to be converted to dsRNAs by RDR2 before being processed to P4siRNAs, so the P4RNAs should exist as dsRNAs in the *dcl234* background. We sought to confirm the dsRNA nature of P4RNAs that were detected through RNA-seq above. We performed strand-specific RT-PCR using region- and strand-specific primers for reverse transcription. Indeed, transcripts corresponding to both DNA strands were detected in *dcl234* and the abundance of the transcripts was greatly reduced in *dcl234 nrpd1* (Figure 2S.3). Therefore, P4RNAs could be potentially enriched by separation of dsRNAs from ssRNAs.

We performed three biological replicates of dsRNA-seq in *dcl234* and *dcl234 nrpd1* to enrich for P4RNAs (Zheng et al. 2010a). Indeed, the percentage of gene-mapping reads was greatly reduced in dsRNA-seq compared to that from RNA-seq (Figure 2S.2B). While 35% of the reads mapped to P4siRNA loci in *dcl234*, only 5% did in *dcl234 nrpd1* (Figure 2S.2B), suggesting that there was differential expression at P4siRNA loci between the two genotypes. Indeed, 24,035 regions were found to have a statistically significant reduction in transcript abundance in *dcl234 nrpd1* (Figures 2.1A, 2.1B, and 2S.1). 22,990 of these regions overlapped with the 47,442 P4siRNA regions (Figure 2.1C). We consider these 22,990 regions as generating detectable P4siRNA precursors.

Having detected P4RNA-generating regions, we next asked whether all these regions produce P4siRNAs. Our sRNA-seq detected P4siRNAs at 22,990 of the 24,035 regions where P4RNAs were detected by dsRNA-seq. For the 1045 regions from which P4siRNAs were not detected, 946 showed a reduction in small RNA read abundance in *nrpd1* relative to wild type, but these regions did not pass our stringent filter for the definition of differential P4siRNA expression (four-fold reduction in *nrpd1* relative to WT with p-value <0.01). Therefore, these regions were also likely to produce P4siRNAs. This suggests that most (if not all) P4RNAs serve as P4siRNA precursors.

A previous study identified 982 genomic loci bound by Pol IV, among which 787 had detectable siRNA production (Law et al. 2013). The P4RNA-generating regions overlapped with 445 of the 982 regions bound by Pol IV and 405 of the 787 regions producing siRNAs. This does not suggest that only half of the Pol IV-occupied regions produce P4RNAs, but rather, this was likely due to the fact that our approach only uncovered P4RNAs at approximately half of the regions generating P4siRNAs in the genome (see below).

RDR2 has a similar effect as Pol IV on the abundance of P4RNAs

We tested whether *RDR2* is required for the accumulation of P4RNAs. We evaluated the effects of loss of function in *RDR2* on P4RNA levels by performing RT-PCR on *dcl234*, *dcl234 nrpd1* and *dcl234 rdr2* at five P4siRNA loci. No transcripts were detected in *dcl234 rdr2* or *dcl234 nrpd1* at these loci (Figure 2.2A), indicating that the P4RNAs were dependent on both Pol IV and RDR2. The complete lack of P4RNAs in *dcl234 rdr2* was

surprising, as we expected to be able to detect P4RNAs in the absence of *RDR2* based on the current RdDM model in which Pol IV generates an ssRNAs that are converted to dsRNAs by RDR2.

To further examine the *in vivo* effects of the *rdr2* mutation, we performed RNA-seq with *dcl234*, *dcl234 nrpd1* and *dcl234 rdr2*. To increase the sensitivity of RNA-seq, we enriched for low abundance transcripts through DSN normalization (see Methods), which resulted in a moderate increase in read coverage at P4siRNA loci (Figure 2.2B). As a result, 864 P4RNA regions were identified by comparing *dcl234* to *dcl234 nrpd1* (four fold difference, p-value<0.01) in RNA-seq-DSN as compared to 698 from RNA-seq (described before). With the same criteria (four fold difference, p-value<0.01), 968 regions were found to produce transcripts in *dcl234* relative to *dcl234 rdr2*. 850 regions were common (Figure 2.2C), suggesting that the transcripts were dependent on both Pol IV and RDR2. Furthermore, at these 850 loci, the abundance of residual reads from *dcl234 rdr2* was not any higher than that from *dcl234 nrpd1* (Figure 2.2D). The results of RT-PCR and RNA-seq-DSN suggest that RDR2 has the same effect on the production of P4RNAs as does Pol IV.

Assembly of Pol IV/RDR2-dependent transcripts and examination of their surrounding genomic features

With the regions generating P4RNAs known, we next assembled P4RNAs using reads from the dsRNA-seq libraries (see Methods;). A total of 17,606 P4RNAs were assembled with most being in the range of 100 to 500 nt (Figure 2S.4).

A profound A/T enrichment was found for regions surrounding P4RNAs. We aligned all P4RNAs at their 5’ or 3’ ends and determined the proportion of A/T at each nucleotide position in the 1000 nucleotide window upstream of the 5’ end or downstream of the 3’ end. Since the P4RNAs were double-stranded and the actual orientation of P4RNAs was unknown, the 5’ ends of transcripts were defined as the beginning nucleotides on the Watson strand of the TAIR10 reference sequence. The A/T composition was obviously much lower in the P4RNA bodies than the surrounding regions (Figures 2.3A, B), which could simply reflect the GC-richness of P4RNA regions. However, in the ~50 nt regions flanking P4RNA ends, there was a clear increase in A/T richness relative to the regions further away, suggesting that the immediate flanking regions of P4RNAs are A/T rich. A closer examination of the 5’ or 3’ ends showed that the ends had the lowest A/T composition while the flanking nucleotides had higher A/T composition (Figures 2.3A, B, insets). Such patterns of A/T distribution were also found for annotated exons (Figures 2.3C, D) and at Pol II transcription start sites (TSS) or termination sites (TTS) (Figures2.3E, F), although the A/T skew at TSS and TTS sites was not as strong.

Nucleosomes, units of chromatin that influence the access of protein factors to the DNA, are known to be enriched on exons and A/T poor regions (Chodavarapu et al. 2010). We determined the nucleosome occupancy at P4RNA regions using published nucleosome sequencing data (Chodavarapu et al. 2010). Nucleosomes were depleted at both the 5’ and 3’ flanking sequences of P4RNAs and enriched at the ends of P4RNAs (Figures 2.3G, H). Such nucleosome distribution patterns resembled those on exons and

at the TSS of genes (Figures 2.3G, H) (Chodavarapu et al. 2010; Ammar et al. 2012). These results suggest that the initiation of Pol IV and/or RDR2 transcription occurs in A/T-rich and nucleosome-depleted regions.

The genomic distribution of P4RNAs was also examined. P4RNAs were mainly present at intergenic regions. 65% of them overlapped with annotated TEs or repeats; only 9% of them overlapped with genes (Figure 2S.5A). We performed GO analysis on the set of genes overlapping with P4RNA loci. Intriguingly, the GO term "endomembrane system" was highly enriched for the gene set (Table 2.2). To determine whether this unexpected association was due to the concentration of "endomembrane system" genes at pericentromeric regions, we examined the chromosomal distributions of the set of genes overlapping with P4RNAs. We found that the gene set resembled the set of all annotated genes in that the genes were dispersed at euchromatic regions and depleted at pericentromeric regions (Figure 2S.6A).

We next examined the association between regions generating P4RNAs and heterochromatic marks. We first examined the relationship among P4RNAs, P4siRNAs, and CHH regions dependent on DRM2 or CMT2. DRM2- and CMT2-dependent CHH methylation regions were defined as the CHH Differentially Methylated Regions (CHH DMRs) with reduced methylation in *drm1 drm2* and *cmt2* relative to WT, respectively, in a published methylome study (Stroud et al. 2013). Similarly, Pol IV-dependent CHH regions were defined as CHH DMRs between WT and *nrpd1* in the same study (Stroud et al. 2013). Although both DRM2- and CMT2-targeted sites strongly overlapped with regions producing P4siRNAs (Figure 2S.5B), the sites targeted by the two

methyltransferases are largely non-overlapping (Figure 2S.5C) (Zemach et al. 2013). Pol IV-dependent CHH regions are mainly targeted by DRM2 (Figure 2S.5C), and the number of Pol IV/DRM2-dependent CHH regions is only half of the number of CMT2-dependent CHH regions. Therefore, loss of P4siRNAs only leads to reduction in CHH methylation at a small proportion of P4siRNA loci, and these loci are distributed along euchromatic chromosomal arms (Figure 2S.6B) (Wierzbicki et al. 2012). We found that the chromosomal distribution of P4RNAs strongly resembled those of total CHH methylation and CMT2-dependent CHH methylation, which peak at pericentromeric regions (Figure 2S.6B) (Lister et al. 2008; Zemach et al. 2013). This suggests that loci with detected P4RNAs are largely contributed by those whose CHH methylation is targeted by CMT2, which will be further examined later.

Besides siRNAs and DNA methylation, H3 lysine 27 monomethylation (H3K27me1) and H3 lysine 9 dimethylation (H3K9me2) are two other common heterochromatic marks, for which the genomic distributions were profiled through ChIP-chip (Roudier et al. 2011; Deleris et al. 2012). We found that these two marks exhibited similar chromosomal distributions as P4RNAs – all were enriched at pericentromeric regions (Figure 2S.6C).

Features of Pol IV/RDR2-dependent transcripts

The 5’ initiating nucleotides of Pol I and Pol III transcripts have triphosphate groups and those of Pol II transcripts contain 7-methylguanosine caps. To determine the 5’ end structure of P4RNAs, we performed enzymatic treatments of total RNAs followed by the

detection of P4RNAs by RT-PCR. First, we treated total RNAs with no enzyme (control), Tobacco Acid Pyrophosphatase (TAP), which converts 5' triphosphate or 5' 7-methylguanylate cap to 5' monophosphate, or T4 Polynucleotide Kinase (PNK), which adds a 5' phosphate group to 5' hydroxyl RNAs. Next, we digested the RNAs with Terminator, a 5' to 3' exonuclease that acts on RNAs with a 5' monophosphate. Finally, RT-PCR was conducted on these treated RNA samples to detect various P4RNAs. The RNAs treated with Terminator alone showed a dramatic reduction in the abundance of P4RNAs (Figure 2.4A), suggesting that a large portion of P4RNAs had a 5' monophosphate. The samples treated with TAP or PNK followed by Terminator showed similar levels of P4RNAs to the sample treated with Terminator alone (Figure 2.4A). The fact that TAP or PNK treatment did not increase the amount of 5' monophosphate RNAs indicated that P4RNAs primarily had a 5' monophosphate.

Introns are a common feature of Pol II-dependent transcripts. To determine whether the P4RNAs have introns, we first analyzed reads from *dcl234* dsRNA-seq libraries with TopHat2 (Kim et al. 2013), a widely-used software to discover splice junctions for canonical introns. 20,521 spliced junctions were reported through TopHat2, with 20,378 junctions being at genic regions and only 59 junctions being at P4RNA regions. As P4RNAs do not necessarily use splice junctions characteristic of Pol II-dependent transcripts, we also employed a naïve method that reports all spliced reads, i.e., reads whose 5' and 3' portions represent nearby genomic sequences separated by a segment (see Methods). This method predicted 16,018 spliced reads, with 12,670 being at genic regions and only 112 being at P4RNA regions. The potential spliced junctions

predicted by the two methods at P4RNA regions were further examined to determine whether they represented true spliced junctions. The levels of transcripts at intron regions should be much lower than those at exon regions. When subjected to the filter that the coverage of "intron" regions is at least five times lower than that of the flanking regions, none of the predicted junctions was retained. This suggests that P4RNAs do not possess introns.

Polyadenylation is part of the maturation process of Pol II-dependent transcripts. To determine whether P4RNAs have polyA tails, total RNAs were separated into polyA+ and polyA- fractions followed by the detection of P4RNAs by RT-PCR. P4RNAs were detectable from total RNAs and polyA- RNAs, but not from polyA+ RNAs, suggesting that P4RNAs do not have polyA tails (Figure 2.4B).

Given that P4RNAs lack polyA tails and Pol II-dependent transcripts are expected to have polyA tails, we sought to distinguish the two types of transcripts at RdDM loci through the presence or absence of polyA tails. polyA- and polyA+ RNAs were first isolated from two biological replicates of *dcl234* and *dcl234 nrpd1* and subjected to an RNA-seq library construction procedure that preserved the strandedness of the transcripts. In polyA+ libraries, a total of 1,639 P4siRNA loci were found to have read coverage above 1 RPM, indicating that they were expressed. Transcript abundance at these loci was similar in *dcl234 nrpd1* and *dcl234* (Figure 2.4C), suggesting that the polyA+ transcripts were made by Pol II rather than Pol IV. Next we examined the read coverage at the 698 Pol IV-dependent regions discovered through the initial RNA-seq experiment (reported at the beginning of the Results section) in the polyA- RNA-seq

libraries. At 98% of the regions where expression was detected from *dcl234*, decreased expression in *dcl234 nrpd1* was also observed (Figure 2.4D). In addition, the expression of these 698 regions as determined by polyA+ RNA-seq was very low and decreased expression in *dcl234 nrpd1* was not observed (Figure 2.4D). This confirmed that P4RNAs are present in the polyA- RNA fraction and absent from the polyA+ RNA fraction.

Our previous studies showed that a partial loss-of-function mutation in a Pol II subunit gene compromised P4siRNA biogenesis at some RdDM loci. This raised the question of whether Pol II-dependent transcripts at RdDM loci are directly channeled to P4siRNA biogenesis or Pol II promotes P4siRNA biogenesis indirectly, such as by recruiting Pol IV (Zheng et al. 2009). The ability to distinguish Pol II-dependent transcripts and P4RNAs at RdDM loci allowed us to address this question. If Pol II-dependent transcripts were channeled to P4siRNA production, we would expect to detect dsRNAs from Pol II-dependent transcripts in *dcl234*. At P4siRNA loci, P4RNAs in polyA- RNA-seq were derived from two strands as expected (Figures 2.4E and 2S.7A). However, transcripts in polyA+ RNA-seq, presumably Pol II-dependent transcripts, appeared to be mainly derived from one strand (Figures 2.4F, 2S.7B, and 2S.8). We calculated the ratio of reads from the two strands in polyA+ RNA-seq at 1,639 P4siRNA loci, where Pol II transcription was detectable. Approximately 99% of polyA+ RNAs at these loci had a ratio of 9:1 or larger between reads derived from the two strands (Figure 2S.7C). This suggests that Pol II-transcribed RNAs were not converted to dsRNAs. We next examined the strand distribution of P4siRNAs at these loci. P4siRNAs were present

at some of the loci in *dcl234*, probably because DCL1 was able to produce P4siRNAs. The reads for P4siRNAs were derived from two strands, while the Pol II transcribed polyA+ RNAs were from one strand (Figure 2S.8). In *dcl234 nrpd1*, the P4siRNAs were depleted, suggesting that the P4siRNAs were derived from Pol IV. The fact that Pol II-dependent RNAs from loci that generate P4siRNAs are only from one strand (Figure 2S.8) and that no P4siRNAs are present in *dcl234 nrpd1* suggests that Pol II-dependent transcripts are not channeled to P4siRNA production.

The decreased CHH DNA methylation in *dcl234* is correlated to compromised Pol IV transcription

Our dsRNA-seq effort uncovered 22,990 regions producing P4RNAs, less than half of the 47,442 regions that produce P4siRNAs. Thus, we interrogated why P4RNAs were not detected from half of the P4siRNA-generating loci. We examined the 24,452 P4siRNA regions from which P4RNAs were not detected and found that 72% of the regions had low read coverage of less than 0.9 RPM in both *dcl234* and *dcl234 nrpd1* dsRNA-seq libraries, which made it impossible to make any comparisons between the two genotypes (Figure 2S.9A). Therefore, low levels of the P4RNAs were the major reason prohibiting their discovery.

We next asked whether the low levels of P4RNAs at these regions in *dcl234* were attributable to the fact that these regions have low Pol IV activity in WT. The output of Pol IV activity is P4siRNAs. We divided all regions producing P4siRNAs in WT into four quartiles according to the abundance of P4siRNAs. The percentage of P4RNAs

discovered was calculated for each quartile. As expected, with the decrease in P4siRNA abundance, the percentage of P4RNAs discovered also decreased. However, even in the first quartile that contained regions with the most abundant P4siRNAs in WT, still 30% of the regions lacked detectable P4RNAs in *dcl234* (Figure 2.5A). Therefore, our approaches failed to detect P4RNAs at some of the loci that generate abundant P4siRNAs and are thus predicted to also generate high levels of P4RNAs.

We next examined whether levels of P4RNAs in *dcl234* were correlated to levels of CHH DNA methylation. The CHH DNA methylation levels were examined separately for P4siRNA loci with or without P4RNAs detected. The average CHH DNA methylation levels at the two types of loci were similar in WT (Figure 2S.9C), but different in *dcl234*; the type without P4RNAs detected had much lower levels of CHH methylation than the type with P4RNAs detected (Figures 2.5B and 2S.9D). Therefore, it appeared that CHH methylation correlated with the production of P4RNAs.

We examined whether P4RNAs were affected differently by CHH methylation at DRM2- and CMT2-targeted sites, which will be referred to as D2 and C2 loci for simplicity. The P4siRNAs produced from these two types of loci will be referred to as D2 and C2 siRNAs. First, the relative abundance of D2 and C2 siRNAs was determined by sRNA-seq in WT. Although the number of C2 loci was larger than that of D2 loci, the total small RNA read number of C2 siRNAs was much smaller than that of D2 siRNAs no matter when total P4siRNAs or only 21nt, 22nt, 23nt, or 24nt P4siRNAs were separately considered (Figure 2S.9B). Next, we calculated the percentage of P4RNA discovery in *dcl234* at D2 and C2 loci separately. Although the average abundance of D2

siRNAs was higher than C2 siRNAs, P4RNAs were detected at 38% of D2 loci vs. 62% of C2 loci. The difference was even more obvious when D2 and C2 loci belonging to the lowest quartile of P4siRNA abundance were considered (Figure 2.5C). We observed a strong correlation between P4RNA discovery and CHH DNA methylation at D2 loci. When D2 sites were divided into four quartiles according to their CHH DNA methylation levels in *dcl234*, the percentage of D2 P4RNA discovery decreased with decreasing CHH DNA methylation (Figure 2.5D). Similarly, the abundance of P4RNAs at D2 sites, as revealed by dsRNA-seq in *dcl234*, also decreased with decreasing CHH DNA methylation (Figure 2S.10A). These trends were not found for C2 sites (Figures 2.5D and 2S.10A). The correlation between P4siRNA abundance and levels of CHH methylation was also examined in *dcl234* (Figures 2S.10B, C) and WT (Figures 2S.10D, E). The abundance of D2 siRNAs but not C2 siRNAs decreased with decreasing CHH DNA methylation. In summary, Pol IV transcription appeared to depend on CHH DNA methylation to a greater extent at D2 sites than at C2 sites.

Genetic requirements for P4siRNA biogenesis

Previous studies demonstrated that Pol IV, RDR2 and DCL3 are responsible for the biogenesis of P4siRNAs and that CHH DNA methylation and H3K9me2 affect P4siRNA accumulation. By utilizing published sRNA-seq, ChIP-seq and methylome data (Table 2.3) (Roudier et al. 2011; Deleris et al. 2012; Lee et al. 2012; Law et al. 2013; Stroud et al. 2013; Stroud et al. 2014), we further explored the genetic requirements for P4siRNA production.

The levels of P4siRNAs were first examined in WT and mutants in genes participating in P4siRNA biogenesis such as *DCL3*, *RDR2*, *NRPD1*, *SSH1*, *CLSY1*, and *DMS4*. D2, C2 and total P4siRNAs were equally affected in *dcl234*, *rdr2* and *nrpd1* (Figure 2.6A). In *clsy1*, *ssh1* and *dms4*, D2 siRNA levels were similarly decreased but not completely eliminated, and the reduction in P4siRNA abundance correlated with a reduction in CHH methylation in the three genotypes (Figures 2.6A and 2S.11A). At C2 loci, P4siRNA levels were decreased in *clsy1* and *ssh1* but increased in *dms4*, and these changes in P4siRNA levels were not accompanied by appreciable changes in CHH methylation (Figures 2.6A and 2S.11B). Therefore, a correlation between P4siRNA accumulation and CHH methylation is only true for D2 loci. Another conclusion is that all these genes, with the exception of *DMS4*, are required for P4siRNA biogenesis at both D2 and C2 loci.

The levels of P4siRNAs were also examined in mutants of genes participating in the RdDM pathway downstream of P4siRNA biogenesis, such as *DMS3*, *DRD1*, *RDM1*, *DRM2*, and *NRPE1*. Mutations in these genes all resulted in a near elimination of CHH methylation at D2 loci (Figure 2S.11A) but had almost no effect on CHH methylation at C2 loci (Figure 2S.11B). P4siRNA levels were also reduced in these mutants at both D2 and C2 loci, but D2 loci were affected to a greater extent; the remaining P4siRNAs were at 20% and 60% of wild-type levels for D2 and C2 loci, respectively (Figure 2.6B). These results were also consistent with a correlation between P4siRNA biogenesis and CHH methylation at D2 loci.

The levels of P4siRNAs were also examined in mutants of genes that confer DNA methylation, such as *DRM2*, *CMT3*, and *CMT2*, or H3K9me2 deposition, such as *SUVH4*, *5*, and *6*. In the *cmt2* mutant, in which CHH methylation was nearly eliminated at C2 loci but unaffected at D2 loci (Figures 2S.11A, B), P4siRNA accumulation was not affected at D2 loci or C2 loci (Figure 2.6C). In *drm1 drm2 cmt2 cmt3* (*drm12cmt23*) in which all non-CG methylation is lost and H3K9me2 cannot be maintained because of the loss of non-CG methylation (Stroud et al. 2014) (Figures 2S.11A, B), D2 siRNA levels were severely reduced but C2 siRNAs were only weakly affected (Figure 2.6C). In *suvh456* in which H3K9me2 is lost (Stroud et al. 2014) and CHH methylation at both D2 and C2 loci is partially reduced (Figures 2S.11A, B), D2 and C2 siRNAs were at 40% and 65% of the levels in WT, respectively (Figure 2.6C).

The above observations support a tight correlation between CHH methylation and P4siRNA abundance at D2 loci but only a weak correlation at C2 loci. To explore possible contributors to P4siRNA biogenesis at C2 loci, we examined the overlap between P4siRNA loci and repressive epigenetic marks H3K9me2 and H3K27me1 (Roudier et al. 2011; Deleris et al. 2012). P4siRNAs were found at 57% of H3K9me2 regions, 67% of H3K27me1 regions, and 75% of the regions harboring both H3K9me2 and H3K27me1, which may suggest that H3K27me1 and H3K9me2 work together in promoting P4siRNA biogenesis (Figure 2.6D). When D2 and C2 loci were separately examined for their overlap with H3K9me2 and H3K27me1, both marks were present at 92% of C2 loci but only 19% of D2 loci (Figure 2.6E). This is consistent with prior

knowledge that D2 loci are primarily on euchromatic arms while C2 loci are in pericentromeric heterochromatin (Zemach et al. 2013).

In summary, D2 and C2 siRNAs share a common biogenesis pathway involving Pol IV, RDR2 and DCL3, but Pol IV transcription at these loci is probably regulated differently by different epigenetic marks (Figure 2.7). Compared to C2 siRNAs, D2 siRNAs are highly abundant and are found at euchromatic regions harboring high levels of CHH methylation but low levels of H3K9me2 or H3K27me1 (Figures 2S.11C, D, E, F). D2 siRNAs and CHH methylation appear to be under tight feedback regulation – D2 siRNAs are required for the maintenance of CHH methylation and their biogenesis (probably at the level of Pol IV transcription) is promoted by CHH methylation. In contrast, C2 siRNAs are less abundant and are found at heterochromatic regions with high levels of repressive marks such as H3K9me2 or H3K27me1 (Figures 2S.11C, D, E, F). C2 siRNAs are not required to maintain CHH methylation and, and their biogenesis is less affected by the loss of CHH methylation, probably because H3K9me2 and H3K27me1 contribute to Pol IV transcription at these loci.

Discussion

Pol IV is thought to generate the precursors to endogenous siRNAs, which are central players in RdDM in plants. However, Pol IV-derived transcripts have not been detected before, probably owing to their short-lived nature and the transcription of RdDM loci by Pol II in a Pol IV loss of function mutant. In this study, we devised a strategy that enabled

the detection of tens of thousands of P4siRNA precursors that we refer to as P4RNAs. The analysis of these P4RNAs provided the following insights into P4siRNA biogenesis.

Specifically, key tenets of the current model of P4siRNA biogenesis have been confirmed. We showed for the first time that Pol IV indeed generates long noncoding RNAs, consistent with the presumed role of Pol IV in transcribing RdDM loci in the current model. Previously, failure to detect Pol IV transcription by a nuclear run-on assay led to the hypothesis that Pol IV in maize is likely a dysfunctional polymerase (Erhard et al. 2009). Our findings are in favor of Arabidopsis Pol IV, and maize Pol IV by inference, as a functional polymerase. The fact that long noncoding P4RNAs are from both DNA strands and are absent in an *rdr2* mutant is consistent with the model that P4siRNA precursors are generated by the concerted actions of Pol IV and RDR2. Our findings may also prompt a re-consideration of the current model. Previous biochemical studies show that RDR2 and Pol IV are in the same complex and, *in vitro*, RDR2 activity requires Pol IV but Pol IV activity does not require RDR2 (Haag et al. 2012). Based on these observations, the current model is that Pol IV transcribes P4siRNA loci and RDR2 converts nascent P4RNAs into dsRNAs (Matzke and Mosher 2014). Our findings not only agree with the notion that Pol IV and RDR2 act together, but also implicate an essential role of RDR2 for Pol IV transcription. If Pol IV activity does not require RDR2 *in vivo*, we expect to detect P4RNAs in *dcl234 rdr2*. However, detection of P4RNAs either by RT-PCR at specific loci or by RNA-seq at the genomic scale showed that *nrpd1* and *rdr2* mutations were equally defective in the production of these transcripts. This

suggests that RDR2 may be required for the recruitment of Pol IV to P4siRNA loci, the transcription activity of Pol IV, or the stability of P4RNAs *in vivo*.

Our findings also provide new insights into RdDM. We show that P4RNAs are non-polyadenylated and lack introns, and thus are different from Pol II-dependent transcripts. Using presence or absence of polyA as the distinguishing feature, we found that P4RNAs are derived from both DNA strands while the de-repression of RdDM loci results in Pol II transcription from a single strand. The single-stranded nature of Pol II-dependent transcripts from RdDM loci in *dcl234* also suggests that Pol II transcripts are not converted to dsRNAs for P4siRNA production. However, our previous studies revealed a reduction in P4siRNA levels from some RdDM loci in a partial loss-of-function Pol II mutant (Zheng et al. 2009). Together, these data imply that Pol II does not contribute to P4siRNA biogenesis by supplying P4siRNA precursors. Instead, Pol IV recruitment to chromatin was compromised in the pol II mutant (Zheng et al. 2009), suggesting that Pol II transcription might act to recruit Pol IV. However, we note that this study only examined loci that are already under surveillance by RdDM. We cannot rule out that Pol II-derived transcripts may be used directly in siRNA production when a naïve element is first introduced into a genome.

The lack of introns in P4RNAs also has implications. Several splicing related proteins were reported to affect both P4siRNA abundance and CHH methylation, although their effects are less prominent than that of Pol IV (Zhang et al. 2013). The absence of introns in P4RNAs suggests that these splicing factors promote P4siRNA

biogenesis either indirectly through their splicing functions on genes or directly through splicing-independent functions on P4RNAs.

A surprising finding was that the 5’ ends of P4RNAs bear a monophosphate. The 5’ end of a primary transcript is expected to bear a 5’ triphosphate, or a cap as in Pol II-derived transcripts. It is possible that the P4RNAs that we detected represent processed transcripts. Alternatively, Pol IV or RDR2 may use 5’monophosphate-containing RNAs as primers to initiate transcription. Regardless, the predominant form of P4RNAs *in vivo* is the form with a 5’ monophosphate. In this respect, P4RNAs resemble rRNAs, which are present *in vivo* as processed forms with a 5’ monophosphate (Dahlber et al. 1978; Unfried and Gruendler 1990). It is of note that the P4RNAs are also products of RDR2, therefore, the features of the 5’ and 3’ ends reflect co- or post-transcriptional events of Pol IV/RDR2.

A striking finding was the higher A/T composition and lower nucleosome occupancy of the flanking sequences of P4RNAs. This raises the possibility that high A/T composition and absence of nucleosomes promote the initiation and termination of Pol IV transcription. Nucleosome depletion in the 5’ flanking region is immediately followed by nucleosome enrichment 3’ to the transcription start site for P4RNAs. Such a pattern of nucleosome distribution is also found around the transcription start sites of protein-coding genes in diverse eukaryotes (Ammar et al. 2012), and thus represents a common feature of transcription initiation sites for Pol II and Pol IV.

CHH DNA methylation and H3K9me2 are repressive marks in the suppression of transposon expression and both are thought to promote P4siRNA biogenesis. Recent

studies have uncovered two parallel pathways of CHH methylation maintenance requiring two different DNA methyltransferases, DRM2 and CMT2. For the DRM2-targeted (D2) sites that are more dispersed within chromosomal arms, P4siRNAs and CHH methylation levels are high, and loss of CHH methylation impedes Pol IV transcription to result in reduced P4siRNA abundance. Therefore, CHH methylation and P4siRNA biogenesis are engaged in a positive feedback loop at D2 sites (Figure 2.7). CMT2-targeted (C2) sites are concentrated at pericentromeric regions, where other repressive marks such as H3K9me2 and H3K27me1 are prevalent (Roudier et al. 2011). At these sites, loss of CHH methylation has a minimal effect on Pol IV transcription as compared to D2 sites, and little impact on P4siRNA abundance (Figure 2.7). While it was found that H3K9me2 promotes P4siRNA accumulation at C2 sites ((Stroud et al. 2014) and this study), C2 siRNAs are only moderately affected in the *suvh456* mutant that lacks H3K9me2 or in *drm12cmt23* that lacks both H3K9me2 and CHH methylation (Figure 2.6C). We found that both H3K9me2 and H3K27me1 are highly prevalent at C2 loci. Thus, our findings implicate a role of H3K27me1 in P4siRNA biogenesis at C2 loci (Figure 2.7).

Materials and Methods

Plant materials

All tissues used in this study are from unopened flower buds and all Arabidopsis strains are in the Columbia ecotype. The *dcl2-1 dcl3-1 dcl4-2* (*dcl234*), *nrpd1-3* (*nrpd1*) and *rdr2-1* (*rdr2*) lines were previously described (Xie et al. 2004; Onodera et al. 2005;

Henderson et al. 2006). The quadruple mutants *dcl234 nrpd1* and *dcl234 rdr2* were obtained by crossing of *dcl234* with *nrpd1* and *rdr2*.

RNA isolation, digestion, and RT-PCR

Total RNAs were extracted from unopened flower buds with TRIzol (Invitrogen, 15596-018) and treated with DNase I (Roche, 04716728001). cDNA was synthesized using random primers with RevertAid Reverse Transcriptase (Fermentas EP0442). To determine the strandedness of the transcripts, reverse transcription was performed with gene-specific primers from each of the two strands. Sequences of primers are in Table 2.4.

To determine the nature of the 5' ends of P4RNAs, 5μg total RNAs from *dcl234* were divided into each of four tubes and were treated as follows. First, the RNAs were incubated at 37°C for 2h with or without enzymes: tube 1 and tube 2 with buffer only; tube 3 with Tobacco Acid Pyrophosphatase (TAP, Epicenter T19250); and tube 4 with T4 Polynucleotide Kinase (PNK, NEB, M0201S). After phenol-chloroform extraction and ethanol precipitation, RNAs in tube 1 were incubated at 30°C for 1h with buffer only, while RNAs in the other three tubes were incubated with Terminator Exonuclease (Epicenter, TER51020) at 30°C for 1h. The RNAs were extracted with phenol-chloroform and precipitated with ethanol before being subjected to RT-PCR.

Construction and sequencing of RNA-seq, RNA-seq-DSN, dsRNA-seq and sRNA-seq libraries

Unopened flower buds from *dcl234* and *dcl234 nrpd1* were collected and were used for RNA extraction using Trizol (Invitrogen, 15596-018). Briefly, 10μg of DNA-free RNAs were subjected to rRNA removal using a Ribomius kit (Invitrogen, A10838-08). For dsRNA-seq libraries, RNase One (Promega, M4261) was used to digest single-stranded RNAs. The treated RNAs were fragmented using Fragmentation Reagents (Ambion, AM8740). T4 Polynucleotide Kinase (NEB, M0201S) was used to phosphorylate the 5' ends as well as to remove the 3' phosphate groups of the RNA fragments. The treated RNAs were resolved in a 15% denaturing polyacrylamide gel and 15-100 nt RNAs were excised and purified. These RNAs were used to construct the RNA-seq and dsRNA-seq libraries using the True-seq small RNA preparation kit (Illumina, RS-200-0012). For some samples, the RNAs were further treated with Duplex-Specific Nuclease (DSN, Evrogen, EA001) to enrich for low abundance transcripts. The RNA-seq libraries treated with DSN are referred to as RNA-seq-DSN libraries. The sRNA-seq libraries were also constructed using the True-seq small RNA preparation kit. The libraries were sequenced through Illumina Hiseq2000 and the data were deposited at NCBI under the accession number GSE57215. All libraries built in this study are listed in Table 2.5, which contains information on the number of biological replicates for each library type and genotype.

Processing and mapping of RNA-seq, RNA-seq-DSN and dsRNA-seq reads

Raw reads were first collapsed into a set of non-redundant reads. All of the non-redundant reads were initially mapped to the *Arabidopsis* TAIR10 reference genome using the short-read alignment tool (BWA) allowing no mismatches (Henderson et al.

2006; Li and Durbin 2009). Unaligned reads were processed further by sequentially trimming off nucleotides at the 3' end with any match to the 5' end of the adapter sequence allowing for 0, 1, 2, and 3 mismatches if the 3' end nucleotides match to less than nine nucleotides, 10-19 nucleotides, 20-29 nucleotides, and 30-33 nucleotides, respectively, of the adapter sequence. The longest allowed match to the adapter sequence is set arbitrarily at 33 nucleotides to maintain the shortest trimmed reads at 18 nucleotides. Adapter-trimmed reads were mapped to the TAIR10 genome allowing no mismatches. All mapped reads (untrimmed and adaptor-trimmed) were combined for further downstream analysis.

To determine the regions that harbor P4RNAs, the genome was tiled into 500 bp bins and the reads whose 5' ends fall within a bin were considered as belonging to this bin. The numbers of reads were counted for each bin for both *dcl234* and *dcl234 nrpd1* and compared between the two genotypes. The fold change and p-value were calculated using edgeR for dsRNA-seq and RNA-seq (Robinson et al. 2010). The Poisson distribution is used to calculate the p-value for RNA-seq-DSN libraries (Marioni et al. 2008). The regions with p-value < 0.01 and four-fold reduction in read counts in *dcl234 nrpd1* relative to *dcl234* were considered as regions that generate P4RNAs.

Processing and mapping of sRNA-seq reads

The reads in sRNA-seq libraries were first trimmed to remove adapters. Each read was queried for the presence of the first 9 nt sequence (TGGAATTCT) of the 5' end adapter. If found, the query sequence plus the flanking 3' end sequence is removed from the read.

Adaptor-free reads between 18 nt and 42 nt in length were mapped to the TAIR10 genome. To calculate and compare small RNA abundance in different genotypes, the genome was tiled into 500 bp windows and reads whose 5' end nucleotides fall within a window were assigned to the window. To identify differentially expressed small RNAs, edgeR was applied to calculate the fold change and p-value. The windows with p-value < 0.01 and four-fold reduction in read counts in *nrpd1* relative to WT were considered as regions that generate P4siRNAs.

Assembly of Pol IV/RDR2-dependent transcripts

In-house R scripts were employed to assemble P4RNAs. The first step was to collect and combine all the reads located at P4RNA regions from the three replicates of dsRNA-seq libraries from *dcl234*. Then neighboring reads no more than 60 nt apart were joined together to form transcripts. The transcripts that passed the following three filters were retained. First, the transcripts must be longer than 60 nt. Second, the normalized read count from the combined three libraries of *dcl234* was above 1RPM. Third, the levels of the transcripts in *dcl234* were at least four fold higher than those in *dcl234 nrpd1*. Finally, the transcripts were overlapped with P4siRNA loci to filter out the transcripts without corresponding P4siRNA expression.

Determination of A/T composition of various genomic regions

Exons and genes were according to TAIR10 annotation; P4RNAs were determined in this study as described above. Only P4RNAs longer than 200 bp were included in this analysis. The start and end sites of P4RNA regions were arbitrarily defined as the 5' and 3' ends of P4RNAs on the Watson strand. Within each category (P4RNAs, exons, or genes), sequences were aligned at the start site of transcription (for P4RNAs and genes) or the beginning of exons, or at the end site of transcription/end of exons. Up to 1 kb of sequences flanking these sites were interrogated. The numbers of A, T, C, or G at each position for all the sequences in each category were counted. The A/T composition was calculated as the proportion of A and T nucleotides in the total.

Determination of nucleosome occupancy at various genomic regions

Nucleosome occupancy was examined at the same exons, genes and P4RNAs interrogated for their A/T composition (described above). The positions of nucleosomes in the genome were obtained by analysis of the dataset (Chodavarapu et al. 2010) using the nucleosome-calling program NOrMAL (Polishko et al. 2012). The sequences of each category (P4RNAs, exons, or genes) were aligned at the start site of transcription (for P4RNAs and genes) or the beginning of exons, or at the end site of transcription/end of exons. For each position, the percentage of sequences with nucleosomes in total sequences was calculated.

A naïve method of identifying spliced reads

To identify reads that represent potential splicing events, the first step was to filter out reads that mapped perfectly to the genome. The unmapped reads were mapped to the genome again using blastall with a minimum mapped length of 15 nt (Zhang and Madden 1997). The reads were kept if both the beginning 15 nt and the end 15 nt of the reads mapped perfectly to the genome. In addition, the mapping positions on the genome of these reads were examined. If both the beginning and the end of the reads were mapped to the same strand within a distance of 1000 nt, the reads were kept as representing a potential splicing event.

The definition of D2 and C2 loci

Differentially methylated regions (DMRs) in wild type vs. *drm2* and wild type vs. *cmt2*, named D2 and C2 DMRs, respectively, were derived from published methylome datasets ((Stroud et al. 2013) with accession numbers listed in Table 2.3. P4siRNA regions overlapping with D2 and C2 DMRs were referred as D2 and C2 siRNA loci, respectively.

The overlap between P4siRNAs with H3K27me1 and H3K9me2

The regions with H3K9me2 modifications were defined through analysis of published ChIP-chip dataset (Roudier et al. 2011; Deleris et al. 2012) using BLOC (Pauler et al. 2009). The regions with H3K27me1 modifications were obtained in a published ChIP-chip dataset (Roudier et al. 2011; Deleris et al. 2012). To calculate the P4siRNA regions with H3K27me1 and H3K9me2 modifications and the regions of H3K27me1 and H3K9me2 with P4siRNAs, the regions of H3K27me1 and H3K9me2 were divided into

500bp arbitrary windows, and the overlap between these windows and those of P4RNAs was determined. Then the percentage of the overlap in total windows was determined.

Data access

The genome-wide datasets generated in this study are available at the NCBI Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) under the accession number GSE57215.

Figures

Figure 2. 1 Genome-wide discovery of P4RNAs as P4siRNA precursors.

A-B, Genome-browser views of small RNA reads and P4RNA reads at two representative P4siRNA loci. The read counts (in rpm – reads per million) include reads from both strands. The top two, middle two, and bottom two rows represent reads from dsRNA-seq, sRNA-seq, and RNA-seq, respectively. In A, P4RNAs were detected by both dsRNA-seq and RNA-seq. In B, P4RNAs were only detected by dsRNA-seq. C, Venn diagram showing the overlap of P4RNA regions discovered through dsRNA-seq or RNA-seq with P4siRNA regions discovered through sRNA-seq. Note that dsRNA-seq and RNA-seq were conducted with *dcl234* and *dcl234 nrpd1*, and sRNA-seq was conducted with WT and *nrpd1*. D, Random-primed RT-PCR analysis of P4RNAs discovered through RNA-seq on RNA samples from *dcl234* and *dcl234 nrpd1*. Genomic DNA was included as the positive control for the PCR. -RT: reverse transcriptase was omitted from the reverse transcription reactions. "-RT" and H_2O (no RNAs in the reactions) served as negative controls. The genomic locations of the loci can be found in Table 2.1.

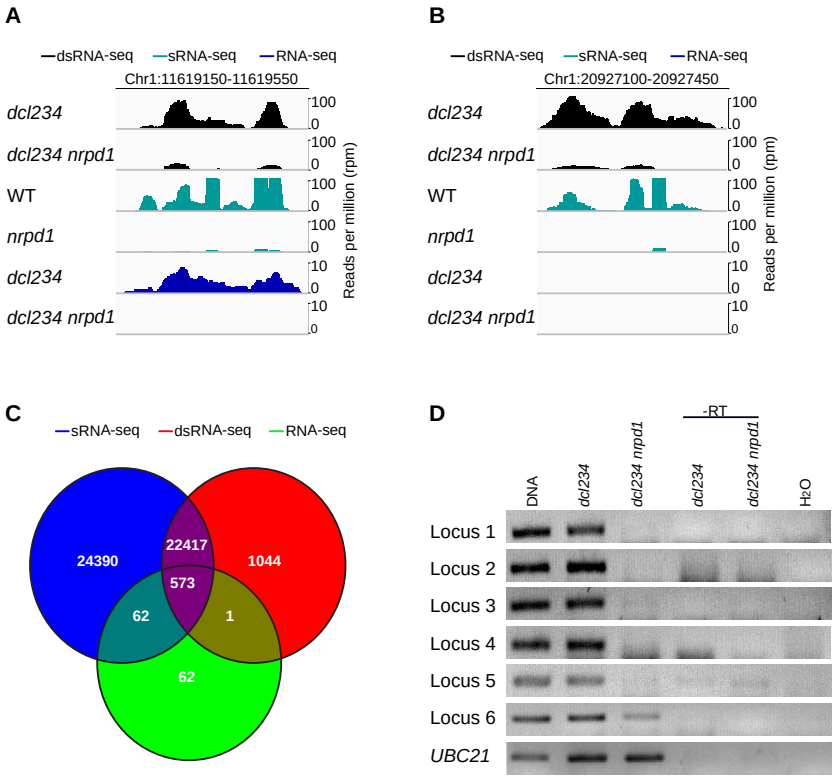

A
dsRNA-seq
sRNA-seq
RNA-seq
Chr1:11619150-11619550
dcl234
dcl234 nrpd1
WT
nrpd1
dcl234
dcl234 nrpd1
100
0
10
Reads per million (rpm)
B
dsRNA-seq
sRNA-seq
RNA-seq
Chr1:20927100-20927450
C
sRNA-seq
dsRNA-seq
RNA-seq
24390
22417
1044
573
62
1
62
D
-RT
DNA
dcl234
dcl234 nrpd1
dcl234
dcl234 nrpd1
H2O
Locus 1
Locus 2
Locus 3
Locus 4
Locus 5
Locus 6
UBC21

Figure 2. 2 RDR2 has a similar effect as Pol IV on the abundance of P4RNAs.

A, Detection of P4RNAs by RT-PCR. Random-primed RT-PCR was performed on *dcl234*, *dcl234 nrpd1* and *dcl234 rdr2* to detect P4RNAs from five loci (Table 2.1). PCR with genomic DNA and H_2O (no RNAs in the reactions) were included as positive and negative controls, respectively. -RT, reverse transcription was performed in the absence of reverse transcriptase. *CBP20*, a genic transcript, was included as a loading control. B, DSN normalization moderately enriched the coverage of reads at P4RNA loci by RNA-seq. The total numbers of normalized reads at 47,442 P4siRNA loci from one replicate of *dcl234* RNA-seq-DSN and three replicates of *dcl234* RNA-seq are shown. C, Venn diagram showing the overlap between regions with Pol IV-dependent transcripts and regions with RDR2-dependent transcripts as determined by RNA-seq-DSN of *dcl234*, *dcl234 nrpd1* and *dcl234 rdr2*. D, Abundance of Pol IV- and RDR2-dependent RNAs at the 850 Pol IV- and RDR2-dependent loci in (C). The total numbers of normalized reads at these loci in RNA-seq-DSN are shown.

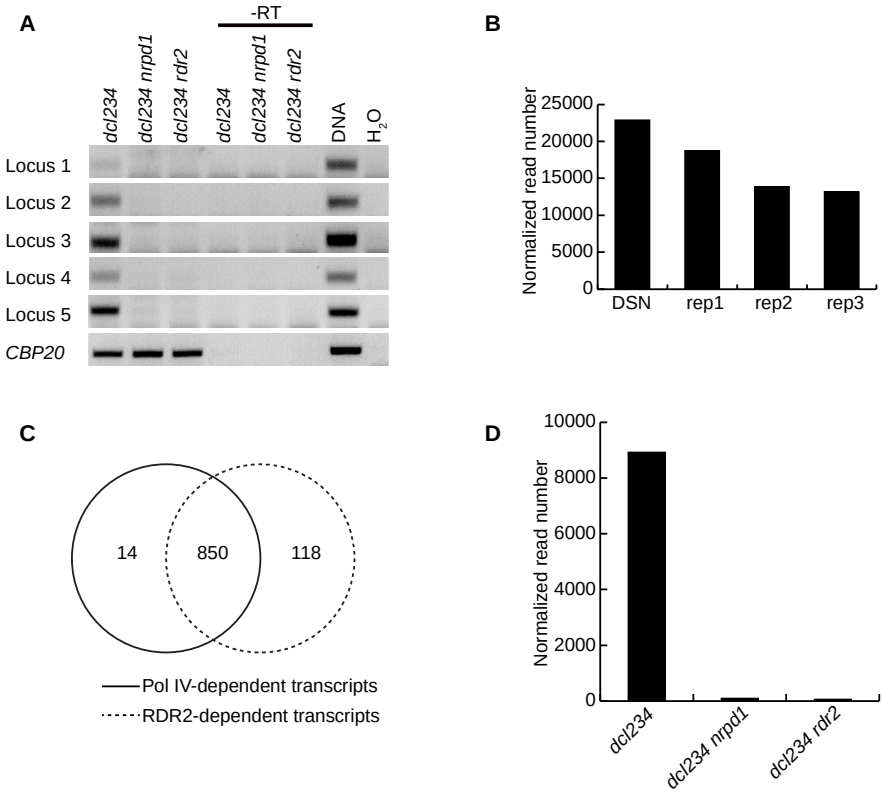

A
-RT
dcl234
dcl234 nrpd1
dcl234 rdr2
dcl234
dcl234 nrpd1
dcl234 rdr2
DNA
H2O
Locus 1
Locus 2
Locus 3
Locus 4
Locus 5
CBP20
B
Normalized read number
25000
20000
15000
10000
5000
0
DSN
rep1
rep2
rep3
C
14
850
118
Pol IV-dependent transcripts
RDR2-dependent transcripts
D
Normalized read number
10000
8000
6000
4000
2000
0
dcl234
dcl234 nrpd1
dcl234 rdr2

Figure 2. 3 Genomic features of P4RNAs and surrounding regions.

A/T composition (A-F) and nucleosome occupancy (G-H) were examined at P4RNAs, exons and genes. Exons and genes were according to TAIR10 annotation. Position 0 refers to the start site of transcription (for P4RNAs and genes) or the beginning of exons (in A, C, E, G), or the end site of transcription/end of exons (in B, D, F, H). Nucleotide positions upstream and downstream of position 0 are represented by negative and positive numbers, respectively. Sequences were aligned at position 0 and the proportion of A/T nucleotides at each position is shown in A-F. A-B, The A/T composition near the P4RNA start sites (A) or end sites (B). C-D, The A/T composition of exons and flanking regions. E-F, The A/T composition near protein-coding gene transcription start sites (E) or termination sites (F). In A-F, the insets display close-up views near position 0. G-H, Average nucleosome occupancy near the start sites (G) or end sites (H) of P4RNAs, exons and genes. The nucleosome positions were derived from published data (Chodavarapu et al. 2010).

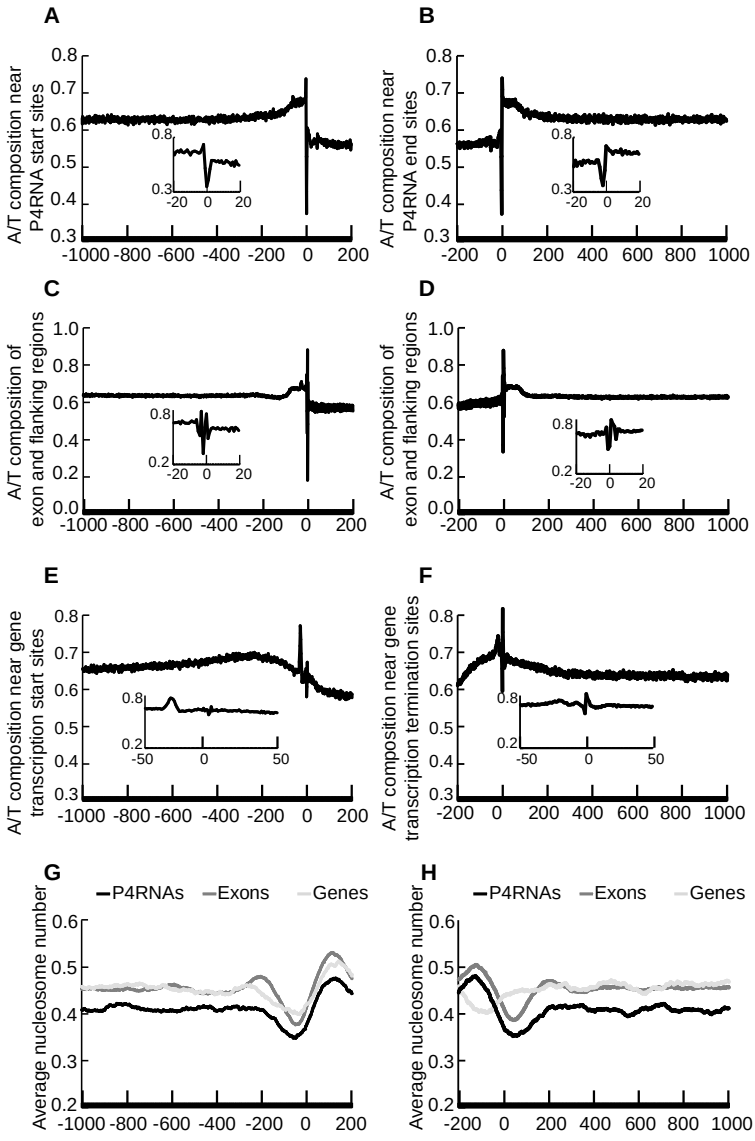

A
B
C
D
E
F
G
H
A/T composition near P4RNA start sites
A/T composition near P4RNA end sites
A/T composition of exon and flanking regions
A/T composition near gene transcription start sites
A/T composition near gene transcription termination sites
Average nucleosome number
P4RNAs
Exons
Genes

Figure 2. 4 Features of P4RNAs. A, Determination of the 5' end structure of P4RNAs.

Total RNAs from *dcl234* were treated (+) or not (-) with various enzymes and subjected to random-primed RT-PCR to detect specific P4RNAs (loci 1-6; Table 2.1) with P4RNA-specific primers. PNK, polynucleotide kinase; TAP, Tobacco Acid Pyrophosphatase; Ter, Terminator Exonuclease. PCR with genomic DNA and H_2O (no RNAs in the reactions) were included as positive and negative controls, respectively. Transcripts from two genes, *CBP20* and *UBC21*, were also detected by RT-PCR as controls. As expected, the levels of these RNAs were only reduced by digestion with both TAP and Ter. B, Determination of the 3' end structure of P4RNAs. Random-primed RT-PCR was performed on total RNAs from *dcl234* and *dcl234 nrpd1*, and polyA-enriched and polyA-depleted RNAs from *dcl234* to detect specific P4RNAs. -RT, reverse transcription was performed in the absence of reverse transcriptase. *CBP20* served as a positive control for polyA+ RNAs. The *CBP20* RT-PCR products in the polyA- fraction probably reflected degradation intermediates. C, Abundance of reads at 1639 P4RNA regions with detectable transcripts in polyA+ RNA-seq. Two replicates of RNA-seq were conducted and the sum of the numbers of normalized reads is shown. D, Abundance of transcripts at 698 P4RNA regions discovered through the initial RNA-seq, as determined by RNA-seq from polyA+ and polyA- RNAs. The reduction in transcript abundance in *dcl234 nrpd1* was only observed in polyA- RNAs, indicating that P4RNAs lack polyA tails. E, A genome-browser view of reads at a P4siRNA locus on chromosome 3 from sRNA-seq and polyA-RNA-seq. Read abundance is shown for both the Watson (top) and Crick (bottom)

strands. F, A genome-browser view of reads at a P4siRNA locus on chromosome 2 from sRNA-seq and polyA+ RNA-seq. Read abundance is shown for both the Watson (top) and Crick (bottom) strands.

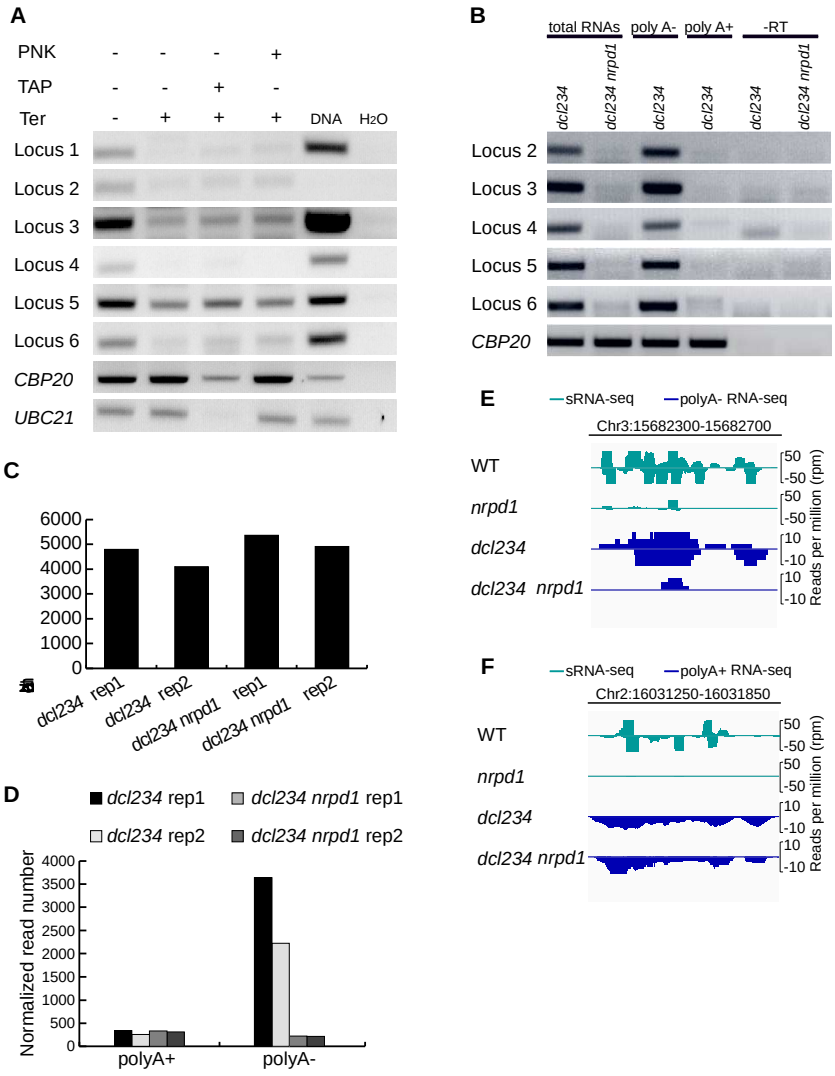

A
PNK - - - +
TAP - - + -
Ter - + + + DNA H2O
Locus 1
Locus 2
Locus 3
Locus 4
Locus 5
Locus 6
CBP20
UBC21
B
total RNAs
poly A-
poly A+
-RT
dcl234
dcl234 nrpd1
dcl234
dcl234
dcl234
dcl234 nrpd1
Locus 2
Locus 3
Locus 4
Locus 5
Locus 6
CBP20
C
6000
5000
4000
3000
2000
1000
0
dcl234 rep1
dcl234 rep2
dcl234 nrpd1 rep1
dcl234 nrpd1 rep2
D
dcl234 rep1
dcl234 nrpd1 rep1
dcl234 rep2
dcl234 nrpd1 rep2
Normalized read number
4000
3500
3000
2500
2000
1500
1000
500
0
polyA+
polyA-
E
sRNA-seq
polyA- RNA-seq
Chr3:15682300-15682700
WT
nrpd1
dcl234
dcl234 nrpd1
50
-50
50
-50
10
-10
10
-10
Reads per million (rpm)
F
sRNA-seq
polyA+ RNA-seq
Chr2:16031250-16031850
WT
nrpd1
dcl234
dcl234 nrpd1
50
-50
50
-50
10
-10
10
-10
Reads per million (rpm)

Figure 2.5 Decreased CHH DNA methylation in *dcl234* compromises Pol IV transcription.

A-C, P4siRNA loci are divided into four quartiles according to P4siRNA abundance in WT, with the first quartile containing loci with the highest levels of P4siRNAs. A, The percentage of P4siRNA loci with and without P4RNAs detected in our dsRNA-seq for the four quartiles. B, The levels of CHH methylation decrease in *dcl234* compared to WT for the four quartiles of P4siRNA loci with and without precursors detected. The decrease in CHH methylation was calculated using published methylome data (Stroud et al. 2013). C, The percentage of P4RNAs detected at D2 and C2 siRNA loci for the four quartiles. 13,479 D2, 19,039 C2, and 47,742 total P4siRNA loci were included in the analysis. D, Correlation between P4RNA discovery and levels of CHH methylation at the siRNA loci. D2 and C2 siRNA loci are divided into four quartiles according to their CHH methylation levels in *dcl234*. The percentage of loci with P4RNAs detected in each quartile is shown. As CHH methylation levels decreases, the success rate of P4RNA discovery also decreases in total P4siRNA loci. In terms of P4RNA discovery, D2 loci are more sensitive to levels of CHH methylation.

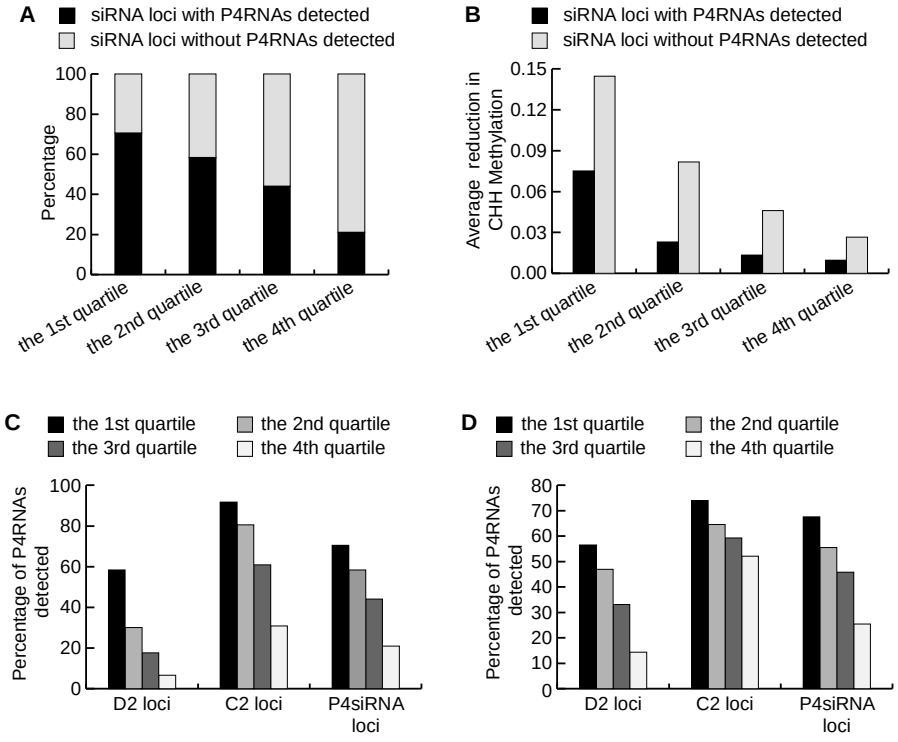

A
siRNA loci with P4RNAs detected
siRNA loci without P4RNAs detected
Percentage
100
80
60
40
20
0
the 1st quartile
the 2nd quartile
the 3rd quartile
the 4th quartile
B
siRNA loci with P4RNAs detected
siRNA loci without P4RNAs detected
Average reduction in CHH Methylation
0.15
0.12
0.09
0.06
0.03
0.00
the 1st quartile
the 2nd quartile
the 3rd quartile
the 4th quartile
C
the 1st quartile
the 2nd quartile
the 3rd quartile
the 4th quartile
Percentage of P4RNAs detected
100
80
60
40
20
0
D2 loci
C2 loci
P4siRNA loci
D
the 1st quartile
the 2nd quartile
the 3rd quartile
the 4th quartile
Percentage of P4RNAs detected
80
70
60
50
40
30
20
10
0
D2 loci
C2 loci
P4siRNA loci

Figure 2. 5~~6~~ RdDM genes, epigenetic marks and P4siRNA biogenesis.

A-C, Effects of mutations in various CHH methylation pathway genes on P4siRNA biogenesis. The relative abundance of D2, C2 and total P4siRNAs in various mutants compared to WT is shown. The analysis was performed with published sRNA-seq data (Lee et al. 2012; Law et al. 2013; Stroud et al. 2014). For each genotype, reads corresponding to P4siRNA loci were normalized against small RNAs from non-P4siRNA loci. P4siRNA loci were defined as those showing differentially expressed siRNAs between WT and *nrpd1* (see Methods). A total of 47,742 total P4siRNA loci, 13,479 D2, and 19,039 C2 loci were used in the analysis. A, Relative siRNA abundance in mutants in genes known to act in P4siRNA biogenesis. B, Relative siRNA abundance in mutants in genes known to act downstream of P4siRNAs in RdDM. C, Relative siRNA abundance in mutants in genes that confer CHH DNA methylation or histone H3K9 methylation. *nrpd1* is included in B and C for comparison. D-E, Overlap between P4siRNA loci and the epigenetic marks H3K9me2 or H3K27me1. Published ChIP-chip data were used to define regions with H3K9me2 or H3K27me1 (Roudier et al. 2011; Deleris et al. 2012). D, Regions with H3K9me2, H3K27me1, or both H3K9me2 and H3K27me1 were divided into 500 bp windows. The numbers of windows where P4siRNAs were present or not were counted, and the percentage of total windows is shown. E, The percentage of P4siRNA loci with H3K9me2, H3K27me1, or both. The numbers of D2, C2 and total P4siRNA loci with H3K9me2, H3K27me1 or both marks were determined, and the percentage of these total loci is shown.

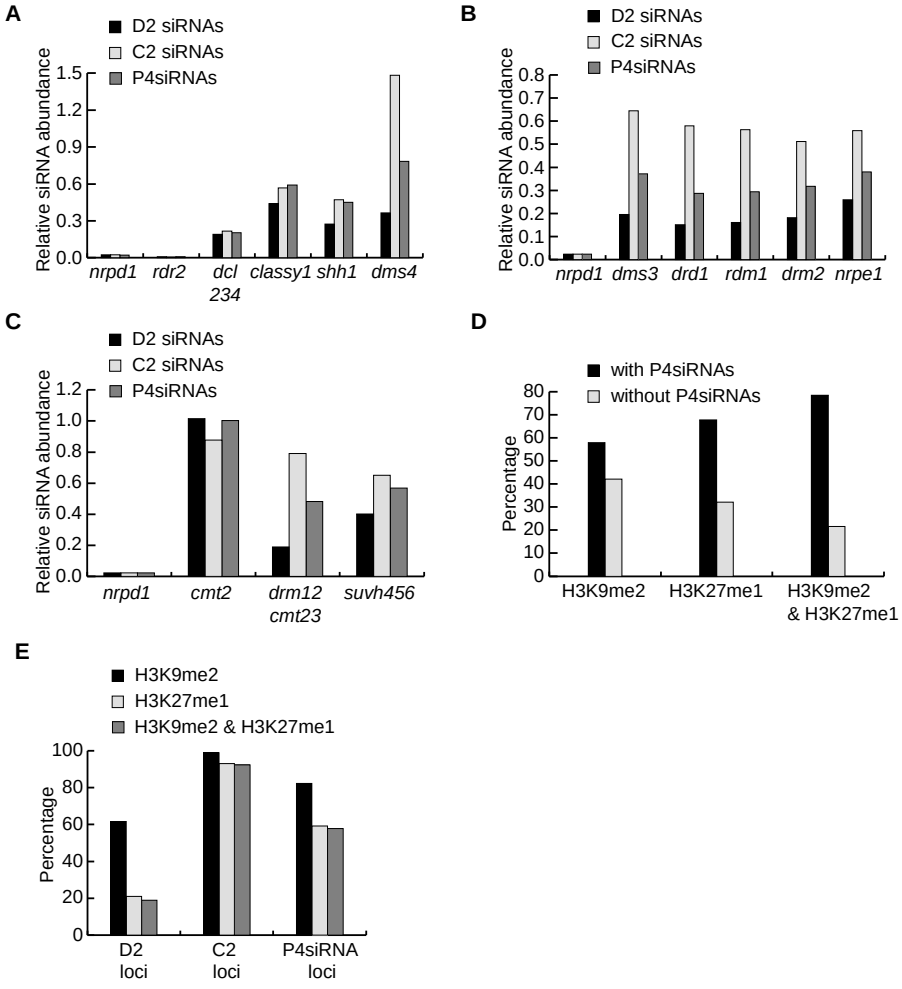

A
D2 siRNAs
C2 siRNAs
P4siRNAs
Relative siRNA abundance
1.5
1.2
0.9
0.6
0.3
0.0
nrpd1
rdr2
dcl 234
classy1
shh1
dms4
B
D2 siRNAs
C2 siRNAs
P4siRNAs
Relative siRNA abundance
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
nrpd1
dms3
drd1
rdm1
drm2
nrpe1
C
D2 siRNAs
C2 siRNAs
P4siRNAs
Relative siRNA abundance
1.2
1.0
0.8
0.6
0.4
0.2
0.0
nrpd1
cmt2
drm12 cmt23
suvh456
D
with P4siRNAs
without P4siRNAs
Percentage
80
70
60
50
40
30
20
10
0
H3K9me2
H3K27me1
H3K9me2 & H3K27me1
E
H3K9me2
H3K27me1
H3K9me2 & H3K27me1
Percentage
100
80
60
40
20
0
D2 loci
C2 loci
P4siRNA loci

Figure 2.7 Models on the feedback regulation between Pol IV transcription and epigenetic marks at D2 and C2 loci.

At both D2 and C2 loci, P4siRNA biogenesis requires Pol IV, RDR2, and DCL3. At D2 loci with high levels of methylated CHH and relatively low levels of H3K9me2 or H3K27me1, P4siRNAs and CHH methylation are in a tight feedback loop in which P4siRNAs guide CHH methylation and CHH methylation in turn promotes siRNA biogenesis, probably by recruiting Pol IV. At C2 loci with relatively low levels of methylated CHH and extensive overlap with H3K9me2 or H3K27me1, P4siRNA biogenesis is only moderately affected by the absence of CHH methylation (in *drm12cmt23* and *cmt2*) or H3K9me2 (in *suvh456*). The high percentage of C2 siRNA loci containing both H3K9me2 and H3K27me1 suggests that both epigenetic marks may contribute to Pol IV recruitment at C2 loci.

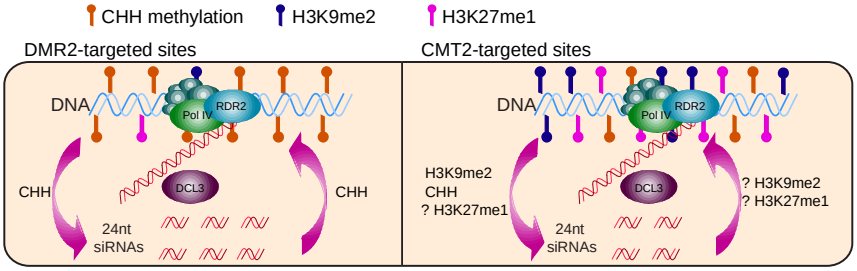

CHH methylation
H3K9me2
H3K27me1
DMR2-targeted sites
CMT2-targeted sites
DNA
Pol IV
RDR2
DCL3
CHH
CHH
24nt
siRNAs
DNA
Pol IV
RDR2
H3K9me2
CHH
? H3K27me1
DCL3
? H3K9me2
? H3K27me1
24nt
siRNAs

Figure 2S. 1 Genome-browser views of P4RNA and small RNA reads at two P4siRNA loci on Chromosome 1.

The two loci are the same as the ones shown in Figure 1A and 1B, except that three biological replicates (rep) are shown separately here. Note that reads from the two strands are not separately displayed.

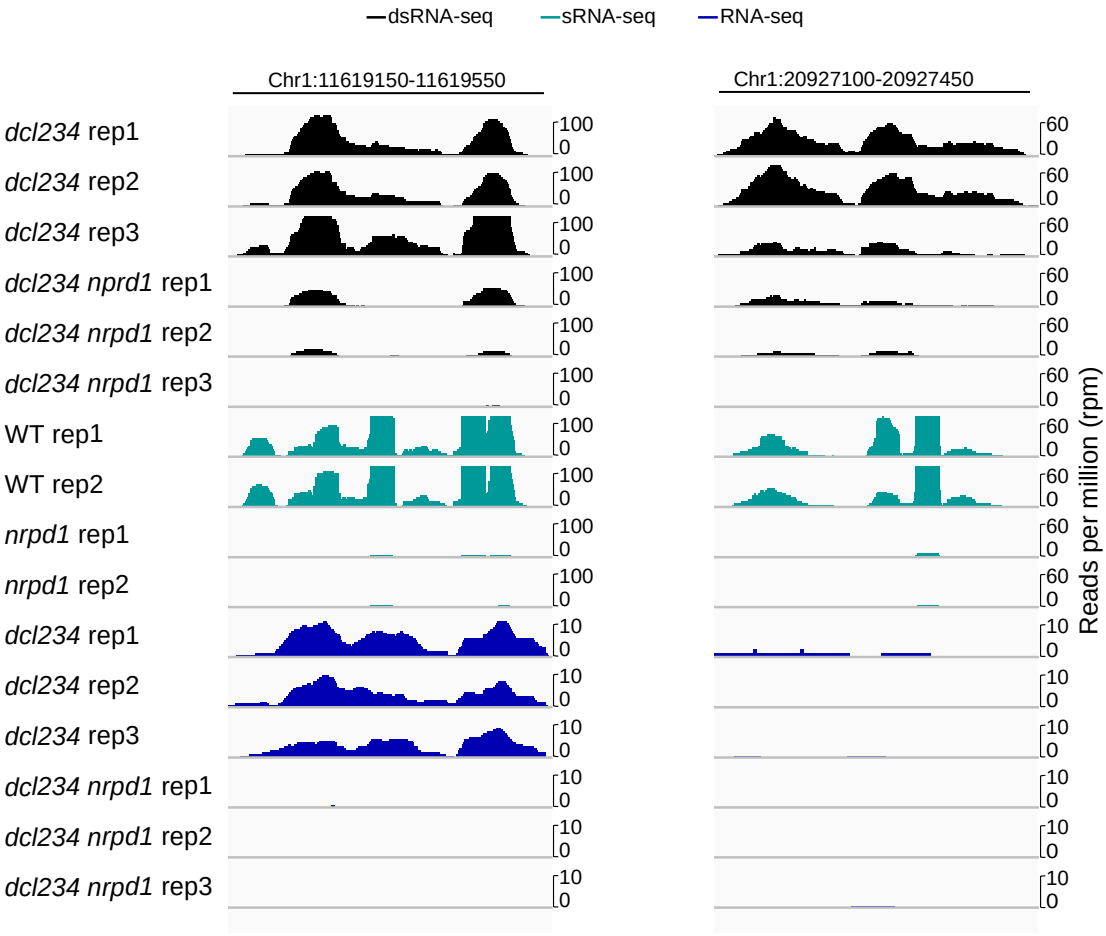


Figure 2S. 2 Detection of P4RNAs.

A, Random-primed RT-PCR to detect P4RNAs at 16 individual loci in *dcl234* and *dcl234 nrpd1*. Genomic DNA and H_2O (no RNAs in the reactions) were included as positive and negative controls, respectively. -RT, reverse transcription was conducted in the absence of reverse transcriptase. B, The percentage of reads that map to genes, intergenic regions and P4siRNA loci in RNA-seq and dsRNA-seq. Three biological replicates (rep) are shown

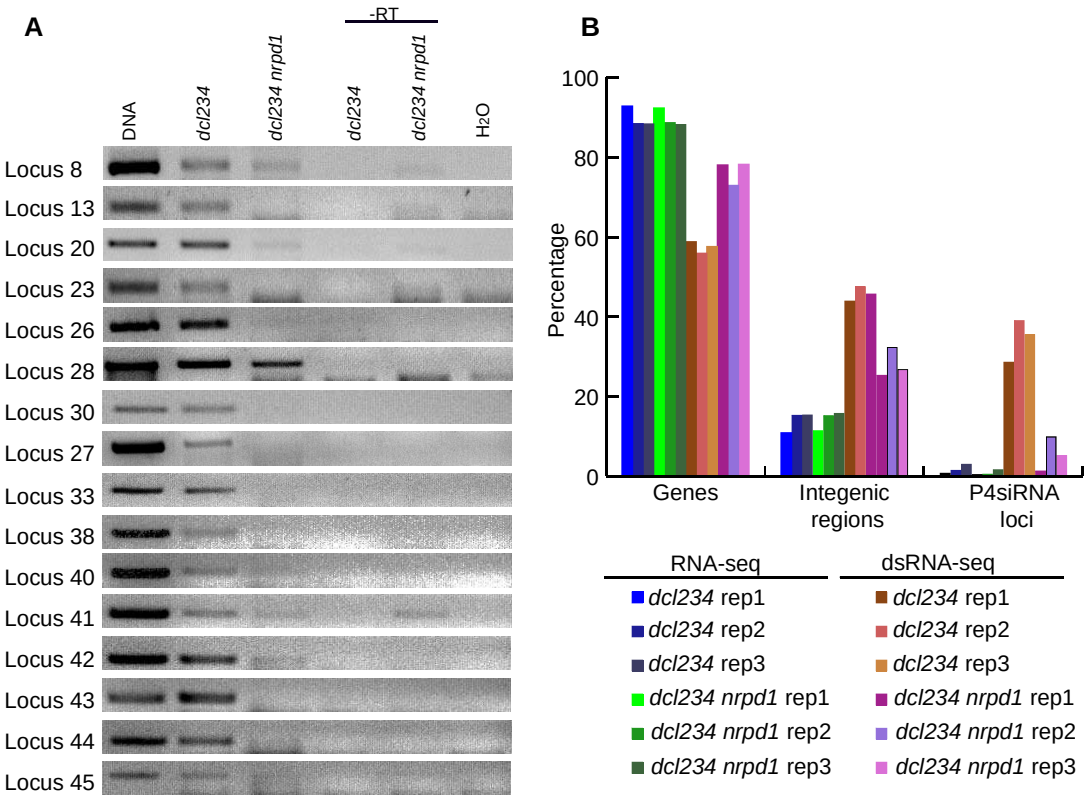


Figure 2S. 3 P4RNAs are derived from both DNA strands.

RT-PCR was performed with random primers or strand-specific primers for reverse transcription (RT) and sequence-specific primers for PCR to detect P4RNAs. The nature of the RT primers is indicated below the gel images. The Watson strand refers to the reference strand in TAIR10 annotation; the Crick strand refers to the reverse complementary strand of the reference. -RT, reverse transcription was performed in the absence of reverse transcriptase.

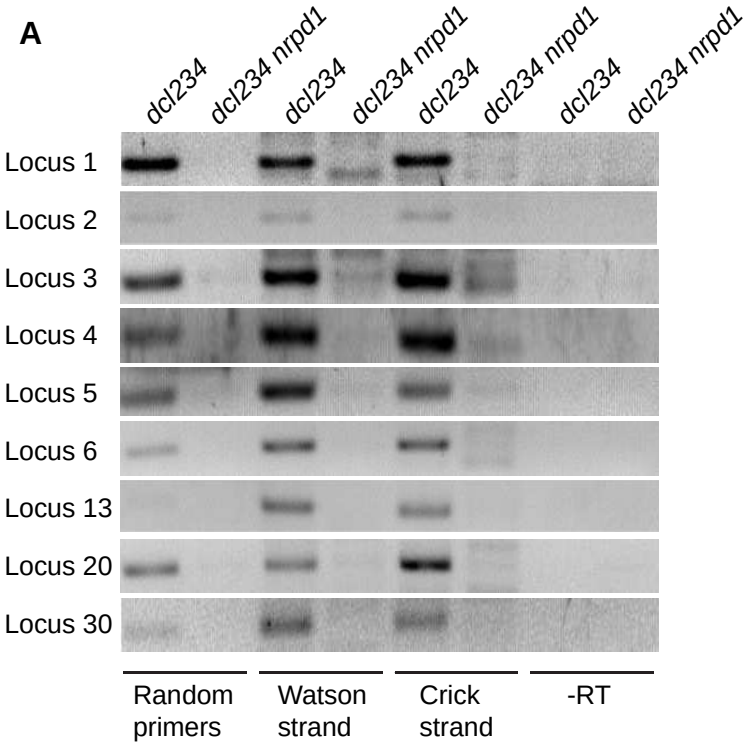


Figure 2S. 4 Size distribution of P4RNAs.

The number of P4RNAs in different size ranges (in nucleotide) is shown.

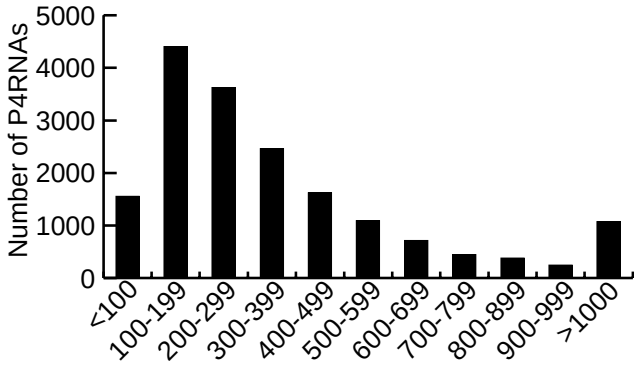


Figure 2S. 5 Relationships among P4RNAs, P4siRNAs, and CHH DNA methylation.

A, The percentage of P4RNA regions that overlap with genes, transposons, repeats, and intergenic regions. B, The presence of P4siRNAs at loci dependent on DRM2 or CMT2 for CHH methylation. The percentage of DRM2- or CMT2-dependent loci with P4siRNAs is shown. DRM2- and CMT2-dependent loci were defined as differentially methylated CHH regions (DMRs) in *drm2* and *cmt2*, respectively, relative to wild type. C, Venn diagram showing the overlap among DMRs dependent on Pol IV, DRM2, or CMT2.

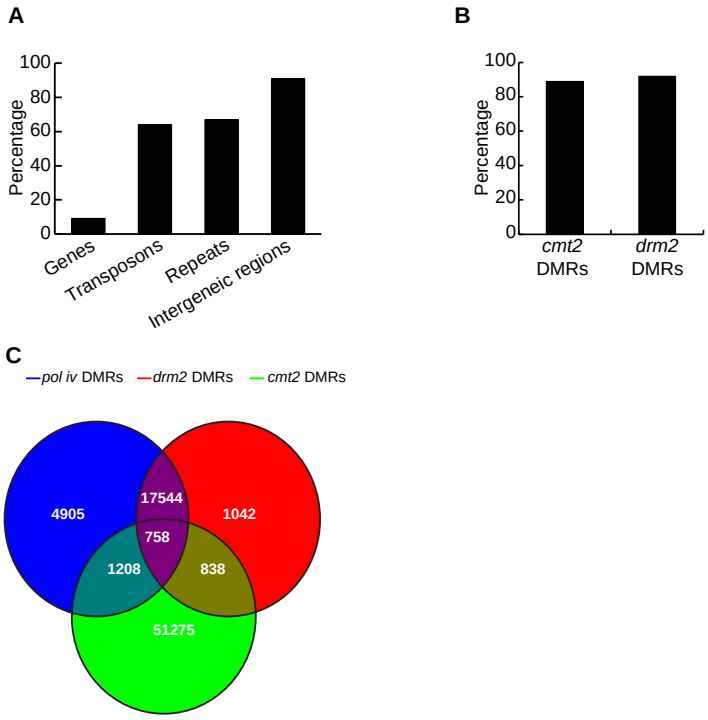
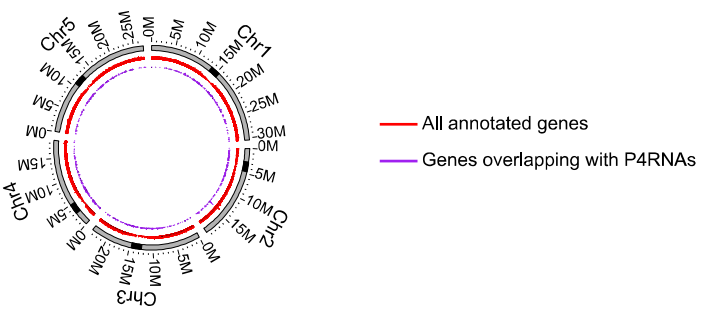
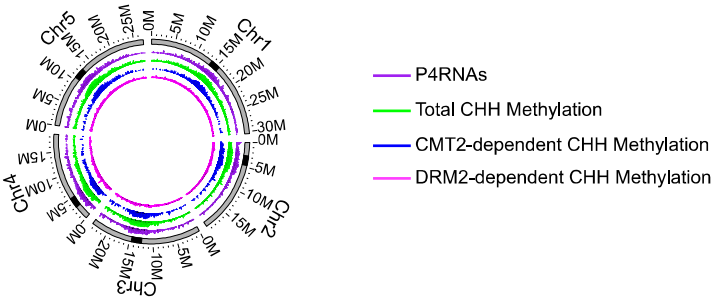

A
100
80
60
40
20
0
Percentage
Genes
Transposons
Repeats
Intergeneic regions
B
100
80
60
40
20
0
Percentage
cmt2 DMRs
drm2 DMRs
C
pol iv DMRs
drm2 DMRs
cmt2 DMRs
4905
17544
1042
758
1208
838
51275

Figure 2S. 6 Chromosomal distributions of P4RNAs and other genomic features. A, The chromosomal distribution of annotated genes and genes overlapping with P4RNAs. B, The chromosomal distribution of P4RNAs and CHH methylated regions. C, The chromosomal distribution of P4RNAs and regions containing H3K27me1 or H3K9me2. In A-C, the outermost layer represents each of the five chromosomes, with the centromeres indicated by the black bands. The inner layers represent the density of the featured regions (color coded) in 5 kb windows.

A


Chr1
Chr2
Chr3
Chr4
Chr5
All annotated genes
Genes overlapping with P4RNAs

B


Chr1
Chr2
Chr3
Chr4
Chr5
P4RNAs
Total CHH Methylation
CMT2-dependent CHH Methylation
DRM2-dependent CHH Methylation

C

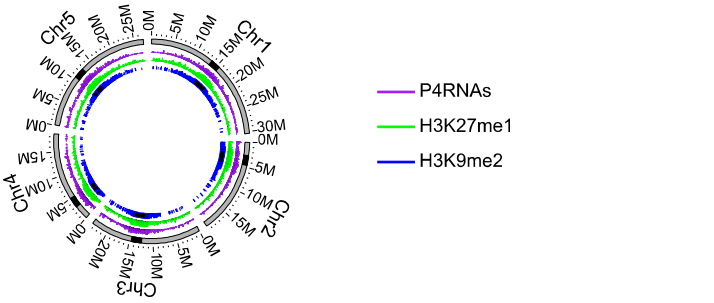

Chr1
Chr2
Chr3
Chr4
Chr5
P4RNAs
H3K27me1
H3K9me2

Figure 2S. 7 Features of P4RNAs and Pol II transcribed RNAs at P4siRNA loci.

A, A genome-browser view of reads from sRNA-seq and polyA- RNA-seq at a P4siRNA locus on chromosome 1. This locus is also shown in Figure 3E; two biological replicates (rep) are shown here. B, A genome-browser view of reads from sRNA-seq and polyA+ RNA-seq at a P4siRNA locus on chromosome 2. This locus is also shown in Figure 3F; two biological replicates are shown here. Normalized read numbers are shown above or below the horizontal lines for reads from the Watson and Crick strands, respectively. C, The percentage of transcripts derived from one major strand at P4siRNA loci in polyA+ RNA-seq. The numbers of reads from each of the two strands at P4siRNA loci were counted in polyA+ RNA-seq. Loci with 90% of the reads derived from one strand were considered as loci with transcripts derived from one major strand.

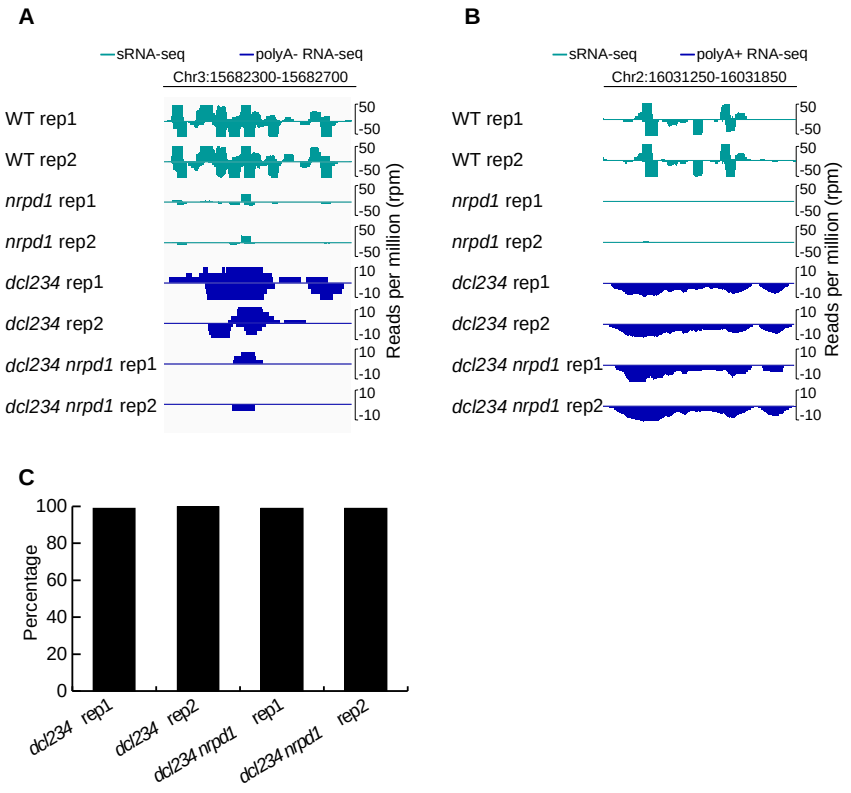

A
sRNA-seq
polyA- RNA-seq
Chr3:15682300-15682700
WT rep1
WT rep2
nrpd1 rep1
nrpd1 rep2
dcl234 rep1
dcl234 rep2
dcl234 nrpd1 rep1
dcl234 nrpd1 rep2
50
-50
10
-10
Reads per million (rpm)
B
sRNA-seq
polyA+ RNA-seq
Chr2:16031250-16031850
C
Percentage
100
80
60
40
20
0
dcl234 rep1
dcl234 rep2
dcl234 nrpd1 rep1
dcl234 nrpd1 rep2

Figure 2S. 8 Plots showing the strandedness of small RNAs and polyA+ RNAs from P4siRNA loci with Pol II transcribed RNAs.

The x-axis and y-axis represent the numbers of raw reads from the Watson and Crick strands, respectively. Each dot represents one P4siRNA locus, with the green and red colors representing small RNAs and polyA+ RNAs, respectively. Results from each of two biological replicates (rep) of polyA+ RNA-seq and the corresponding sRNA-seq from *dcl234* and *dcl234 nrpd1* are shown as indicated.

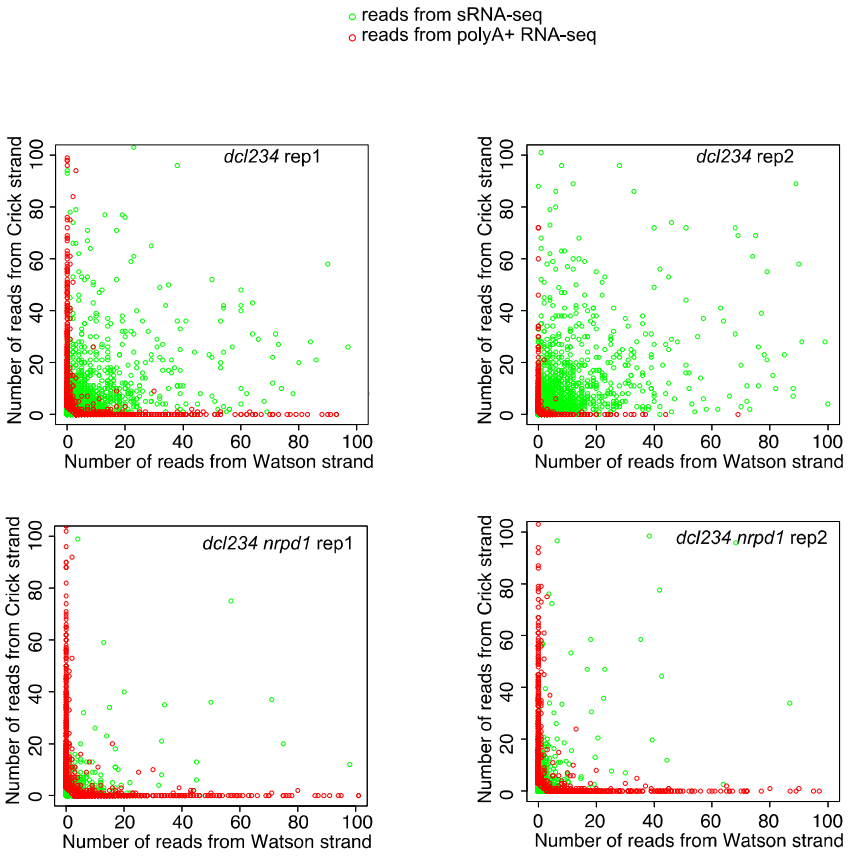

reads from sRNA-seq
reads from polyA+ RNA-seq
dcl234 rep1
dcl234 rep2
dcl234 nrpd1 rep1
dcl234 nrpd1 rep2
Number of reads from Crick strand
Number of reads from Watson strand
0
20
40
60
80
100

Figure 2S. 9 The presence of P4RNAs in *dcl234* is correlated with the levels of CHH methylation but not P4siRNA abundance.

A, A pie chart showing the reasons why P4RNAs were not detected by comparing *dcl234* to *dcl234 nrpd1* in dsRNA-seq. Low read abundance was defined as a total read count of less than 0.9RPM in all three *dcl234* libraries at a particular P4siRNA locus. The loci with p-value > 0.01 showed a consistent reduction in read abundance in *dcl234 nrpd1* but did not pass the p-value filter for the annotation of P4RNAs. B, The relative abundance of D2 and C2 siRNAs as determined by two replicates of sRNA-seq. P4siRNAs of 21nt, 22nt, 23nt, 24nt and 18-42nt (total) are shown. C and D, A lack of correlation between the ability to detect P4RNAs and the abundance of siRNAs at the corresponding loci in WT. P4siRNA loci were divided into four quartiles according to P4siRNA abundance in WT with the first quartile being loci containing the most abundant P4siRNAs. C, The CHH methylation level in WT for the four quartiles of loci. siRNA loci with or without P4RNA detected in our dsRNA-seq (comparing *dcl234* and *dcl234 nrpd1*) are shown separately; the two types of loci do not show drastic differences in their levels of CHH methylation in WT. D, The CHH methylation level for the four quartiles of loci in *dcl234*. P4siRNA loci without P4RNAs detected in our dsRNA-seq have lower CHH methylation in *dcl234* relative to loci with P4RNAs detected.

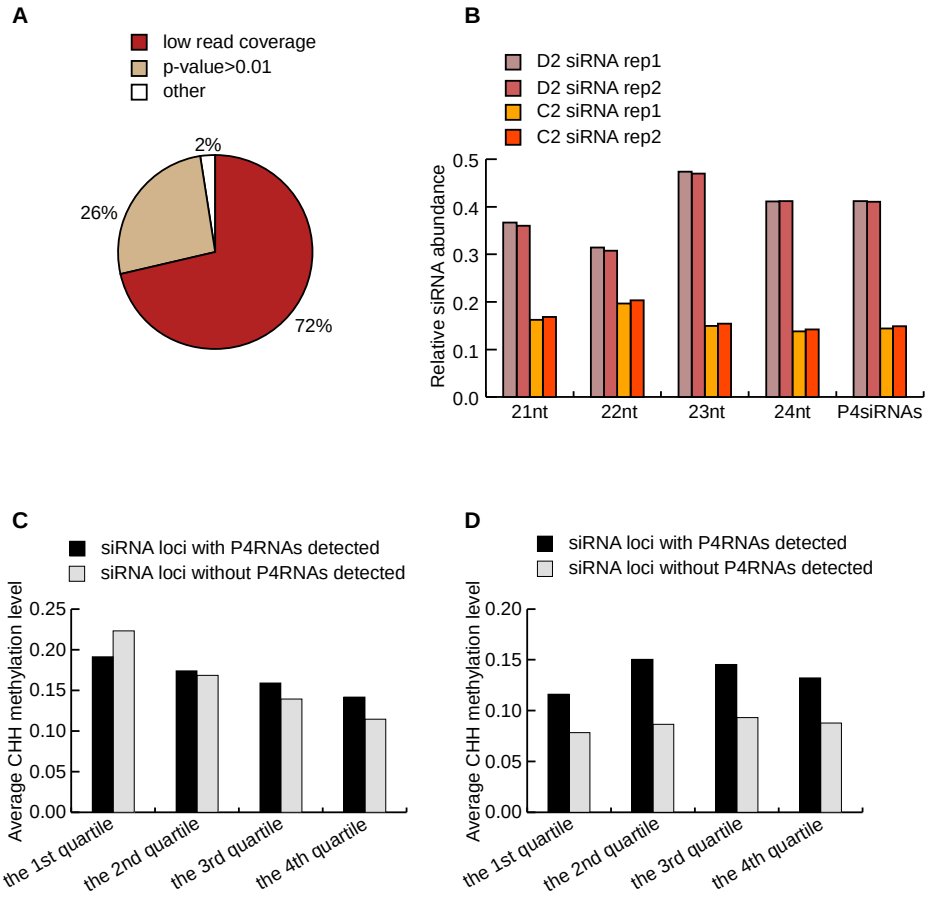

A
low read coverage
p-value>0.01
other
2%
26%
72%
B
D2 siRNA rep1
D2 siRNA rep2
C2 siRNA rep1
C2 siRNA rep2
Relative siRNA abundance
0.0
0.1
0.2
0.3
0.4
0.5
21nt
22nt
23nt
24nt
P4siRNAs
C
siRNA loci with P4RNAs detected
siRNA loci without P4RNAs detected
Average CHH methylation level
0.00
0.05
0.10
0.15
0.20
0.25
the 1st quartile
the 2nd quartile
the 3rd quartile
the 4th quartile
D
siRNA loci with P4RNAs detected
siRNA loci without P4RNAs detected
Average CHH methylation level
0.00
0.05
0.10
0.15
0.20
the 1st quartile
the 2nd quartile
the 3rd quartile
the 4th quartile

Figure 2S. 10 Differences between D2 and C2 loci in P4RNA discovery, P4siRNA levels, and CHH methylation levels.

A-C, D2 and C2 siRNA loci were divided into four quartiles according to their CHH methylation levels in *dcl234*. A, The relative abundance of P4RNAs in *dcl234* in each quartile. B, Average CHH methylation levels in *dcl234* in the four quartiles. C, The relative abundance of P4siRNAs in *dcl234* in each quartile. D-E, D2 and C2 loci were divided into four quartiles according to their CHH methylation levels in WT. D, The average CHH methylation levels in WT in the four quartiles. E, The relative abundance of P4siRNAs in WT in each quartile.

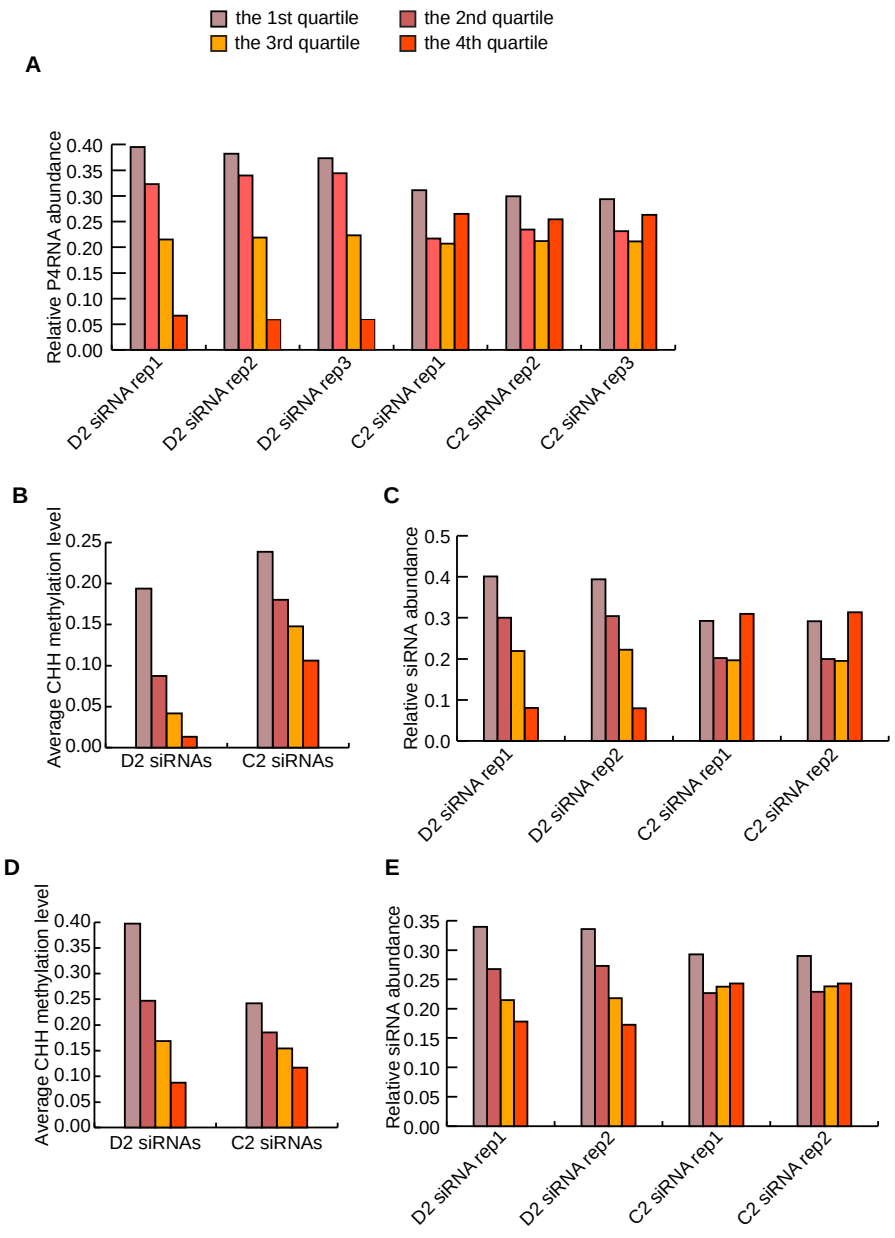
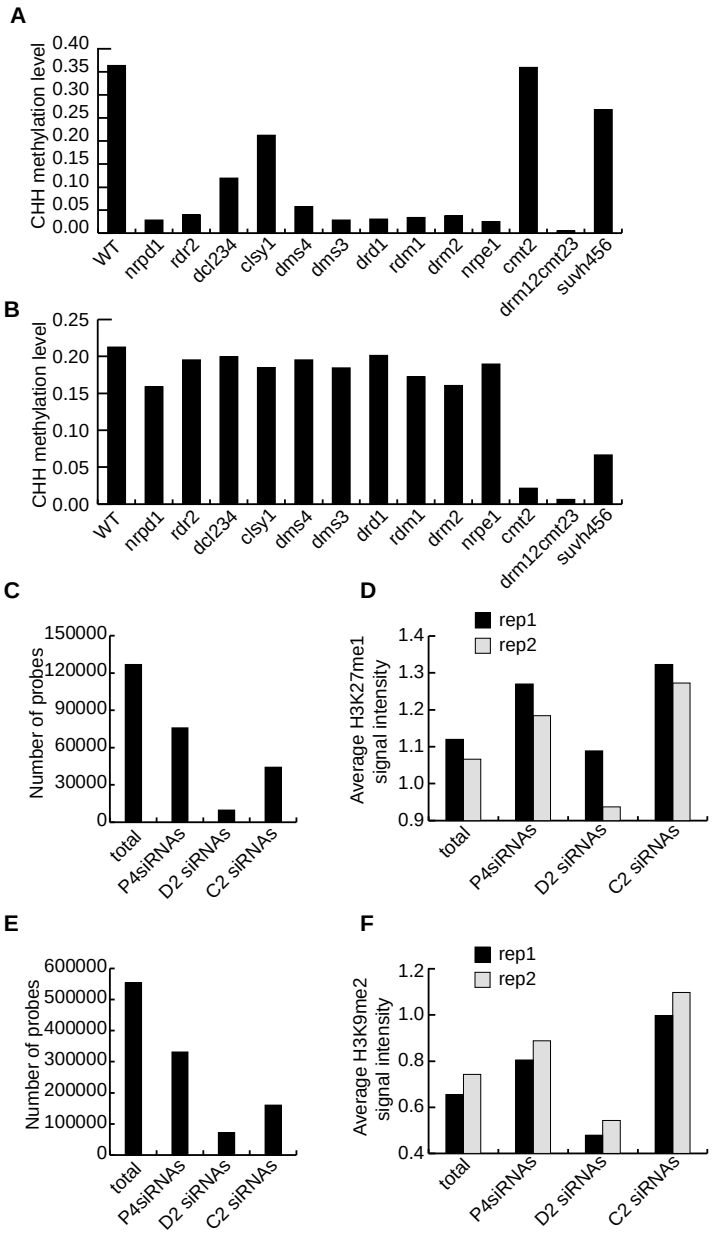

the 1st quartile
the 2nd quartile
the 3rd quartile
the 4th quartile
A
Relative P4RNA abundance
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
D2 siRNA rep1
D2 siRNA rep2
D2 siRNA rep3
C2 siRNA rep1
C2 siRNA rep2
C2 siRNA rep3
B
Average CHH methylation level
0.25
0.20
0.15
0.10
0.05
0.00
D2 siRNAs
C2 siRNAs
C
Relative siRNA abundance
0.5
0.4
0.3
0.2
0.1
0.0
D2 siRNA rep1
D2 siRNA rep2
C2 siRNA rep1
C2 siRNA rep2
D
Average CHH methylation level
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
D2 siRNAs
C2 siRNAs
E
Relative siRNA abundance
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
D2 siRNA rep1
D2 siRNA rep2
C2 siRNA rep1
C2 siRNA rep2

Figure 2S. 11 CHH methylation, H3K27me1, and H3K9me2 levels in WT and various mutants.

A-B, CHH methylation levels in various genotypes at D2 (A) and C2 (B) loci. CHH methylation levels were determined (see Methods) using published methylome data (Stroud et al. 2013; Stroud et al. 2014). C-D, H3K27me1 levels at D2 and C2 loci as determined by ChIP-chip (Roudier et al. 2011). C, The number of probes that show H3K27me1 signals at various genomic features in the published ChIP-chip study. D, The average H3K27me1 ChIP-chip signal intensity at the indicated genomic regions corresponding to the probes in C. Results from two biological replicates (rep) are shown separately. E-F, H3K9me2 levels at D2 and C2 loci as determined by ChIP-chip (Deleris et al. 2012). E, The number of probes with H3K9me2 signals at various genomic features in the published ChIP-chip study. F, The average H3K9me2 signal intensity at the indicated genomic regions corresponding to the probes in E. Results from two biological replicates (rep) are shown separately.


A
CHH methylation level
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
WT
nrpd1
rdr2
dcl234
clsy1
dms4
dms3
drd1
rdm1
drm2
nrpe1
cmt2
drm12cmt23
suvh456
B
CHH methylation level
0.25
0.20
0.15
0.10
0.05
0.00
WT
nrpd1
rdr2
dcl234
clsy1
dms4
dms3
drd1
rdm1
drm2
nrpe1
cmt2
drm12cmt23
suvh456
C
Number of probes
150000
120000
90000
60000
30000
0
total
P4siRNAs
D2 siRNAs
C2 siRNAs
D
rep1
rep2
Average H3K27me1 signal intensity
1.4
1.3
1.2
1.1
1.0
0.9
total
P4siRNAs
D2 siRNAs
C2 siRNAs
E
Number of probes
600000
500000
400000
300000
200000
100000
0
total
P4siRNAs
D2 siRNAs
C2 siRNAs
F
rep1
rep2
Average H3K9me2 signal intensity
1.2
1.0
0.8
0.6
0.4
total
P4siRNAs
D2 siRNAs
C2 siRNAs

Tables

Table 2. 1 Chromosomal positions of the P4siRNA loci examined by RT-PCR in this study.

Name	Chromosome	Start Position	End Position
Locus 1	Chr1	11619088	11619830
Locus 2	Chr3	5780028	5780762
Locus 3	Chr3	7419920	7421330
Locus 4	Chr3	10691074	10691841
Locus 5	Chr3	10747222	10748309
Locus 6	Chr1	4506452	4507032
Locus 8	Chr2	5661047	5661660
Locus 13	Chr2	2865442	2866452
Locus 20	Chr3	11042663	11043163
Locus 23	Chr3	14729148	14731788
Locus 26	Chr3	15682149	15682550
Locus 27	Chr3	17842320	17843219
Locus 28	Chr3	20030863	20031378
Locus 30	Chr1	23453816	23455008
Locus 33	Chr4	12841422	12842570
Locus 38	Chr4	272801	273244
Locus 40	Chr5	9800868	9801476
Locus 41	Chr5	1410300	1410550
Locus 42	Chr5	17174556	17175364
Locus 43	Chr5	20313464	20314875
Locus 44	Chr5	22706840	22707252
Locus 45	Chr5	22707688	22708114

Table 2. 2 GO annotation of genes overlapping with P4RNAs.

GO ID	Gene[1] number	Gene[2] number	p-value	Adjusted p-value	GO Term	GO category
GO:0004565	40	12	9.37E-10	7.59E-08	beta-galactosidase activity	molecular function
GO:0015925	44	12	3.18E-09	2.57E-07	galactosidase activity	molecular function
GO:0016798	442	33	1.02E-06	8.25E-05	hydrolase activity, acting on glycosyl bonds	molecular function
GO:0005199	39	9	1.43E-06	1.16E-04	structural constituent of cell wall	molecular function
GO:0004553	412	31	1.83E-06	1.48E-04	hydrolase activity, hydrolyzing O-glycosyl compounds	molecular function
GO:0030145	39	8	1.41E-05	1.14E-03	manganese ion binding	molecular function
GO:0030599	147	15	2.80E-05	2.27E-03	pectinesterase activity	molecular function
GO:0004650	71	10	3.94E-05	3.19E-03	polygalacturonase activity	molecular function
GO:0007047	165	16	3.50E-05	1.58E-03	cell wall organization	biological process
GO:0045229	183	16	1.21E-04	5.45E-03	external encapsulating structure organization	biological process
GO:0070882	229	18	1.82E-04	8.18E-03	cell wall organization or biogenesis	biological process
GO:0012505	4063	207	2.10E-15	4.62E-14	endomembrane system	cellular component

GO:0009341	30	10	9.42E-09	2.07E-07	beta-galactosidase complex	cellular component

1 All annotated genes

2 Genes overlapping with P4RNAs

Table 2. 3 Published genomic datasets used in this study.

Library	Genotype	Geo ID	Publication
ChIP-chip (H3K9me2)	Col	GSE37075	Deleris et al. 2012
ChIP-chip (H3K27me1)	Col	GSE24710	Roudier et al. 2011
BS-seq	Col rep1	GSM938370	Stroud et al. 2013
BS-seq	Col rep2	GSM980986	Stroud et al. 2013
BS-seq	Col rep3	GSM980987	Stroud et al. 2013
BS-seq	*clsy1*	GSM981000	Stroud et al. 2013
BS-seq	*cmt2*	GSM981002	Stroud et al. 2013
BS-seq	*dcl234*	GSM981008	Stroud et al. 2013
BS-seq	*dms3*	GSM981010	Stroud et al. 2013
BS-seq	*dms4*	GSM981011	Stroud et al. 2013
BS-seq	*drd1*	GSM981014	Stroud et al. 2013
BS-seq	*drm12*	GSM981015	Stroud et al. 2013
BS-seq	*nrpd1*	GSM981039	Stroud et al. 2013
BS-seq	*nrpe1*	GSM981040	Stroud et al. 2013
BS-seq	*rdr2*	GSM981044	Stroud et al. 2013
BS-seq	*rdm1*	GSM981042	Stroud et al. 2013
BS-seq	*suvh456*	GSM981060	Stroud et al. 2013
BS-seq	Col	GSM1242401	Stroud et al. 2014
BS-seq	*drm12cmt23*	GSM1242404	Stroud et al. 2014
sRNA-seq	Col	GSM1242406	Stroud et al. 2014
sRNA-seq	*cmt2*	GSM1242407	Stroud et al. 2014
sRNA-seq	*drm12cmt23*	GSM1242409	Stroud et al. 2014
sRNA-seq	*suvh456*	GSM1242410	Stroud et al. 2014
sRNA-seq	Col	GSM893118	Lee et al. 2012
sRNA-seq	*dms4*	GSM893119	Lee et al. 2012
sRNA-seq	*drd1*	GSM893120	Lee et al. 2012
sRNA-seq	*dms3*	GSM893121	Lee et al. 2012
sRNA-seq	*rdm1*	GSM893122	Lee et al. 2012
sRNA-seq	Col rep1	GSM1103235	Law et al. 2013
sRNA-seq	Col rep2	GSM1103236	Law et al. 2013
sRNA-seq	*nrpe1*	GSM1103238	Law et al. 2013
sRNA-seq	*drm2*	GSM1103240	Law et al. 2013

Table 2. 4 Primers used in this study.

Name	Sequence	Purpose
Locus 1F	AATACAAGCAACATAGGGAAG	RT-PCR for locus 1 Watson strand primer for RT
Locus 1R	AACCAAGCCACAAATCTCT	RT-PCR for locus 1 Crick strand primer for RT
Locus 2F	TATCGTATTGTCGTCCTTGA	RT-PCR for locus 2 Watson strand primer for RT
Locus 2R	GTCCCACTCCACTTTCATT	RT-PCR for locus 2 Crick strand primer for RT
Locus 3F	GGGAAACGACTTTGTATGTT	RT-PCR for locus 3 Watson strand primer for RT
Locus 3R	ATTGCTCTGGTGTTCTCACT	RT-PCR for locus 3 Crick strand primer for RT
Locus 4F	AGCATCCCCAATAACAAAT	RT-PCR for locus 4 Watson strand primer for RT
Locus 4R	ATCTACGAGGTCAGTCAAGG	RT-PCR for locus 4 Crick strand primer for RT
Locus 5F	CGAACAGCACCACTAAGC	RT-PCR for locus 5 Watson strand primer for RT
Locus 5R	GAAGGAAAAGCAACTCACTC	RT-PCR for locus 5 Crick strand primer for RT
Locus 6F	GCATCATTCACAGTATCCAA	RT-PCR for locus 6 Watson strand primer for RT
Locus 6R	GTTCTTCTTCTTCGGGTATC	RT-PCR for locus 6 Crick strand primer for RT
Locus 8F	AAAGAGATGTTGGTGAAAGG	RT-PCR for locus 8
Locus 8R	CTTGATGGGTGGAATGAC	RT-PCR for locus 8
Locus 13F	TAAGATTGATGTAACTGGGAAG	RT-PCR for locus 13 Watson strand primer for RT
Locus 13R	TCGGTAGAGATGACTTGAGA	RT-PCR for locus 13 Crick strand primer for RT
Locus 20F	GAACAAGGCTACTGTGGTG	RT-PCR for locus 20 Watson strand primer for RT
Locus 20R	GGAAGGCATCCATTTGAT	RT-PCR for locus 20 Crick strand primer for RT
Locus 23F	AAGAAAGCCCAAGTAGAAGA	RT-PCR for locus 23
Locus 23R	AGCGTATCAACCCAAATG	RT-PCR for locus 23
Locus 26F	AACTACCCCAATCCTTTCTA	RT-PCR for locus 26

Locus 26R	CTGGTCACTTCTCCGATG	RT-PCR for locus 26
Locus 27F	TACTCTTGGCTTCTCAAAAC	RT-PCR for locus 27
Locus 27R	CATTGTGTCCTCCTGTTACC	RT-PCR for locus 27
Locus 28F	TGGATACTTGCCTCGTGT	RT-PCR for locus 28
Locus 28R	CCAGATGGAGACATTATTG	RT-PCR for locus 28
Locus 29F	CTTATGGCGGTTCTCAGT	RT-PCR for locus 29
Locus 29R	TCCTTCTCTCTCTTCTCCAG	RT-PCR for locus 29
Locus 30F	ATAGCCTTCAACACTTGCTT	RT-PCR for locus 30 Watson strand primer for RT
Locus 30R	GAGTTCATTCTCCGACTTTC	RT-PCR for locus 30 Crick strand primer for RT
Locus 33F	CCAGAAGAATAGCATAGAAGC	RT-PCR for locus 33
Locus 33R	TAGGAATACAAGACCTCAAATG	RT-PCR for locus 33
Locus 34F	ATGTTGAATGGCTCTATGC	RT-PCR for locus 34
Locus 34R	ACGCTCTTGCTCATCTTC	RT-PCR for locus 34
Locus 35F	TCCTCCTCATTCTCCTACAT	RT-PCR for locus 35
Locus 35R	AACTTTTCAGACCTAACATCAA	RT-PCR for locus 35
Locus 38F	GATGGACTCTCTGGCTTG	RT-PCR for locus 38
Locus 38R	AACGGTGGTGATTATGGA	RT-PCR for locus 38
Locus 40F	ATTATTCAAACTCACCACAAAG	RT-PCR for locus 40
Locus 40R	AATCGCCTTCACAACATTA	RT-PCR for locus 40
Locus 41F	TGCTTTTCCTTCACTCTTCT	RT-PCR for locus 41
Locus 41R	TAACGGCTCTATCACTTTTG	RT-PCR for locus 41
Locus 42F	AGGGAGTAATAGATGTGATGG	RT-PCR for locus 42
Locus 42R	ATTTAGGAGGAGCAAAAGC	RT-PCR for locus 42
Locus 43F	GGTGTTGGATAAAGGGTAGA	RT-PCR for locus 43
Locus 43R	CATCTTGTGAGCAGGAAAA	RT-PCR for locus 43
Locus 44F	GTAAATAAACCCAAGAACCAC	RT-PCR for locus 44
Locus 44R	TGCGAAACTAATGGAAGAAT	RT-PCR for locus 44
Locus 45F	TTTGGTAGAATAGAAGGAATGA	RT-PCR for locus 45
Locus 45R	TGAAATAAGATGGGGACAAT	RT-PCR for locus 45
UBC-F	TACAGCGAGAGAAAGTAGCA	RT-PCR for locus *UBC21*
UBC-R	GCAAAGGATAAGGTTCAGG	RT-PCR for locus *UBC21*
CBP20-F	TCAGGAACACAAGAGGAGTT	RT-PCR for locus *CBP20*
CBP20-R	AGAACAGGACGAAACAAAAG	RT-PCR for locus *CBP20*

Table 2. 5 Genomic datasets generated in this study[1].

Library	Genotype
dsRNA-seq	*dcl2-1 dcl3-1 dcl4-2* rep1[2]
dsRNA-seq	*dcl2-1 dcl3-1 dcl4-2* rep2
dsRNA-seq	*dcl2-1 dcl3-1 dcl4-2* rep3
dsRNA-seq	*dcl2-1 dcl3-1 dcl4-2 nrpd1-3* rep1
dsRNA-seq	*dcl2-1 dcl3-1 dcl4-2 nrpd1-3* rep2
dsRNA-seq	*dcl2-1 dcl3-1 dcl4-2 nrpd1-3* rep3
RNA-seq	*dcl2-1 dcl3-1 dcl4-2* rep1
RNA-seq	*dcl2-1 dcl3-1 dcl4-2* rep2
RNA-seq	*dcl2-1 dcl3-1 dcl4-2* rep3
RNA-seq	*dcl2-1 dcl3-1 dcl4-2 nrpd1-3* rep1
RNA-seq	*dcl2-1 dcl3-1 dcl4-2 nrpd1-3* rep2
RNA-seq	*dcl2-1 dcl3-1 dcl4-2 nrpd1-3* rep3
RNA-seq-DSN	*dcl2-1 dcl3-1 dcl4-2*
RNA-seq-DSN	*dcl2-1 dcl3-1 dcl4-2 nrpd1-3*
RNA-seq-DSN	*dcl2-1 dcl3-1 dcl4-2 rdr2-1*
RNA-seq (poly A+)	*dcl2-1 dcl3-1 dcl4-2* rep1
RNA-seq (poly A+)	*dcl2-1 dcl3-1 dcl4-2* rep2
RNA-seq (poly A+)	*dcl2-1 dcl3-1 dcl4-2 nrpd1-3* rep1
RNA-seq (poly A+)	*dcl2-1 dcl3-1 dcl4-2 nrpd1-3* rep2
RNA-seq (poly A-)	*dcl2-1 dcl3-1 dcl4-2* rep1
RNA-seq (poly A-)	*dcl2-1 dcl3-1 dcl4-2* rep2
RNA-seq (poly A-)	*dcl2-1 dcl3-1 dcl4-2 nrpd1-3* rep1
RNA-seq (poly A-)	*dcl2-1 dcl3-1 dcl4-2 nrpd1-3* rep2
sRNA-seq	Col rep1
sRNA-seq	Col rep2
sRNA-seq	*nrpd1-3* rep1
sRNA-seq	*nrpd1-3* rep2
sRNA-seq	*dcl2-1 dcl3-1 dcl4-2* rep1
sRNA-seq	*dcl2-1 dcl3-1 dcl4-2* rep2
sRNA-seq	*dcl2-1 dcl3-1 dcl4-2 nrpd1-3* rep1
sRNA-seq	*dcl2-1 dcl3-1 dcl4-2 nrpd1-3* rep2
sRNA-seq	*rdr2-1* rep1
sRNA-seq	*rdr2-2* rep2
sRNA-seq	*dcl3-1* rep1
sRNA-seq	*dcl3-1* rep2
sRNA-seq	*clsy1*

1 The datasets have been deposited in the Gene Expression Omnibus at National Center for Biotechnology Information under the accession number GSE57215.

2 rep: biological replicate

References

Allen E, Xie Z, Gustafson AM, Carrington JC. 2005. microRNA-directed phasing during trans-acting siRNA biogenesis in plants. *Cell* **121**(2): 207-221.

Ammar R, Torti D, Tsui K, Gebbia M, Durbic T, Bader GD, Giaever G, Nislow C. 2012. Chromatin is an ancient innovation conserved between Archaea and Eukarya. *eLife* **1**: e00078.

Cao X, Jacobsen SE. 2002. Locus-specific control of asymmetric and CpNpG methylation by the DRM and CMT3 methyltransferase genes. *Proceedings of the National Academy of Sciences of the United States of America* **99 Suppl 4**: 16491-16498.

Cho SH, Addo-Quaye C, Coruh C, Arif MA, Ma Z, Frank W, Axtell MJ. 2008. Physcomitrella patens DCL3 is required for 22-24 nt siRNA accumulation, suppression of retrotransposon-derived transcripts, and normal development. *PLoS genetics* **4**(12): e1000314.

Chodavarapu RK, Feng S, Bernatavichute YV, Chen PY, Stroud H, Yu Y, Hetzel JA, Kuo F, Kim J, Cokus SJ et al. 2010. Relationship between nucleosome positioning and DNA methylation. *Nature* **466**(7304): 388-392.

Dahlber AE, Dahlber JE, Lund E, Tokimatsu H, Rabson AB, Calvert PC, ReynoldS F, Zahalak M. 1978. Processing of the 5'end of Escherichia coli 16S ribosomal RNA. *PNAS* **75**(2): 3598-3602.

Deleris A, Stroud H, Bernatavichute Y, Johnson E, Klein G, Schubert D, Jacobsen SE. 2012. Loss of the DNA methyltransferase MET1 Induces H3K9 hypermethylation at PcG target genes and redistribution of H3K27 trimethylation to transposons in Arabidopsis thaliana. *PLoS genetics* **8**(11): e1003062.

Erhard KF, Jr., Stonaker JL, Parkinson SE, Lim JP, Hale CJ, Hollick JB. 2009. RNA polymerase IV functions in paramutation in Zea mays. *Science* **323**(5918): 1201-1205.

Haag JR, Ream TS, Marasco M, Nicora CD, Norbeck AD, Pasa-Tolic L, Pikaard CS. 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**(5): 811-818.

Henderson IR, Zhang X, Lu C, Johnson L, Meyers BC, Green PJ, Jacobsen SE. 2006. Dissecting Arabidopsis thaliana DICER function in small RNA processing, gene silencing and DNA methylation patterning. *Nature genetics* **38**(6): 721-725.

Herr AJ, Jensen MB, Dalmay T, Baulcombe DC. 2005. RNA polymerase IV directs silencing of endogenous DNA. *Science* **308**(5718): 118-120.

Jia Y, Lisch DR, Ohtsu K, Scanlon MJ, Nettleton D, Schnable PS. 2009. Loss of RNA-dependent RNA polymerase 2 (RDR2) function causes widespread and unexpected changes in the expression of transposons, genes, and 24-nt small RNAs. *PLoS genetics* **5**(11): e1000737.

Johnson LM, Du J, Hale CJ, Bischof S, Feng S, Chodavarapu RK, Zhong X, Marson G, Pellegrini M, Segal DJ et al. 2014. SRA- and SET-domain-containing proteins link RNA polymerase V occupancy to DNA methylation. *Nature* **507**(7490): 124-128.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology* **14**(4): R36.

Law JA, Ausin I, Johnson LM, Vashisht AA, Zhu JK, Wohlschlegel JA, Jacobsen SE. 2010. A protein complex required for polymerase V transcripts and RNA-directed DNA methylation in Arabidopsis. *Current biology : CB* **20**(10): 951-956.

Law JA, Du J, Hale CJ, Feng S, Krajewski K, Palanca AM, Strahl BD, Patel DJ, Jacobsen SE. 2013. Polymerase IV occupancy at RNA-directed DNA methylation sites requires SHH1. *Nature* **498**(7454): 385-389.

Law JA, Jacobsen SE. 2010. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature reviews Genetics* **11**(3): 204-220.

Law JA, Vashisht AA, Wohlschlegel JA, Jacobsen SE. 2011. SHH1, a homeodomain protein required for DNA methylation, as well as RDR2, RDM4, and chromatin remodeling factors, associate with RNA polymerase IV. *PLoS genetics* **7**(7): e1002195.

Lee TF, Gurazada SG, Zhai J, Li S, Simon SA, Matzke MA, Chen X, Meyers BC. 2012. RNA polymerase V-dependent small RNAs in Arabidopsis originate from small, intergenic loci including most SINE repeats. *Epigenetics* **7**(7): 781-795.

Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR. 2008. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* **133**(3): 523-536.

Liu Q, Feng Y, Zhu Z. 2009. Dicer-like (DCL) proteins in plants. *Funct Integr Genomics* **9**(3): 277-286.

Matzke MA, Mosher RA. 2014. RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. *Nature reviews Genetics* **15**(6): 394-408.

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research* **18**(9): 1509-1517.

Mosher RA, Schwach F, Studholme D, Baulcombe DC. 2008. PolIVb influences RNA-directed DNA methylation independently of its role in siRNA biogenesis. *Proceedings of the National Academy of Sciences of the United States of America* **105**(8): 3145-3150.

Onodera Y, Haag JR, Ream T, Costa Nunes P, Pontes O, Pikaard CS. 2005. Plant nuclear RNA polymerase IV mediates siRNA and DNA methylation-dependent heterochromatin formation. *Cell* **120**(5): 613-622.

Pauler FM, Sloane MA, Huang R, Regha K, Koerner MV, Tamir I, Sommer A, Aszodi A, Jenuwein T, Barlow DP. 2009. H3K27me3 forms BLOCs over silent genes and intergenic regions and specifies a histone banding pattern on a mouse autosomal chromosome. *Genome research* **19**(2): 221-233.

Polishko A, Ponts N, Le Roch KG, Lonardi S. 2012. NORMAL: accurate nucleosome positioning using a modified Gaussian mixture model. *Bioinformatics* **28**(12): i242-249.

Pontier D, Yahubyan G, Vega D, Bulski A, Saez-Vasquez J, Hakimi MA, Lerbs-Mache S, Colot V, 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**(17): 2030-2040.

Qi Y, He X, Wang XJ, Kohany O, Jurka J, Hannon GJ. 2006. Distinct catalytic and non-catalytic roles of ARGONAUTE4 in RNA-directed DNA methylation. *Nature* **443**(7114): 1008-1012.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**(1): 139-140.

Roudier F, Ahmed I, Berard C, Sarazin A, Mary-Huard T, Cortijo S, Bouyer D, Caillieux E, Duvernois-Berthet E, Al-Shikhley L et al. 2011. Integrative epigenomic mapping defines four main chromatin states in Arabidopsis. *The EMBO journal* **30**(10): 1928-1938.

Smith LM, Pontes O, Searle I, Yelina N, Yousafzai FK, Herr AJ, Pikaard CS, Baulcombe DC. 2007. An SNF2 protein associated with nuclear RNA silencing and the spread of a silencing signal between cells in Arabidopsis. *The Plant cell* **19**(5): 1507-1521.

Stroud H, Do T, Du J, Zhong X, Feng S, Johnson L, Patel DJ, Jacobsen SE. 2014. Non-CG methylation patterns shape the epigenetic landscape in Arabidopsis. *Nature structural & molecular biology* **21**(1): 64-72.

Stroud H, Greenberg MV, Feng S, Bernatavichute YV, Jacobsen SE. 2013. Comprehensive analysis of silencing mutants reveals complex regulation of the Arabidopsis methylome. *Cell* **152**(1-2): 352-364.

Unfried I, Gruendler P. 1990. Nucleotide sequence of the 5.8S and 25S rRNA genes and of the internal transcribed spacers from Arabidopsis thaliana. *Nucleic acids research* **18**(13): 4011.

Wang XB, Jovel J, Udomporn P, Wang Y, Wu Q, Li WX, Gasciolli V, Vaucheret H, Ding SW. 2011. The 21-nucleotide, but not 22-nucleotide, viral secondary small interfering RNAs direct potent antiviral defense by two cooperative argonautes in Arabidopsis thaliana. *The Plant cell* **23**(4): 1625-1638.

Wierzbicki AT, Cocklin R, Mayampurath A, Lister R, Rowley MJ, Gregory BD, Ecker JR, Tang H, Pikaard CS. 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.

Wierzbicki AT, Haag JR, Pikaard CS. 2008. Noncoding transcription by RNA polymerase Pol IVb/Pol V mediates transcriptional silencing of overlapping and adjacent genes. *Cell* **135**(4): 635-648.

Wierzbicki AT, Ream TS, Haag JR, Pikaard CS. 2009. RNA polymerase V transcription guides ARGONAUTE4 to chromatin. *Nature genetics* **41**(5): 630-634.

Xie Z, Johansen LK, Gustafson AM, Kasschau KD, Lellis AD, Zilberman D, Jacobsen SE, Carrington JC. 2004. Genetic and functional diversification of small RNA pathways in plants. *PLoS biology* **2**(5): E104.

Zemach A, Kim MY, Hsieh PH, Coleman-Derr D, Eshed-Williams L, Thao K, Harmer SL, Zilberman D. 2013. The Arabidopsis nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin. *Cell* **153**(1): 193-205.

Zhang CJ, Zhou JX, Liu J, Ma ZY, Zhang SW, Dou K, Huang HW, Cai T, Liu R, Zhu JK et al. 2013. The splicing machinery promotes RNA-directed DNA methylation and transcriptional silencing in Arabidopsis. *The EMBO journal* **32**(8): 1128-1140.

Zhang J, Madden TL. 1997. PowerBLAST: a new network BLAST application for interactive or automated sequence analysis and annotation. *Genome research* **7**(6): 649-656.

Zheng B, Wang Z, Li S, Yu B, Liu JY, 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**(24): 2850-2860.

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

Zhong X, Du J, Hale CJ, Gallego-Bartolome J, Feng S, Vashisht AA, Chory J, Wohlschlegel JA, Patel DJ, Jacobsen SE. 2014. Molecular mechanism of action of plant DRM de novo DNA methyltransferases. *Cell* **157**(5): 1050-1060.

Zhong X, Hale CJ, Law JA, Johnson LM, Feng S, Tu A, Jacobsen SE. 2012. DDR complex facilitates global association of RNA polymerase V to promoters and evolutionarily young transposons. *Nature structural & molecular biology* **19**(9): 870-875.

Chapter 3. SUVH1, a histone methyltransferase, is required for the expression of genes targeted by DNA methylation

Abstract

Transposons and repeats are found throughout the genomes of all organisms. To prevent the harmful effects of these elements, repressive marks such as DNA methylation and H3K9me2 have evolved to control transposon activity and ultimately maintain genome integrity. However, how silencing mechanisms are themselves regulated to avoid stochastic silencing of genes remains unclear. Here, negative regulators of silencing were identified using a forward-genetic screen on a reporter line that harbors a *LUCIFERASE* (*LUC*) gene driven by a double *35S* promoter. SUVH1, a SU(VAR)3-9 homolog, was isolated as a factor promoting the expression of the *LUC* gene. Treatment with a cytosine methylation inhibitor abolished the effect of the *suvh1* mutation, indicating that SUVH1 is dispensable for *LUC* expression in the absence of DNA methylation. However, the *suvh1* mutation did not alter DNA methylation levels at the *LUC* region or on a genome-wide scale; thus, SUVH1 may function downstream of DNA methylation. Histone methylation analysis revealed that *suvh1* led to decreased H3K4me3 levels; in contrast, H3K9me2 levels remained unchanged. Moreover, characterization of endogenous genes indicated that SUVH1 functions at genes with repressive marks in the promoter region. Taken together, these findings shed light on the regulatory network acting at genes with various epigenetic marks.

Introduction

Chromatin structure, histone modifications and DNA methylation regulate expression and influence transposon activity. The model plant *Arabidopsis* has been used to uncover the molecular framework of DNA methylation, which is critical for the regulation of transposon activity and the maintenance of genome integrity. The RNA-directed DNA methylation (RdDM) pathway is responsible for establishing DNA methylation at CG, CHG and CHH contexts (H = A, C, T) and maintaining asymmetrical CHH methylation (Law and Jacobsen 2010; Matzke and Mosher 2014). To maintain symmetrical CG and CHG methylation during DNA replication, DNA methyltransferases MET1 and CMT3 methylate the newly synthesized strand using the old, methylated strand as a guide (Chan et al. 2005; Stroud et al. 2013). DNA methylation can also be actively erased through demethylation. Four DNA glycosylases involved in DNA demethylation are known (Morales-Ruiz et al. 2006; Penterman et al. 2007): DME, which functions primarily in the seed (Choi et al. 2002), and three DME homologs (ROS1, DML2 and DML3) with broader domains of activity in the plant (Gong et al. 2002; Lister et al. 2008; Ortega-Galisteo et al. 2008).

Histone modifications also influence gene expression. H3K4me3 is a well-recognized active mark deposited by the SET domain proteins ATX1 and ATXR3 (Alvarez-Venegas and Avramova 2005; Berr et al. 2010; Guo et al. 2010). H3K9me2 is a repressive mark that mediates the chromatin association of CMT3 through its bromo-adjacent homology and chromo domains (Du et al. 2012). Human, murine and yeast Su(var)3-9 proteins were shown to have histone methyltransferase activity (Rea et al.

2000). In *Arabidopsis*, there are ten Su(var)3-9 homologs, which can be divided into the four following subgroups: SUVH1, SUVH2, SUVH4 and SUVH5 (Naumann et al. 2005). SUVH4, SUVH5 and SUVH6 belonging to the SUVH4 and SUVH5 subgroups are active H3K9me2 methyltransferases (Ebbs and Bender 2006; Stroud et al. 2014). SUVH2 and SUVH9 in the SUVH2 subgroup are players in RdDM; they are required for Pol V occupancy at regions with DNA methylation (Johnson et al. 2014; Liu et al. 2014). No functions have been reported for any of the SUVH1 subgroup proteins, which include SUVH1, SUVH3, SUVH7, SUVH8 and SUVH10 (Naumann et al. 2005).

Although DNA methylation and H3K9me2 largely occur in heterochromatic regions, they are also found in euchromatic regions where genes are located. In fact, when such epigenetic modifications are close to genes, the expression of the nearby genes could be repressed (Soppe et al. 2000; Liu et al. 2004; Henderson and Jacobsen 2008). This raises the question of how genes with nearby transposable elements can overcome the effects of epigenetic silencing to be expressed. With the goal of identifying negative regulators of gene silencing, a forward genetic screen was performed using a reporter line named *YJ11-3F* (hereafter referred to as *YJ*), which harbors a luciferase gene (*LUC*) driven by a double *35S* promoter, which harbors DNA methylation. A mutation causing decreased luciferase activity was mapped to the *SUVH1* locus. Treatment of the *YJ* and *YJ suvh1* lines with the cytosine methylation inhibitor 5-Aza-2'-deoxycytidine compromised the effect of the *suvh1* mutation, indicating that SUVH1 functions in the DNA methylation pathway. However, the results of McrBC treatment and genome-wide methylome profiling data revealed that the *suvh1* mutation did not lead to changes in

DNA methylation levels; thus, SUVH1 may function downstream of DNA methylation. ChIP analyses of various repressive and active marks showed that the *suvh1* mutation led to decreased H3K4me3 levels, with no changes observed for H3K9me2. The present findings suggest that SUVH1 counteract the effect of DNA methylation through H3K4me3 to promote gene expression.

Results

Two reporter lines with a *LUC* gene driven by the double 35S promoter

To identify new factors in DNA methylation and transcriptional gene silencing (TGS), particularly negative factors, two reporter lines with a *LUC* gene driven by the dual cauliflower mosaic virus *35S* promoter (*d35S*) was employed in our lab in forward genetic screens. To avoid the posttranscriptional silencing of the transgenes, both transgenes were introduced into the *rdr6-11* background. One of the reporter lines is named *LUCH* (*LUC* repressed by CHH methylation), in which high levels of DNA methylation and small RNAs are present at the *d35S* promoter (Won et al. 2012). In *LUCH, LUC* expression is strongly de-repressed by decreased DNA methylation in RdDM mutants such as *ago4, drd1*, *nrpe1*, *drm2*, and further suppressed by increased DNA methylation in a *ros1* mutant (Won et al. 2012). The other line is named *YJ*, where *LUC* is also driven by a *d35S* promoter but the site of transgene insertion in *YJ* is different from that in *LUCH*. Although similar levels of DNA methylation are detected at *d35S* promoter in *YJ* and *LUCH, LUC* expression levels are much higher in *YJ* than in *LUCH*. When *LUC* expression was determined in RdDM mutants and a *ros1* mutant, no

de-repression was observed in RdDM mutants (such as *drd1*, *ago4*, and *nrpe1*) but a suppression of *LUC* expression was observed in the *ros1* mutant (unpublished results), which suggests that *LUC* in *YJ* is regulated by DNA methylation but not by the CHH methylation maintained by the RdDM pathway. To further examine the effect of DNA methylation on *LUC* expression in the two reporter lines, the plants were grown on media containing the cytosine methylation inhibitor 5-aza-2'-deoxycytidine. LUC expression in all of the plants including *YJ*, *LUCH*, *LUCH ago4*, was increased by 5-aza-2'-deoxycytidine treatment (unpublished results). In conclusion, the *LUC* transgene in both reporter lines is under repression by DNA methylation, with *LUC* in *LUCH* being sensitive to CHH methylation and *LUC* in *YJ* not sensitive to CHH methylation.

Identification of a *suvh1* mutant involved in the DNA methylation pathway

To identify new factors in DNA methylation and transcriptional gene silencing (TGS), particularly negative factors, the YJ line was treated with ethyl methanesulfonate (EMS) for a forward genetic screen. A mutant exhibiting reduced luminescence was isolated, and qRT-PCR confirmed the reduced expression of the transgene (Figure 3.1A). Traditional map-based cloning revealed a G to A mutation that caused a Q to E substitution in the SET domain of SUVH1 (Figure 3.1B). A wild-type *SUVH1* genomic fragment introduced into this mutant completely rescued the reduced *LUC* expression in 19 out of 20 T2 transgenic lines (Figure 3.1A, data not shown), thereby confirming that the *suvh1* mutation was responsible for the observed decrease in *LUC* transcripts. This mutation was designated *suvh1-1* and is hereafter referred to as *suvh1*. The equal expression of

SUVH1 in *YJ* and *YJ suvh1* indicated that the *suvh1* mutation affects *SUVH1* function at the protein level (Figure 3S.1). The introduction of the *suvh1* mutation into *LUCH* also led to decreased *LUC* expression; thus, the *suvh1* mutation decreased *LUC* expression in both the *YJ* and *LUCH* backgrounds (Figure 3.1C). These studies show that SUVH1 is required for the expression of two transgenes. This was unexpected as three other SUVH genes, SUVH4, SUVH5, and SUVH6 belonging to another subgroup, are required for gene silencing.

To determine whether SUVH1 regulates *LUC* expression through the DNA methylation pathway, *LUC* expression levels were analyzed in *YJ suvh1*, *LUCH suvh1* and control plants (*YJ* and *LUCH*, respectively) treated with the cytosine methylation inhibitor 5-aza-2'-deoxycytidine. Luminescence imaging and qRT-PCR revealed that the decreases in *LUC* expression observed with *suvh1* were completely eliminated in both the *YJ* and *LUCH* backgrounds following chemical treatment (Figures 3.1D and 3.1E). Therefore, eliminating the DNA methylation of the *LUC* reporter gene completely suppressed the *suvh1* phenotype, indicating that SUVH1 functions through the DNA methylation pathway.

The *suvh1* mutation does not affect DNA methylation

The next question addressed was whether the *suvh1* mutation leads to increased DNA methylation levels. First, the methylation level was analyzed at the *LUC* transgene using a qPCR-based assay. DNA was digested by the methylation-sensitive restriction enzyme McrBC that cleaves methylated DNA, and realtime PCR was performed on the digested

DNA. Weaker PCR bands are expected at hypermethylated regions following McrBC treatment. Surprisingly, despite the drastic decrease in *LUC* expression in both *YJ suvh1* and *LUCH suvh1*, no differences were observed for the methylation levels at the double *35S* promoter when comparing *YJ* to *YJ suvh1* or *LUCH* to *LUCH suvh1* (Figures 3.2A and 3.2B). For the *LUC* coding region, methylation levels were low in both *YJ* and *LUCH*, and increased DNA methylation was not observed in the *suvh1* mutant (Figures 3.2A and 3.2B). To further assess whether SUVH1 influences DNA methylation levels, MethylC-seq was performed to interrogate the status of DNA methylation at the genomic scale. Two biological replicates were performed for *YJ* and *YJ suvh1*; the bisulfite conversion efficiency and coverage are listed in Tables 3.1 and 3.2. The methylation levels of the double *35S* promoter and *LUC* were determined. As shown in Figures 3.2C and 3.2D, there were no methylation level differences at either the highly methylated double *35S* promoter or the unmethylated *LUC* coding region when comparing *YJ* and *YJ suvh1*. These results confirmed that the decreased *LUC* expression observed in the *suvh1* mutant was not attributable to increased DNA methylation, indicating that SUVH1 functions downstream of DNA methylation.

We next examined whether SUVH1 influences DNA methylation at endogenous loci. No significant changes in the levels of DNA methylation at the genome-wide scale were found when comparing *YJ* and *YJ suvh1* (Figure 3.2E). To determine whether SUVH1 influences DNA methylation at a subset of genomic loci, differentially methylated regions (DMRs) between *YJ* and *YJ suvh1* were identified. There were 144, 4 and 314 CG, CHG and CHH DMRs, respectively, with reduced DNA methylation, and

274, 80 and 276 CG, CHG and CHH DMRs, respectively, with increased DNA methylation in *YJ suvh1* as compared to *YJ*. In light of the total number of regions analyzed (1196682 regions, of which 252111, 136201 and 142622 are CG, CHG and CHH methylated regions, respectively), the possibility that the identified DMRs reflected random noise was considered. Specifically, the DMRs obtained in the present study were compared to the DMRs previously obtained by another group (Stroud et al. 2013) to identify overlapping DMRs. In their study, they used a salk line with T-DNA insertion (SALK_003675) in the exon of SUVH1. The analysis eliminated most of the DMRs identified in the present study, leaving only 12, 1 and 10 hypo CG, CHG and CHH DMRs, respectively, and 10, 16 and 66 hyper CG, CHG and CHH DMRs, respectively. Moreover, correlation analysis of the methylation levels in *YJ* and *YJ suvh1* was performed. As shown in Figure 3S.2, there was a tight linear correlation between *YJ* and *YJ suvh1* when levels of methylated CG, CHG and CHH were examined. Taken together, the McrBC and methylome profiling data indicate that *suvh1* does not affect DNA methylation levels either at the *LUC* region or on a genome-wide scale. Instead, the effect of *suvh1* on *LUC* expression may have reflected activity downstream of DNA methylation.

The *suvh1* mutation causes decreased H3K4me3 levels without affecting H3K9me2 levels

Because SUVH1 is a member of the H3K9me2 methyltransferase family, the effect of the *suvh1* mutation on H3K9me2 levels was analyzed. In *YJ*, H3K9me2 was found at the double *35S* promoter but not at the *LUC* coding region. The *suvh1* mutation did not result

in a significant increase in H3K9me2 levels at the *LUC* coding region or the double *35S* promoter (Figure 3.3A). This indicates that SUVH1 does not impact H3K9me2 levels. We next examined the status of histone modifications associated with gene expression, namely, histone acetylation marks and H3K4me3. The decreased *LUC* expression in *YJ suvh1* was not accompanied by decreased H3K9Ac or H3K14Ac levels (Figure 3.3C). For H3K4me3, no differences were observed in the double *35S* promoter region, but there was a significant decrease in the *LUC* coding region (Figure 3.3B). These results suggest that SUVH1 promotes *LUC* expression through H3K4me3, either directly as an H3K4me3 methyltransferase or by affecting the function of H3K4me3 methyltransferases. Alternatively, the reduced H3K4me3 levels are a consequence of reduced LUC expression in *suvh1*.

SUVH1 has an anti-silencing role at certain endogenous loci

In light of the anti-silencing function of SUVH1 on transgenic *LUC* expression, its effect on the expression of endogenous loci was also investigated. Specifically, mRNA-seq libraries were constructed to profile the transcriptomes of *YJ* and *YJ suvh1*. To identify differentially expressed genes, the fold change between *YJ* and *YJ suvh1* RPKM-normalized read abundance was calculated (where RPKM indicates reads per kilobase of a gene per million mapped reads), and the p-value was calculated using the Poisson distribution (Marioni et al. 2008). Considering the effect of noise on the calculations based on the genomic data, different combinations of p-values and fold changes were considered when assessing the effect of the *suvh1* mutation (Table 3.3). Regardless of the

cutoff used, the number of genes with decreased transcript levels always exceeded the number of genes with increased transcript levels as a result of the *suvh1* mutation. To analyze the effect of the *suvh1* mutation on transcripts located at the intergenic regions, the genome was divided into 500 bp static windows, and transcript level comparison was performed for each window. As shown in Table 3.3, the predominant effect of the *suvh1* mutation was decreased expression. To validate the library data, eight loci were selected (six genes and two un-annotated transcripts) and analyzed using qRT-PCR (Figure 3.4A). At four of the eight loci (three genes and one un-annotated transcript), decreased transcript levels were consistently detected in *YJ suvh1*. Moreover, these four loci were tested in the *LUCH* background, and decreased expression was consistently observed with the *suvh1* mutation (Figure 3S.3). These results suggest that SUVH1 promotes the expression of certain endogenous genes.

SUVH1 may promotes gene expression at DNA-methylated loci through H3K4me3

To follow up on the finding that the role of SUVH1 in promoting *LUC* expression involved the maintenance of H3K4me3 levels, the DNA and histone methylation of the four confirmed endogenous loci were also assessed. The whole-genome methylome data were used to determine the DNA methylation levels at the four endogenous loci. As shown in Figure 3.4B and Figures 3S.4A, 3S.4B and 3S.4C, the promoter regions of the endogenous loci exhibited high levels of DNA methylation that remained unchanged in *YJ suvh1*, consistent with the observation for the double *35S* promoter and *LUC* transgene. H3K4me3 and H3K9me2 ChIP assays were performed to assess the histone methylation

levels of the endogenous loci. Although H3K9me2 and H3K4me3 occurred at the promoter regions, no differences were observed between *YJ* and *YJ suvh1* (Figure 3.4C). The coding regions contained almost no H3K9me2. However, decreased H3K4me3 levels in the coding regions were detected in *YJ suvh1* (Figure 3.4C). Considering the presence of an SRA domain and a SET domain in SUVH1, these results support a role of SUVH1 in promoting H3K4me3 at regions with DNA methylation.

The genetic relationships between *SUVH1* and DNA methylation factors

The findings that SUVH1 functions at genes with DNA-methylated promoters prompted the question of how SUVH1 is related to the RdDM pathway. It has been proposed that NRPE1, the largest Pol V subunit, produces non-coding scaffold transcripts that recruit the AGO4-siRNA complex (Wierzbicki et al. 2008; Wierzbicki et al. 2009). With the mutations in *NRPE1*, the RdDM pathway is disrupted and CHH methylation cannot be maintained (Stroud et al. 2013); in contrast, CHG methylation and CG methylation are virtually unaffected. To determine whether the SUVH1-targeted loci are regulated by RdDM and whether CHH methylation is required for SUVH1 function, qRT-PCR was performed to detect the transcript levels of the SUVH1-targeted loci in *YJ nrpe1* and *YJ suvh1 nrpe1*. In *YJ nrpe1*, in which CHH methylation cannot be maintained, the expression of SUVH1-targeted loci was de-repressed (Figure 3.5A), indicating that these loci are also under the regulation of RdDM. The expression levels of the SUVH1-targeted loci in *YJ suvh1 nrpe1* were greatly reduced relative to *YJ nrpe1* (Figure 3.5A), indicating

that loss of CHH methylation does not alleviate the requirement for SUVH1 in the expression of these genes.

The DNA glycosylase/lyase ROS1 is a DNA demethylase (Agius et al. 2006), and *ros1* mutants exhibit increased DNA methylation (at CG, CHG and CHH) (Lister et al. 2008; Stroud et al. 2013). The transcript levels of the SUVH1-targeted loci were examined in *YJ ros1* by qRT-PCR to determine whether they are regulated by ROS1. The results showed decreased transcript levels for the four loci in *YJ ros1* relative to the *YJ* control (Figure 3.5B), indicating that the SUVH1-targeted loci are also regulated by the ROS1 demethylation pathway. Next, the transcript levels of the SUVH1-targeted loci were examined in the *YJ suvh1 ros1* double mutant to determine whether ROS1 and SUVH1 function in the same pathway. Decreased transcript levels were observed for the SUVH1-targeted loci in *YJ suvh1 ros1* compared to *YJ ros1*. Additionally, *SUVH1* transcript levels were unchanged in the RdDM mutants (Figure 3S.1), which contrasts the decreased expression of ROS1 pathway factors when the RdDM is disrupted (Huettel et al. 2006; Qian et al. 2012). These results suggest that ROS1 and SUVH1 are not in the same pathway, which is consistent with the previous finding that the *suvh1* mutation does not alter DNA methylation levels.

Lack of anti-correlation between promoter DNA methylation and gene expression

DNA methylation is an important mechanism for suppressing the expression of transposons and is established through the RdDM pathway. A possible consequence of transposon insertion into the promoter of a gene is suppression of the gene through DNA

methylation. Using existing methylome and gene expression datasets, we explored whether there is any anti-correlation between gene expression levels and promoter DNA methylation. We determined the DNA methylation level at 1 kb regions upstream of genes from methylome data and derived the corresponding gene expression levels from mRNA-seq data. As shown in Figure 3S.5, DNA methylation levels tended to be low in gene promoter regions, regardless of whether CG, CHG or CHH methylation was considered, and there was no strong anti-correlation between gene expression and promoter DNA methylation level. Genes with or without DNA methylation in their promoters were found to have high, medium or low expression levels. Despite the role of DNA methylation in suppressing gene expression, genes with DNA methylation at the promoter region are not necessarily suppressed, indicating that some regulatory mechanism must exist to override this suppressive mark.

Discussion

Since the initial discovery of transposons by Barbara McClintock, the regulation of transposons has been widely investigated. DNA methylation is a well-recognized epigenetic mark for the suppression of transposon transcription, and numerous effectors involved in the DNA methylation pathway, from initial establishment to maintenance, have been characterized. However, the understanding of opposing mechanisms and the negative regulation of DNA methylation is very limited. In the present study, a forward-genetic screening approach was used to identify negative regulators with an anti-silencing function. *SUVH1*, which encodes a SET-domain protein, was identify and found to

promote the expression of reporter gene expression only when their promoters harbor DNA methylation.

Although DNA methylation deposition has been well studied, subsequent processes downstream of DNA methylation function have not been as thoroughly explored. At present, there are two known types of conserved domains capable of binding methylated DNA: the SET and RING-associated (SRA) domain (Rajakumara et al. 2011) and the METHYL-CpG-BINDING domain (MBD) (Fournier et al. 2012). In animals, MBD proteins have been implicated in the establishment of repressive chromatin marks through the promotion of histone deacetylase and histone methyltransferase activity (Jones et al. 1998; Nan et al. 1998; Zhang et al. 1999; Fuks et al. 2003a). One family of SRA proteins, the RING-associated VARIANT IN METHYLATION (VIM)/ORTHRUS (ORTH) family and their homologs in animals, Ubiquitin-like PHD and RING finger domain (UHRF1), have all been found to be critical for DNA methylation maintenance through binding methylated CG sites (Rajakumara et al. 2011). The SU(var)3-9 homologs, which constitute another family of SRA proteins, are associated with the SET domain. Several SRA proteins have been shown to have H3K9me2 methyltransferase activity or to participate in the RdDM pathway (Rea et al. 2000; Naumann et al. 2005). Ultimately, all of these DNA-methylation-associated proteins function in connecting DNA methylation to repressive chromatin marks (namely, DNA methylation, H3K9me2 and histone deacetylation). In contrast, SRA proteins have not been associated with active chromatin marks or gene silencing suppression.

In the present study, a mutation leading to decreased reporter expression was mapped to the *SUVH1* locus, indicating that SUVH1 promoted *LUC* expression. This contradicts the known roles of SUVH homologs, which have been found to regulate gene expression by promoting silencing (Naumann et al. 2005; Rajakumara et al. 2011; Johnson et al. 2014). Thus, a loss of function *suvh* mutant would be predicted to exhibit high *LUC* expression. The low *LUC* expression in *YJ suvh1* suggests that SUVH1 has a different role than its homologs with currently known functions. Given that none of the SUVH1 subgroup homologs have been associated with silencing roles, this raises the possibility that this particular subgroup is characterized by anti-silencing functions. ChIP analysis of histone modification levels did not reveal any changes in H3K9me2 abundance in the *suvh1* mutant, providing a second line of evidence that SUVH1 function may be distinct from those of other SUVH proteins associated with RdDM or H3K9me2. The decreased levels of H3K4me3 in *YJ suvh1* suggest that SUVH1 may regulate H3K4me3 abundance either directly as an H3K4me3 methyltransferase or by affecting the functions of other H3K4me3 methyltransferases. The fact that the *suvh1* mutation leads to an amino acid substitution in the SET domain without affecting *SUVH1* transcript levels raises the possibility that SUVH1 functions as an H3K4me3 methyltransferase.

Among the SUVH1-targeted loci, Pol IV-dependent siRNAs were detected at the promoter regions along with CG, CHG and CHH methylation and transposons (Figure 3.6). A model for SUVH1 function is proposed based on the present findings. With transposons inserting into different positions in the genome over the course of evolution,

Pol IV-generated siRNAs function as guides directing DNA methylation at the sites of insertion to inhibit the harmful effects of the active transposon. While this is necessary for genome stability, this silencing mechanism could cause a gene to be suppressed if a transposon inserts into its promoter region. To counteract this suppression, however, SUVH1, a protein with a DNA methylation binding domain and a histone methylation domain, is recruited to these loci to promote gene expression through promoting H3K4me3 levels. The proposed regulatory model is novel insofar as the effects of both repressive and active marks at a given locus are jointly considered.

Materials and Methods

Plant materials

All tissues used in the present study were from 8- to 10-day-old seedlings, and all *Arabidopsis* strains were in the Columbia ecotype. The reporter lines *LUCH* (Won et al. 2012) and *YJ* are in the *rdr6-11* mutant background (Peragine et al. 2004b). *ros1-5*, *ago4-6* and *drd1-12* were isolated in the *LUCH* background (Won et al. 2012) and subsequently introduced into *YJ* and *YJ suvh1*. *nrpe1-1* was described previously (Kanno et al. 2005) and was also introduced into *YJ* and *YJ suvh1*.

RT-PCR

Total RNA was extracted from seedlings with Trizol (Invitrogen, 15596-018) then treated with DNase I (Roche, 04716728001). cDNA was synthesized using oligo-dT primers and

RevertAid Reverse Transcriptase (Fermentas, EP0442). qRT-PCR was performed with three technical replicates on a Bio-Rad C1000 thermal cycler equipped with a CFX detection module using iQ™ SYBR® (Bio-Rad, 170-8880). The primers used in the study are listed in Table 3.4.

Luciferase live imaging and 5-Aza-2'-deoxycytidine treatment

For luciferase live imaging, 8- to 10-day-old seedlings growing on plates with half-strength Murashige and Skoog (MS) media supplemented with 8% agar and 1% sucrose were sprayed with 1 mM luciferin (Promega) in 0.01% Triton X-100. After a 5 min incubation in the dark, the plants were placed in a Stanford Photonics Onyx Luminescence Dark Box equipped with a Roper Pixis 1024B camera controlled by WinView32 software then imaged with a 1 min exposure time. For 5-Aza-2'-deoxycytodine (5-aza-2'-dC) (Sigma, A3656) treatment, plants were grown in MS media with 7 μg/ml 5-aza-2'-dC for 2 weeks.

EMS mutagenesis of the YJ line

A 1 ml volume of seeds (around 10,000 seeds) was pre-washed with 0.1% Tween 20 for 15 min then treated with 0.2% EMS for 12 h, followed by three washes with 10 ml water for 1 h with gentle agitation. The M0 seedlings were planted in soil to obtain the M1 seeds. Mutants with reduced *LUC* activity, based on *LUC* live imaging, were isolated in the M2 generation. The isolated mutants were backcrossed to the parental line (*YJ*) two times prior to further analysis.

Mapping of the *suvh1-1* mutation

To identify genes responsible for low *LUC* expression, the mutants were crossed to *YJ* in the L*er* background to generate the mapping populations. The F2 mapping populations were used to narrow the mapping regions. For the mapping of *YJ suvh1*, a 44 kb region encompassing 11 genes on Chromosome 5 was further narrowed using a combination of SSLP and dCAPS markers. Sequencing of AT5G04940 revealed a G to A mutation resulting in a Q to E amino acid substitution in the SET domain.

Plasmid construction

To generate the *SUVH1:SUVH1-3XFLAG* transgene, the *SUVH1* coding region including 1.5 kb of the endogenous promoter region and lacking the stop codon was amplified from *YJ* genomic DNA and cloned into the PJL-Blue entry vector. The genomic fragment was then introduced into a binary vector containing a PEG301 backbone and a C-terminal 3XFLAG tag using Gateway® LR Clonase® Enzyme mix (Invitrogen, Cat 11791-019).

McrBC-PCR

Genomic DNA was extracted using the CTAB method (Rogers and Bendich 1985), and ribonuclease A (Sigma, R4875-100MG) was used to digest the RNAs. A volume containing 100 ng DNA was treated with 2 units of McrBC (New England Biolabs, M0272S) at 37°C for 30 min, and a mix without McrBC was performed in parallel as the control. The mixtures were incubated at 65°C for 20 min to inactivate the McrBC. qPCR was performed using iQ™ SYBR® (Bio-Rad, 170-8880) to quantify the remaining DNA,

with the ratio between the McrBC mix and the mix without McrBC as an indicator of the methylation level. *UBQ5*, which lacks DNA methylation, was used as a control.

Bisulfite sequencing library construction

To generate the whole-genome bisulfite sequencing (BS-seq) libraries, genomic DNA was extracted using the DNeasy Plant Mini Kit (Qiagen, 69104) and quantified using a Qubit fluorometer. One microgram of genomic DNA was sonicated into fragments 100 to 300 bp in length using a Diagenode Bioruptor for four cycles with the following parameters: intensity = high, on = 30 s, off = 30 s and time = 15 min. The sonicated DNA fragments were purified using the PureLink PCR Purification Kit (Invitrogen, K3100-01). End repair was performed at room temperature for 45 min using the End-It™ DNA End-Repair Kit (Epicenter, ER0720), with the substitution of the dNTP with a mixture of dATP, dGTP and dTTP. Following the incubation, the Agencourt AMPure XP-PCR Purification system (Beckman Coulter, A63881) was used for purification. 3'-end adenylation was performed at 37°C for 30 min using dATP and Klenow Fragment (3'→5' exo-) (New England Biolabs, M0212), followed by purification using the Agencourt AMPure XP-PCR Purification system. The purified DNA was ligated with methylated adapters from the TruSeq DNA Sample Preparation Kit (Illumina, FC-121-2001) at 16°C overnight using T4 DNA ligase (New England Biolabs, M0202). The ligation products were purified with AMPure XP beads twice. Less than 400 ng ligated product was used for bisulfite conversion using the MethylCode Kit (Invitrogen, MECOV-50) according to the manufacturer's guidelines, except for the addition of 12 μg

carrier RNA (Qiagen, 1068337) to the conversion product before column purification. The final conversion product was amplified using Pfu Cx Turbo (Agilent, 600414) under the following PCR conditions: 2 min at 95°C; 9 cycles of 15 s at 98°C, 30 s at 60°C and 4 min at 72°C; and 10 min at 72°C. The PCR product was purified using AMPure XP beads prior to a 101-cycle sequencing run (single end) on an Illumina HiSeq 2000.

Data analysis of the BS-seq libraries

The raw reads that passed the Illumina quality control steps were retained, and duplicated reads were removed prior to mapping. The reads were mapped to the TAIR10 genome using BS Seeker (Chen et al. 2010), and in-house R and Perl scripts were employed to convert the BS Seeker-aligned reads to every cytosine. DMRs (differentially methylated regions) were calculated according to previously described methodology (Stroud et al. 2013). The *Arabidopsis* genome was divided into 100 bp windows, and the methylation level at each window was calculated separately. The methylation level was defined as the number of methylated cytosines sequenced divided by the total number of cytosines sequenced. To avoid the skew caused by few cytosines and low coverage, only windows with at least four cytosines covered by at least four reads were counted. Windows with an absolute methylation difference greater than 0.4 (CG), 0.2 (CHG) and 0.1 (CHH) and an adjusted p-value (FDR) < 0.01 (Fisher's exact test) were considered DMRs. DMRs identified from both replicates of *YJ* and *YJ suvh1* were considered SUVH1 DMRs.

Chromatin immunoprecipitation (ChIP)

The ChIP experiments were performed as previously described (Gendrel et al. 2005) using H3K4me3 (abcam, ab8580) and H3K9me2 (abcam, ab1220) antibody.

mRNA-seq library construction and data processing

Ten-day-old seedlings from *YJ* and *YJ suvh1* were collected for RNA extraction using Trizol (Invitrogen, 15596-018), and the extracted RNA was treated with DNase I (Roche, 04716728001). Two micrograms of the DNase I-treated RNA and the TruSeq RNA Sample Preparation Kit v2 (Illumina, FC-122-1002) were used for library construction. The libraries were sequenced on an Illumina HiSeq 2000.

The raw reads that passed the Illumina quality control steps were collapsed into a set of non-redundant reads. These non-redundant reads were mapped to the TAIR10 *Arabidopsis* genome using TopHat v2.0.4 with default settings (Kim et al. 2013). For the quantification of a given gene or window, reads whose 5' ends were within the gene or window were counted. The fold change was calculated using the RPKM-normalized read values, and the p-value was calculated based on the Poisson distribution (Marioni et al. 2008).

Figures

Figure 3. 1 Identification of a *suvh1* mutant affecting the DNA methylation pathway. (A) The *suvh1* mutation led to decreased expression of the luciferase gene (*LUC*) in the *YJ* background. *YJ* SUVH1-FLAG indicates the *YJ SUVH1:SUVH1-3XFLAG suvh-1* line. In *YJ* SUVH1-FLAG, the phenotype of the *YJ suvh1* mutant was rescued by a transgene containing a wild-type *SUVH1* genomic region and a 3XFLAG tag at the C-terminal end. (Left panel) *LUC* luminescence of 8-day-old *YJ*, *YJ suvh1* and *YJ* SUVH1-FLAG seedlings grown on MS media. (Right panel) qRT-PCR revealed decreased *LUC* transcript levels in the *suvh1* mutant in the *YJ* background. Three biological replicates were performed. (B) A diagram of the SUVH1 protein and the substitution caused by the *suvh1-1* (*suvh1*) mutation. The SUVH1 protein contains an SRA domain, a Pre-SET domain and a SET domain. The G to E substitution caused by *suvh1* occurs in the SET domain. (C) The *suvh1* mutation led to decreased *LUC* expression in the *LUCH* background. (Left panel) *LUC* luminescence of 8-day-old *LUCH* and *LUCH suvh1* seedlings grown on MS media. (Right panel) qRT-PCR showed decreased *LUC* expression in the *suvh1* mutant in the *LUCH* background. Three biological replicates were performed. (D-E) The decreased *LUC* expression phenotype associated with *suvh1* was suppressed by 5-Aza-2'-deoxycidine treatment in both the *YJ* (D) and *LUCH* (E) backgrounds. (Left panels) *LUC* luminescence of seedlings grown on MS media with 7 μg/ml 5-Aza-2'-deoxycidine for 14 days for *YJ* and *YJ suvh1* (D) and *LUCH* and *LUCH suvh1* (E). (Right panels) qRT-PCR showed rescued *LUC* transcript levels in the treated

YJ suvh1 (D) and *LUCH suvh1* (E) seedlings. Three biological replicates were performed. All of the luciferase images were captured using a CCD camera.

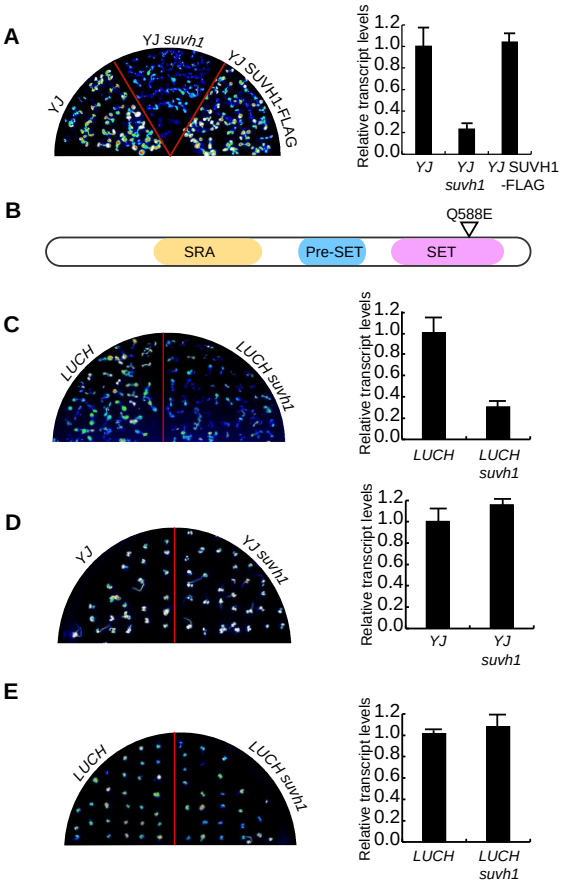

A
YJ suvh1
YJ SUVH1-FLAG
YJ
Relative transcript levels
1.2
1.0
0.8
0.6
0.4
0.2
0.0
YJ
YJ suvh1
YJ SUVH1 -FLAG
B
Q588E
SRA
Pre-SET
SET
C
LUCH
LUCH suvh1
Relative transcript levels
LUCH
LUCH suvh1
D
YJ
YJ suvh1
Relative transcript levels
YJ
YJ suvh1
E
LUCH
LUCH suvh1
Relative transcript levels
LUCH
LUCH suvh1

Figure 3. 2 The *suvh1* mutation does not affect DNA methylation.

(A-B) McrBC-PCR analysis of DNA methylation levels at the double *35S* promoter and the *LUC* coding region in *YJ* (A) and *LUCH* (B). qPCR was performed using genomic DNA treated with or without McrBC. The relative levels of amplified transcripts for *UBQ5*, *LUC* and *35S* in samples treated with McrBC compared to untreated samples. Three biological replicates were performed. (C) The levels of CG, CHG and CHH DNA methylation level at the double *35S* promoter and *LUC* in *YJ* and *YJ suvh1* determined through whole-genome bisulfite sequencing. The results from two biological replicates are shown. (D) The whole-genome CG, CHG and CHH DNA methylation level data for *YJ* and *YJ suvh1* obtained through whole-genome bisulfite sequencing.

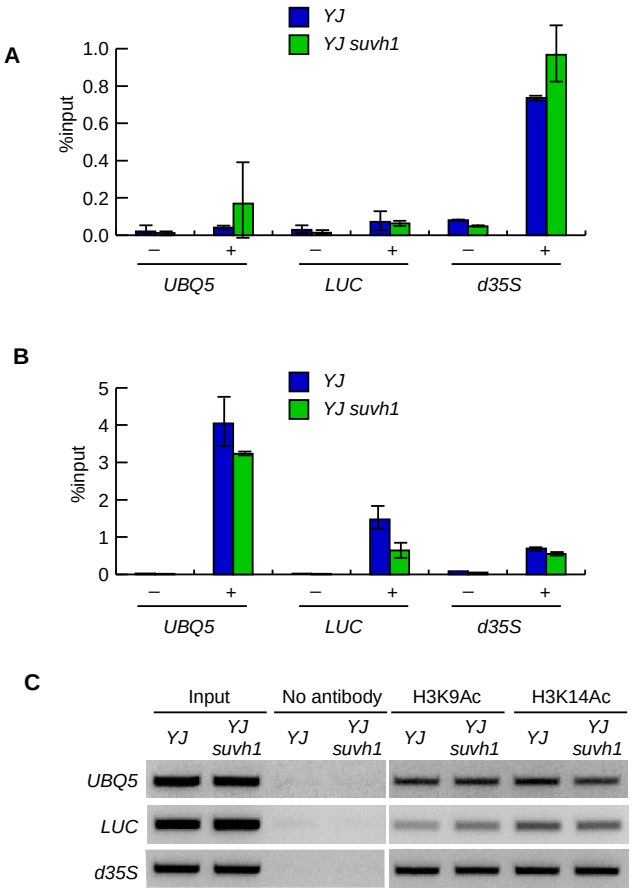

A
YJ
YJ suvh1
%input
1.0
0.8
0.6
0.4
0.2
0.0
–
+
UBQ5
LUC
d35S
B
YJ
YJ suvh1
%input
5
4
3
2
1
0
UBQ5
LUC
d35S
C
Input
No antibody
H3K9Ac
H3K14Ac
YJ
YJ suvh1
UBQ5
LUC
d35S

Figure 3. 3 ChIP analysis of histone methylation and acetylation marks in *suvh1*.

(A-B) ChIP-qPCR was performed to measure H3K9me2 (A) and H3K4me3 (B) levels in *YJ* and *YJ suvh1*. No changes in H3K9me2 levels were observed. Reduced H3K4me3 levels were observed in *YJ suvh1* at the *LUC* coding region but not at the double *35S* promoter. (C) ChIP-PCR revealed no changes in H3K9Ac or H3K14Ac levels in *suvh1*. For (A-C), *UBQ5*, whose expression level was not changed in *suvh1*, was used as a control, and three biological replicates were performed for all analyses.

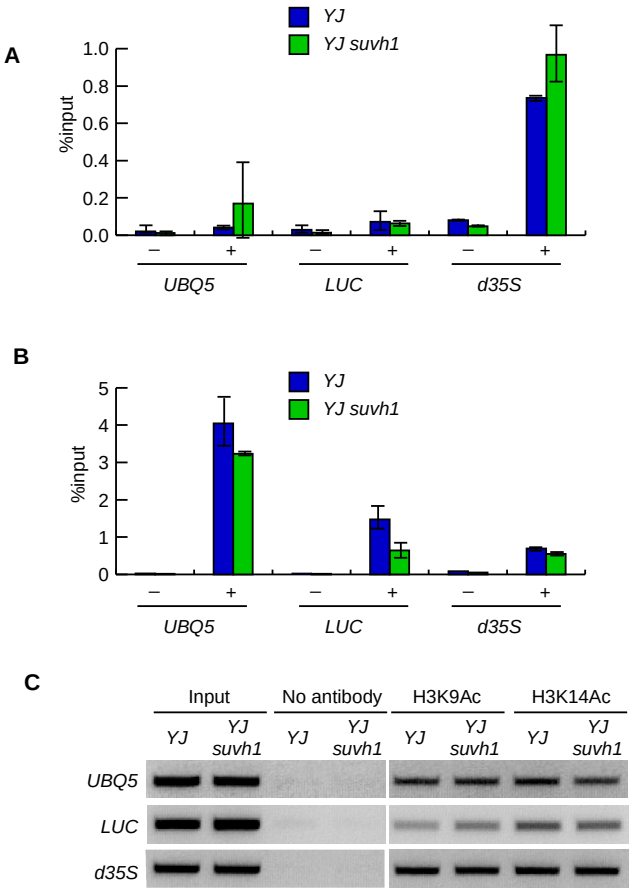

A
YJ
YJ suvh1
%input
1.0
0.8
0.6
0.4
0.2
0.0
–
+
UBQ5
LUC
d35S
B
YJ
YJ suvh1
%input
5
4
3
2
1
0
UBQ5
LUC
d35S
C
Input
No antibody
H3K9Ac
H3K14Ac
YJ
YJ suvh1
UBQ5
LUC
d35S

Figure 3. 4 The *suvh1* mutation leads to the reduced expression of endogenous loci with corresponding reductions in H3K4me2 levels.

(A) The expression of four SUVH1-targeted endogenous loci was confirmed by qPCR, and the decreased expression observed in *YJ suvh1* was rescued in *YJ* SUVH1-FLAG transgenic lines for all four loci. Three biological replicates were performed. (B) The DNA methylation level of the 1 kb promoter of locus 1 determined from the two biological replicates of the *YJ* and *YJ suvh1* methylome data. In all four libraries, CG, CHG and CHH methylation was detected, and there were no consistent differences between *YJ* and *YJ suvh1*. (C) H3K4m3 and H3K9me2 methylation levels at the four endogenous loci and their promoter regions. *UBQ5*, whose expression level was not changed in *suvh1*, was used as a control. Locus 1P, 2P, 3P and 4P refer to the promoter regions of the corresponding loci. Three biological replicates were performed.

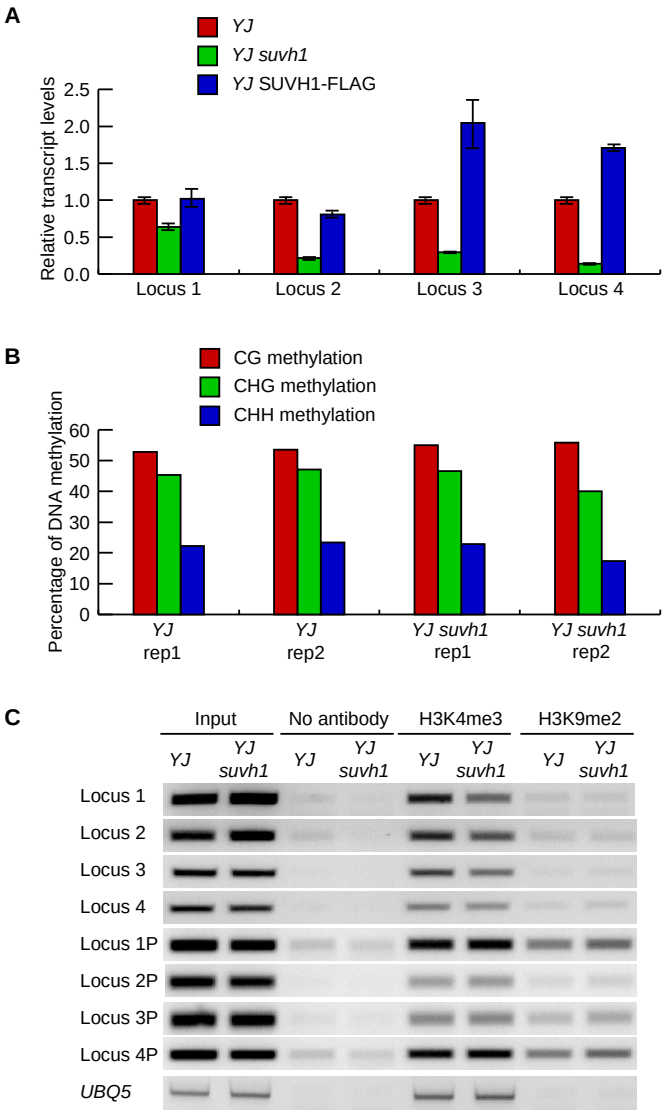

A
YJ
YJ suvh1
YJ SUVH1-FLAG
Relative transcript levels
2.5
2.0
1.5
1.0
0.5
0.0
Locus 1
Locus 2
Locus 3
Locus 4
B
CG methylation
CHG methylation
CHH methylation
Percentage of DNA methylation
60
50
40
30
20
10
0
YJ rep1
YJ rep2
YJ suvh1 rep1
YJ suvh1 rep2
C
Input
No antibody
H3K4me3
H3K9me2
YJ
YJ suvh1
Locus 1
Locus 2
Locus 3
Locus 4
Locus 1P
Locus 2P
Locus 3P
Locus 4P
UBQ5

Figure 3. 5 The expression of SUVH1-targeted loci in the *nrpe1* and *ros1* mutant backgrounds.

qPCR was used to detect the transcript levels of the four SUVH1-targeted endogenous loci in the *nrpe1* mutant background (A) and the *ros1* mutant background (B). Three biological replicates were performed for all analyses.

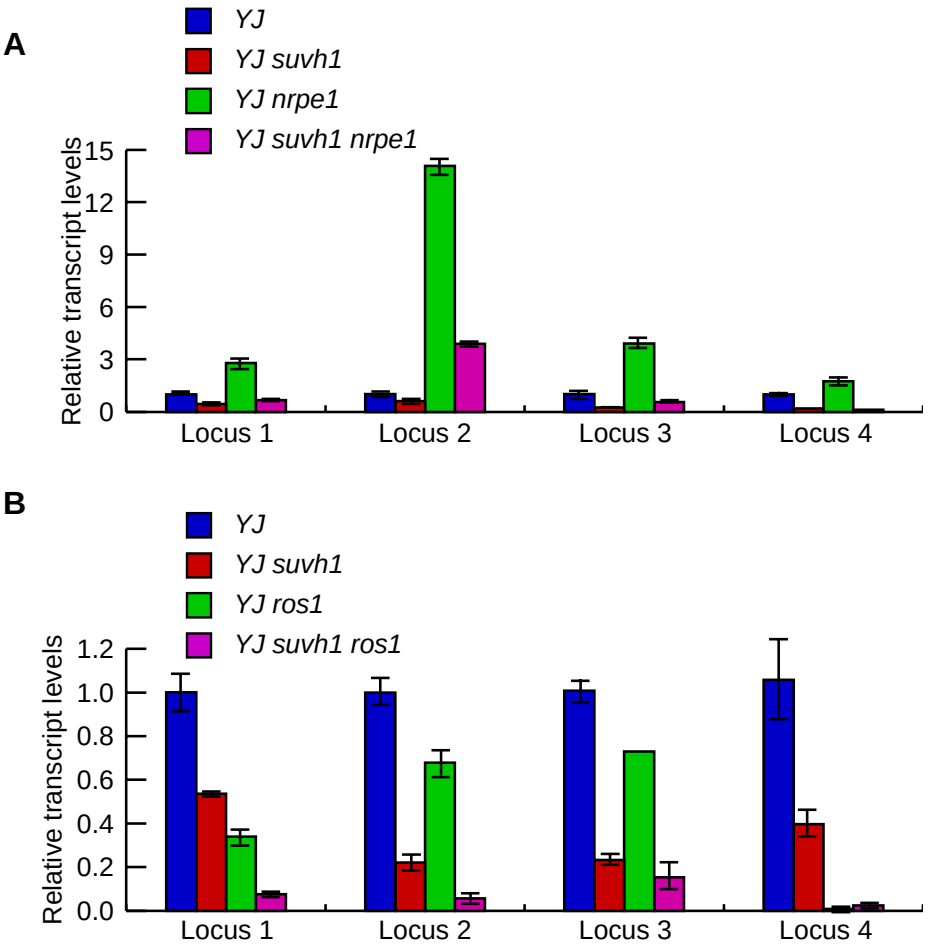


Figure 3. 6 The epigenetic modifications at a SUVH1-targeted locus.

The genome browser view of SUVH1-targeted Locus 1. The top row (row 1) is a gene model with TAIR 10 annotations, where AT1G52040 is a gene, and AT1TE64100 and AT1TE64110 are transposons. Rows 2 and 3 in green represent the reads from the mRNA-seq libraries (no strand information). Rows 4 and 5 in blue show the reads from the small RNA-seq (sRNA-seq) libraries, and read abundance is shown for both the Watson (top) and Crick (bottom) strands. Rows 6 and 7, rows 8 and 9 and rows 10 and 11 represent the CG, CHG and CHH methylation levels derived from the BS-seq libraries, respectively.

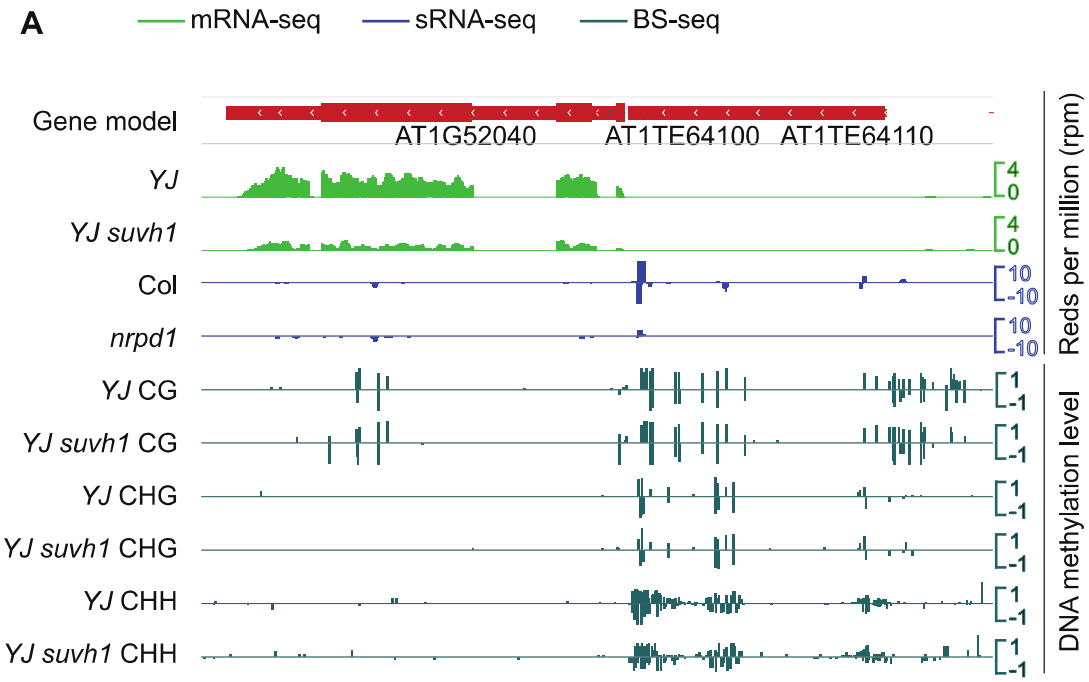


Figure 3S. 1 *SUVH1* transcript levels in various mutants.

qRT-PCR was performed in *YJ* plants with one or two of the following mutations: *suvh1*, *ago4*, *drd1* and *drd3*. Three biological replicates were performed.

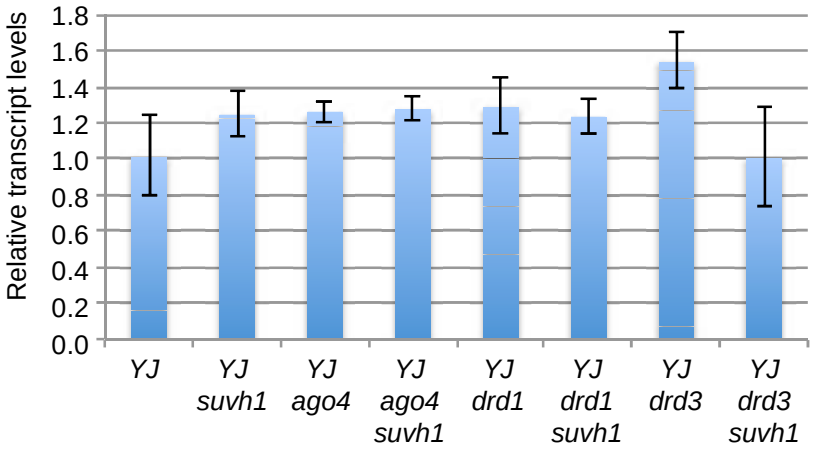


Figure 3S. 2 Correlation plots of CH, CHG and CHH DNA methylation in *YJ* and *YJ suvh1*.

Two biological replicates of whole-genome bisulfite sequencing were performed. Each spot represents the data for a 100 bp window, and for each window, the methylation level was calculated as the total methylated cytosines divided by the total sequenced cytosines.

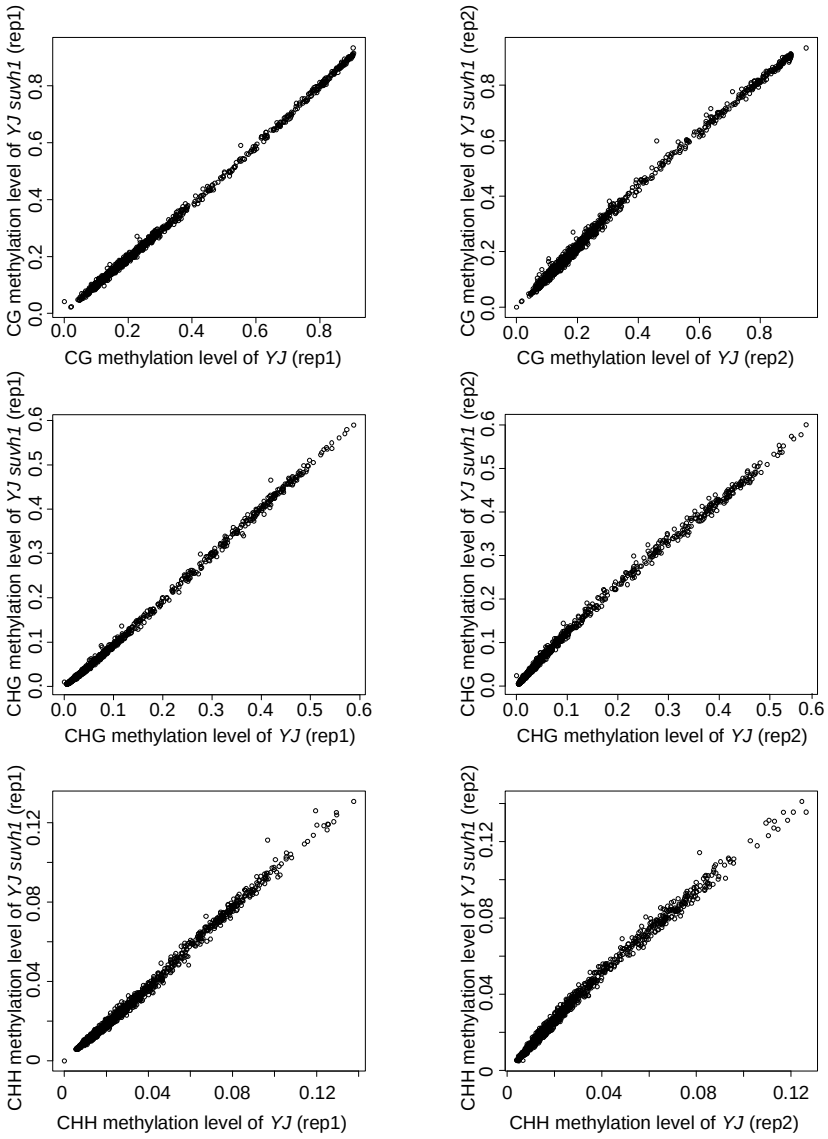

CG methylation level of YJ suvh1 (rep1)
CG methylation level of YJ (rep1)
CG methylation level of YJ suvh1 (rep2)
CG methylation level of YJ (rep2)
CHG methylation level of YJ suvh1 (rep1)
CHG methylation level of YJ (rep1)
CHG methylation level of YJ suvh1 (rep2)
CHG methylation level of YJ (rep2)
CHH methylation level of YJ suvh1 (rep1)
CHH methylation level of YJ (rep1)
CHH methylation level of YJ suvh1 (rep2)
CHH methylation level of YJ (rep2)

Figure 3S. 3 The validation of SUVH1-targeted loci in *LUCH* background.

The decreased transcript levels of the four SUVH1-targeted endogenous loci observed in *YJ suvh1* compared to *YJ* were confirmed by qPCR in *LUCH* background. Three biological replicates were performed.

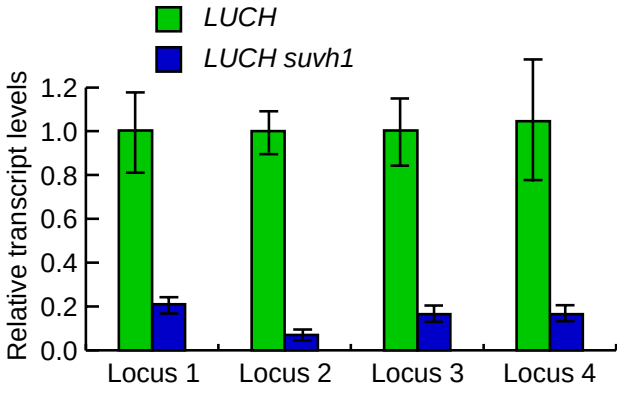


Figure 3S. 4 The DNA methylation level at the promoter of SUVH1-targeted loci.

The DNA methylation level of the 1 kb promoter regions of SUVH1-targeted loci in the two biological replicates of the *YJ* and *YJ suvh1* methylome data. (A) Locus 2. (B) Locus 3. (C) Locus 4. In all four libraries, CG, CHG and CHH methylation was present, and there were no consistent changes between *YJ* and *YJ suvh1*.

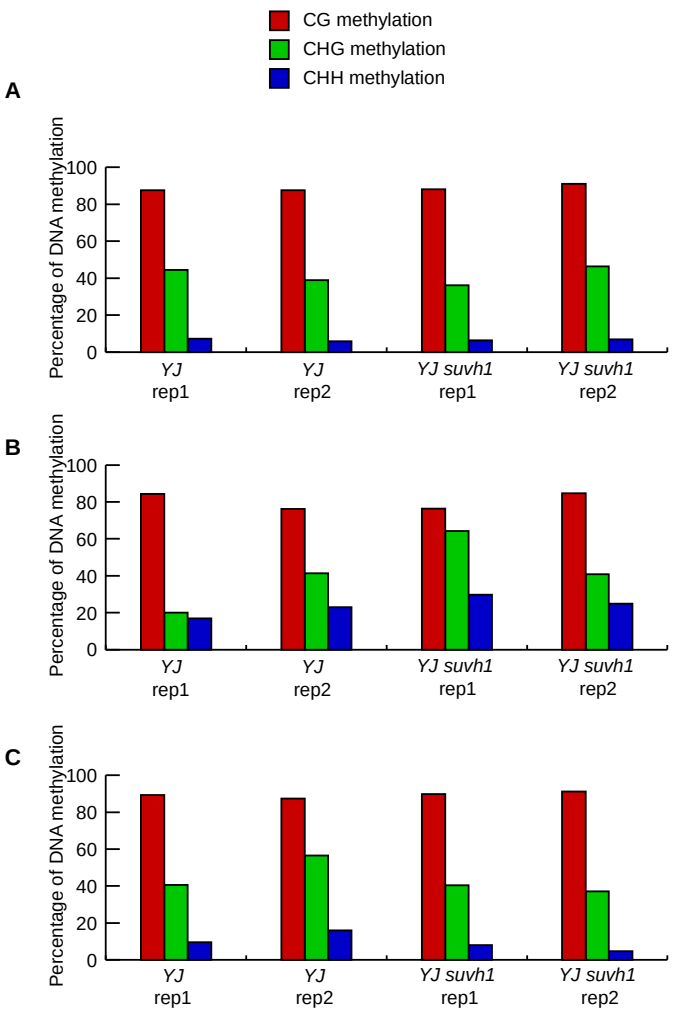
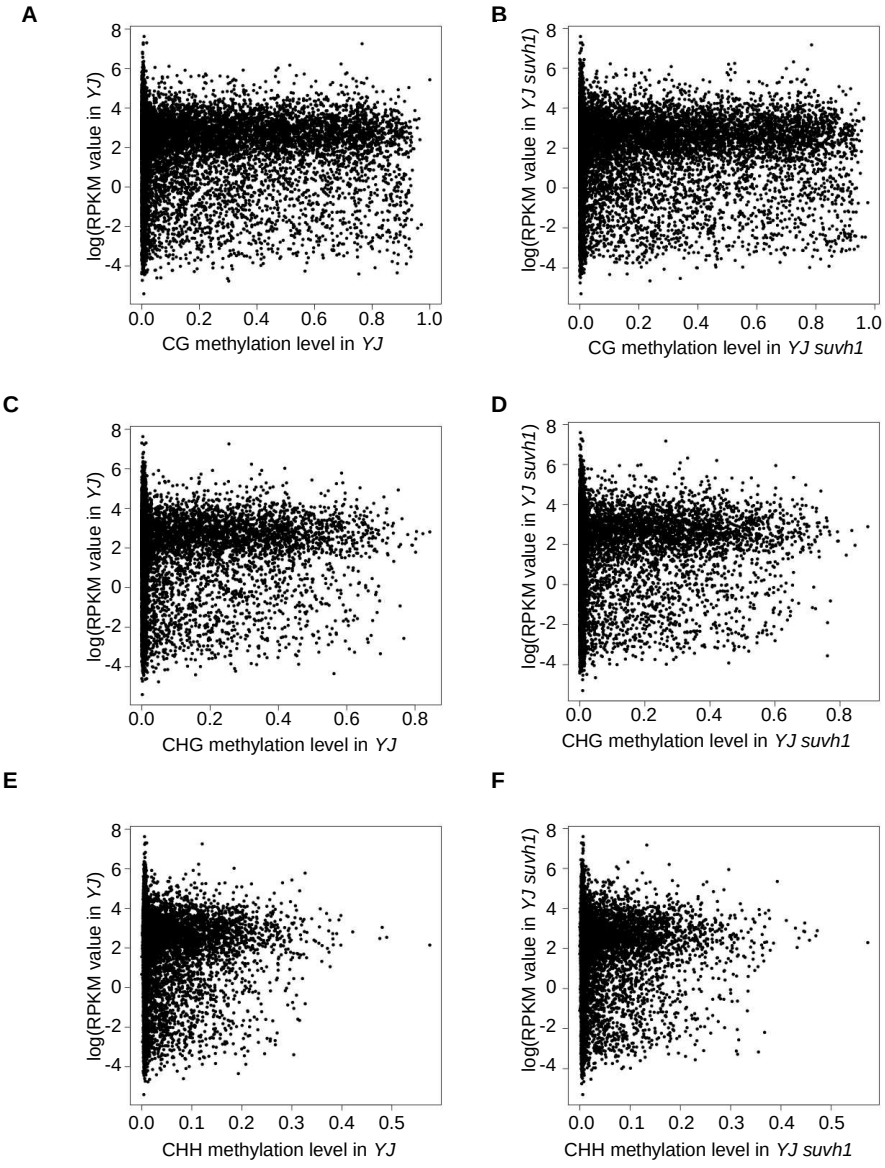

CG methylation
CHG methylation
CHH methylation
A
B
C
Percentage of DNA methylation
100
80
60
40
20
0
YJ rep1
YJ rep2
YJ suvh1 rep1
YJ suvh1 rep2

Figure 3S. 5 Correlations plots of DNA methylation level and gene expression in *YJ* and *YJ suvh1*.

The x-axis represents the level of DNA methylation, and the y-axis represents the natural logarithm of the RPKM (reads per kilobase per million) value for genes from the mRNA-seq libraries. DNA methylation level was calculated at 1 kb of the gene promoter region. (A-B) Correlation plot of CG methylation level with gene expression in *YJ* (A) and *YJ suvh1* (B). (C-D) Correlation plot of CHG methylation level with gene expression in *YJ* (C) and *YJ suvh1* (D). (E-F) Correlation plot of CHH methylation level with gene expression in *YJ* (E) and *YJ suvh1* (F).


A
log(RPKM value in YJ)
CG methylation level in YJ
B
log(RPKM value in YJ suvh1)
CG methylation level in YJ suvh1
C
log(RPKM value in YJ)
CHG methylation level in YJ
D
log(RPKM value in YJ suvh1)
CHG methylation level in YJ suvh1
E
log(RPKM value in YJ)
CHH methylation level in YJ
F
log(RPKM value in YJ suvh1)
CHH methylation level in YJ suvh1

Tables

Table 3. 1 Summary of bisulfite conversion efficiency for each bisulfite sequencing library.

	CG	CHG	CHH	Total C
YJ rep1	97.9%	97.8%	97.7%	97.7%
YJ rep2	97.7%	97.7%	97.5%	97.6%
YJ suvh1 rep1	98.4%	98.3%	98.2%	98.2%
YJ suvh1 rep2	97.9%	97.9%	97.7%	97.8%

Table 3. 2 Read coverage of the whole-genome bisulfite sequencing libraries.

CG	# of sequenced ^{m}C	# of total sequenced C	5567714 *
			Coverage[&]
YJ rep1	14910568	47465516	8.525
YJ rep2	14812260	48766749	8.759
YJ suvh1 rep1	9089283	32953841	5.919
YJ suvh1 rep2	12926127	42052472	7.553
CHG	# of sequenced ^{m}C	# of total sequenced C	6093657 **
			Coverage[&]
YJ rep1	5780812	49895644	8.188
YJ rep2	5634536	51889154	8.515
YJ suvh1 rep1	3249672	35506043	5.827
YJ suvh1 rep2	4945165	44518366	7.306
CHH	# of sequenced ^{m}C	# of total sequenced C	31198380 ***
			Coverage[&]
YJ rep1	9434110	269824938	8.649
YJ rep2	8857244	276503635	8.863
YJ suvh1 rep1	4866332	186035580	5.963
YJ suvh1 rep2	8048085	238759260	7.653
Total	# of sequenced ^{m}C	# of total sequenced C	42859751 ****
			Coverage[&]
YJ rep1	30125490	367186098	8.567
YJ rep2	29304040	377159538	8.800
YJ suvh1 rep1	17205287	254495464	5.938
YJ suvh1 rep2	25919377	325330098	7.591

*, **, *** and **** indicate the total number of CG, CHG, CHH and C sites in the genome, respectively.

[&] Coverage was calculated as the total number of methylated C divided by the total number of sequenced C in the genome.

Table 3. 3 The number of differentially expressed genes and static windows in *YJ suvh1* compared to *YJ*.

p-value	fold change	decreased*	increased*	decreased#	increased#
0.05	2	118	50	109	31
0.05	4	48	19	53	12
0.01	2	81	41	74	23
0.01	4	35	15	36	8

* indicates the number of genes with decreased or increased expression.

# indicates the number of 500 bp static windows with decreased or increased expression.

Table 3. 4 Primers used in the present study.

Name	Sequence	Purpose
35SF1	GAGCACGACACACTTGTCTAC	qRT-PCR, ChIP-PCR for the double *35S* promoter
35SR1	ATGATGGCATTTGTAGGAGC	
LUCmF5	CTCCCCTCTCTAAGGAAGTCG	qRT-PCR, ChIP-PCR for *LUC*
LUCmR5	CCAGAATGTAGCCATCCATC	
N_UBQ5	GGTGCTAAGAAGAGGAAGAAT	qRT-PCR, ChIP-PCR for the *UBQ5* promoter
C_UBQ5	CTCCTTCTTTCTGGTAAACGT	
SUVH1-NlaIVF	CCCTTTCAAGTGGAACTACG	Genotyping of *suvh1*, NlaIV cuts wild-type bands
SUVH1-NlaIVR	ACTATGATTCATGAATCGGGCAA**G**GTTC	
SUVHsmaI	TCCCCCGGGACTGCTCCAAGATTCACG	*SUVH1* genomic fragment amplification
SUVHclaI	CCATCGATTCCAAATGAGCCACGGCAATAC	
SUVH1-RTF	CAAGTGGAACTACGAACCTG	qRT-PCR for *SUVH1*
SUVH1-RTR	ATGTGAGAAATGGCAAAGAA	
S1F	GGGAAAAGAGAAACAAGAGACC	qRT-PCR, ChIP-PCR for Locus 1
S1R	GAACACAAGAGCAGTGACGA	
S2F	AGGTATGGCCTGATCTCAAT	qRT-PCR, ChIP-PCR for Locus 2
S2R	GACAGTGGCAGCAGTATAGG	
S3F	CTTACCGATCGTGAGACAAG	qRT-PCR, ChIP-PCR for Locus 3
S3R	ACGGTGAACTGAAAACCATA	
S4F	CTTCGTCCAATTGTTGGTAA	qRT-PCR, ChIP-PCR for Locus 4
S4R	TCGAAGCAGTCTTCAGAGAA	
S1PF	TTGAGTTACAGTATCTTGTCGGAAAC	ChIP-PCR for the promoter of Locus 1
S1PR	AAAAGAGGATATTATGTTATCGCATGT	
S2PF	GTACACCGCGGAGACAATTC	ChIP-PCR for the promoter of Locus 2
S2PR	CAGGACGGGTTTGACAGA	
S3P1F	GGTTGTGGTCGCTAGCAAAT	ChIP-PCR for the promoter of Locus 3
S3P1R	CATGGTTAAAAATGACAAAATTGA	
S4PF	TCGTCCGACGTATTGCATAG	ChIP-PCR for the promoter of Locus 4
S4PR	AAGGAGACATTTTGGAGCAA	

References

Agius F, Kapoor A, Zhu JK. 2006. Role of the Arabidopsis DNA glycosylase/lyase ROS1 in active DNA demethylation. Proceedings of the National Academy of Sciences of the United States of America 103(31): 11796-11801.

Alvarez-Venegas R, Avramova Z. 2005. Methylation patterns of histone H3 Lys 4, Lys 9 and Lys 27 in transcriptionally active and inactive Arabidopsis genes and in atx1 mutants. Nucleic acids research 33(16): 5199-5207.

Berr A, McCallum EJ, Menard R, Meyer D, Fuchs J, Dong A, Shen WH. 2010. Arabidopsis SET DOMAIN GROUP2 is required for H3K4 trimethylation and is crucial for both sporophyte and gametophyte development. The Plant cell 22(10): 3232-3248.

Chan SW, Henderson IR, Jacobsen SE. 2005. Gardening the genome: DNA methylation in Arabidopsis thaliana. Nature reviews Genetics 6(5): 351-360.

Chen PY, Cokus SJ, Pellegrini M. 2010. BS Seeker: precise mapping for bisulfite sequencing. BMC bioinformatics 11: 203.

Choi YH, Gehring M, Johnson L, Hannon M, Harada JJ, Goldberg RB, Jacobsen SE, Fischer RL. 2002. DEMETER, a DNA glycosylase domain protein, is required for endosperm gene imprinting and seed viability in Arabidopsis. Cell 110(1): 33-42.

Du J, Zhong X, Bernatavichute YV, Stroud H, Feng S, Caro E, Vashisht AA, Terragni J, Chin HG, Tu A et al. 2012. Dual binding of chromomethylase domains to H3K9me2-containing nucleosomes directs DNA methylation in plants. Cell 151(1): 167-180.

Ebbs ML, Bender J. 2006. Locus-specific control of DNA methylation by the Arabidopsis SUVH5 histone methyltransferase. The Plant cell 18(5): 1166-1176.

Fournier A, Sasai N, Nakao M, Defossez PA. 2012. The role of methyl-binding proteins in chromatin organization and epigenome maintenance. Briefings in functional genomics 11(3): 251-264.

Fuks F, Hurd PJ, Wolf D, Nan X, Bird AP, Kouzarides T. 2003. The Methyl-CpG-binding Protein MeCP2 Links DNA Methylation to Histone Methylation. Journal of Biological Chemistry 278(6): 4035-4040.

Gendrel AV, Lippman Z, Martienssen R, Colot V. 2005. Profiling histone modification patterns in plants using genomic tiling microarrays. Nat Methods 2(3): 213-218.

Gong Z, Morales-Ruiz T, Ariza RR, Roldan-Arjona T, David L, Zhu JK. 2002. ROS1, a repressor of transcriptional gene silencing in Arabidopsis, encodes a DNA glycosylase/lyase. Cell 111(6): 803-814.

Guo L, Yu YC, Law JA, Zhang XY. 2010. SET DOMAIN GROUP2 is the major histone H3 lysie 4 trimethyltransferase in Arabidopsis. Proceedings of the National Academy of Sciences of the United States of America 107(43): 18557-18562.

Henderson IR, Jacobsen SE. 2008. Tandem repeats upstream of the Arabidopsis endogene SDC recruit non-CG DNA methylation and initiate siRNA spreading. *Genes & development* **22**(12): 1597-1606.

Huettel B, Kanno T, Daxinger L, Aufsatz W, Matzke AJ, Matzke M. 2006. Endogenous targets of RNA-directed DNA methylation and Pol IV in Arabidopsis. The EMBO journal 25(12): 2828-2836.

Johnson LM, Du J, Hale CJ, Bischof S, Feng S, Chodavarapu RK, Zhong X, Marson G, Pellegrini M, Segal DJ et al. 2014. SRA- and SET-domain-containing proteins link RNA polymerase V occupancy to DNA methylation. Nature 507(7490): 124-128.

Jones PL, Jan Veenstra GC, Wade PA, Vermaak D, Kass SU, Landsberger N, Strouboulis J, Wolffe AP. 1998. Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. Nat Genet 19(2): 187-191.

Kanno T, Huettel B, Mette MF, Aufsatz W, Jaligot E, Daxinger L, Kreil DP, Matzke M, Matzke AJ. 2005. Atypical RNA polymerase subunits required for RNA-directed DNA methylation. Nature genetics 37(7): 761-765.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome biology 14(4): R36.

Law JA, Jacobsen SE. 2010. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nature reviews Genetics 11(3): 204-220.

Li X, Qian W, Zhao Y, Wang C, Shen J, Zhu JK, Gong Z. 2012. Antisilencing role of the RNA-directed DNA methylation pathway and a histone acetyltransferase in Arabidopsis. Proceedings of the National Academy of Sciences of the United States of America 109(28): 11425-11430.

Liu J, He Y, Amasino R, Chen X. 2004. siRNAs targeting an intronic transposon in the regulation of natural flowering behavior in Arabidopsis. *Genes & development* **18**(23): 2873-2878.

Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR. 2008. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell 133(3): 523-536.

Liu ZW, Shao CR, Zhang CJ, Zhou JX, Zhang SW, Li L, Chen S, Huang HW, Cai T, He XJ. 2014. The SET domain proteins SUVH2 and SUVH9 are required for Pol V occupancy at RNA-directed DNA methylation loci. PLoS genetics 10(1): e1003948.

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome research 18(9): 1509-1517.

Matzke MA, Mosher RA. 2014. RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. Nature reviews Genetics 15(6): 394-408.

Morales-Ruiz T, Ortega-Galisteo AP, Ponferrada-Marin MI, Martinez-Macias MI, Ariza RR, Roldan-Arjona T. 2006. DEMETER and REPRESSOR OF SILENCING 1 encode 5-methylcytosine DNA glycosylases. Proceedings of the National Academy of Sciences of the United States of America 103(18): 6853-6858.

Nan X, Ng H-H, Johnson CA, Laherty CD, Turner BM, Eisenman RN, Bird A. 1998. Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. Nature 393(6683): 386-389.

Naumann K, Fischer A, Hofmann I, Krauss V, Phalke S, Irmler K, Hause G, Aurich AC, Dorn R, Jenuwein T et al. 2005. Pivotal role of AtSUVH2 in heterochromatic histone methylation and gene silencing in Arabidopsis. The EMBO journal 24(7): 1418-1429.

Ortega-Galisteo AP, Morales-Ruiz T, Ariza RR, Roldan-Arjona T. 2008. Arabidopsis DEMETER-LIKE proteins DML2 and DML3 are required for appropriate distribution of DNA methylation marks. Plant molecular biology 67(6): 671-681.

Penterman J, Zilberman D, Huh JH, Ballinger T, Henikoff S, Fischer RL. 2007. DNA demethylation in the Arabidopsis genome. Proceedings of the National Academy of Sciences of the United States of America 104(16): 6752-6757.

Peragine A, Yoshikawa M, Wu G, Albrecht HL, Poethig RS. 2004. SGS3 and SGS2/SDE1/RDR6 are required for juvenile development and the production of trans-acting siRNAs in Arabidopsis. Genes & Development 18(19): 2368-2379.

Qian W, Miki D, Lei M, Zhu X, Zhang H, Liu Y, Li Y, Lang Z, Wang J, Tang K et al. 2014. Regulation of active DNA demethylation by an alpha-crystallin domain protein in Arabidopsis. Molecular cell 55(3): 361-371.

Qian W, Miki D, Zhang H, Liu Y, Zhang X, Tang K, Kan Y, La H, Li X, Li S et al. 2012. A histone acetyltransferase regulates active DNA demethylation in Arabidopsis. Science 336(6087): 1445-1448.

Rajakumara E, Law JA, Simanshu DK, Voigt P, Johnson LM, Reinberg D, Patel DJ, Jacobsen SE. 2011. A dual flip-out mechanism for 5mC recognition by the Arabidopsis SUVH5 SRA domain and its impact on DNA methylation and H3K9 dimethylation in vivo. Genes & development 25(2): 137-152.

Rea S, Eisenhaber F, O'Carroll D, Strahl BD, Sun ZW, Schmid M, Opravil S, Mechtler K, Ponting CP, Allis CD et al. 2000. Regulation of chromatin structure by site-specific histone H3 methyltransferases. Nature 406(6796): 593-599.

Rogers SO, Bendich AJ. 1985. Extraction of DNA from milligram amounts of fresh, herbarium and mummified plant tissues. Plant molecular biology 5(2): 69-76.

Soppe WJJ, Jacobsen SE, Alonso-Blanco C, Jackson JP, Kakutani T, Koornneef M, Peeters AJM. 2000. The late flowering phenotype of fwa mutants is caused by gain-of-function epigenetic alleles of a homeodomain gene. *Molecular cell* **6**(4): 791-802.

Stroud H, Do T, Du J, Zhong X, Feng S, Johnson L, Patel DJ, Jacobsen SE. 2014. Non-CG methylation patterns shape the epigenetic landscape in Arabidopsis. Nature structural & molecular biology 21(1): 64-72.

Stroud H, Greenberg MV, Feng S, Bernatavichute YV, Jacobsen SE. 2013. Comprehensive analysis of silencing mutants reveals complex regulation of the Arabidopsis methylome. Cell 152(1-2): 352-364.

Wierzbicki AT, Haag JR, Pikaard CS. 2008. Noncoding transcription by RNA polymerase Pol IVb/Pol V mediates transcriptional silencing of overlapping and adjacent genes. Cell 135(4): 635-648.

Wierzbicki AT, Ream TS, Haag JR, Pikaard CS. 2009. RNA polymerase V transcription guides ARGONAUTE4 to chromatin. Nature genetics 41(5): 630-634.

Won SY, Li S, Zheng B, Zhao Y, Li D, Zhao X, Yi H, Gao L, Dinh TT, Chen X. 2012. Development of a luciferase-based reporter of transcriptional gene silencing that enables bidirectional mutant screening in Arabidopsis thaliana. Silence 3(1): 6.

Zhang Y, Ng H-H, Erdjument-Bromage H, Tempst P, Bird A, Reinberg D. 1999. Analysis of the NuRD subunits reveals a histone deacetylase core complex and a connection with DNA methylation. Genes & Development 13(15): 1924-1935.

Zhao Y, Xie S, Li X, Wang C, Chen Z, Lai J, Gong Z. 2014. REPRESSOR OF SILENCING5 Encodes a Member of the Small Heat Shock Protein Family and Is Required for DNA Demethylation in Arabidopsis. The Plant cell 26(6): 2660-2675.

Zheng X, Pontes O, Zhu J, Miki D, Zhang F, Li WX, Iida K, Kapoor A, Pikaard CS, Zhu JK. 2008. ROS3 is an RNA-binding protein required for DNA demethylation in Arabidopsis. Nature 455(7217): 1259-1262.

Conclusions

Over the course of evolution, the genomes of higher eukaryotes increased in size, containing a greater fraction of transposons compared to lower organisms. Transposons are a double-edged sword: while they may increase species diversity, they also have detrimental effects. DNA methylation has been found to have an indispensable role in controlling transposon expression. How DNA methylation is established and maintained is particularly well studied in *Arabidopsis*. However, two important questions remain to be addressed. The first question concerns the biogenesis of siRNAs as the guidance signal of *de novo* methylation; the lack of knowledge about the primary transcripts that function as siRNA precursors has hampered our understanding of the very first step in the establishment of DNA methylation. The second question addresses how DNA methylation, as a silencing mark, is prevented from stochastically silencing genes. My Ph.D. projects were aimed at resolving these two important problems of DNA methylation.

Project 1. Detection of Pol IV/RDR2-dependent transcripts at the genomic scale in *Arabidopsis* reveals features and regulation of siRNA biogenesis

Although RNA polymerase Pol IV has been proposed to generate siRNA precursor transcripts, Pol IV-dependent transcripts have never previously been reported due to two situations. The first one is that Pol IV-dependent transcripts are short-lived and quickly cleaved by DCL proteins once converted into dsRNAs by RDR2. The second one is that siRNA-generating loci are silenced in wild type and de-repressed in Pol IV

mutants due to decreased DNA methylation. This makes it impossible to detect Pol IV-dependent transcripts by comparing the transcriptomes of wild type and *nrpd1* (*nrpd1* is a Pol IV mutant). We therefore attempted the detection of Pol IV-dependent transcripts in a mutant with greatly compromised DCL function. Pol IV-dependent transcripts were successfully detected when the transcripts from *dcl234* (the *dcl2 dcl3 dcl4* mutant) were compared with those from *dcl234 nrpd1*. Genome-wide detection of Pol IV-dependent transcripts was also achieved after the enrichment of Pol IV-dependent transcripts through elimination of single-stranded RNAs. After assembly and analysis of the Pol IV-dependent transcripts, the Pol IV-transcribed regions, like Pol II-transcribed regions, were found to be flanked by A/T-rich sequences depleted in nucleosomes. However, Pol IV-dependent transcripts were found to differ from Pol II-dependent transcripts in terms of RNA structure, with the former having a 5' monophosphate, lacking introns, lacking a polyA tail and corresponding to both strands. In contrast, Pol II-dependent transcripts have a 5' CAP, introns and a polyA tail and derive from only one strand. The common genomic features (i.e., regions flanked by A/T-rich sequences) raised the possibility that Pol IV transcription initiation may require Pol II, while the contrasting features of the RNAs generated by Pol IV and Pol II may reflect the different functions of these transcripts. Utilizing the available genome-wide DNA methylation and small RNA datasets, the regulation of siRNAs by CHH DNA methylation was also discussed.

Project 2. SUVH1, a histone methyltransferase, is required for the expression of genes targeted by DNA methylation

Compared to the well-studied DNA methylation deposition process, much less is known about how DNA methylation is under control to prevent the stochastic silencing of genes. To identify new factors with negative roles in gene silencing, a forward genetic screen was carried out using a reporter line with a *LUC* gene driven by a double *35S* promoter. One mutant with decreased *LUC* expression was found to disrupt *SUVH1*, which encodes a SET domain protein. Treatment with a DNA inhibitor abolished the decreased *LUC* expression phenotype of the *suvh1* mutant, indicating that SUVH1 function requires DNA methylation. The unaltered DNA methylation at the *LUC* locus and throughout the genome suggested that SUVH1 functions downstream of DNA methylation. Although SUVH1 is a homolog of the H3K9me2 methyltransferase SUVH4, H3K9me2 levels were not affected in *suvh1*. However, a decrease in H3K4me3 was observed in *suvh1*, which was consistent with the observed decrease in *LUC* expression; the finding also raises the possibility that SUVH1 functions as an H3K4me3 methyltransferase. The presence of transposons and DNA methylation in the promoter regions of SUVH1-targeted loci indicates that SUVH1-targeted loci are subject to silencing by DNA methylation. Possibly to ensure gene expression while maintaining transposon silencing, SUVH1 may bind methylated promoters and methylate genic H3K4 without altering the silencing marks (DNA methylation and H3K9me2) at the promoter. This functional analysis of SUVH1 reveals one possible mechanism by which the silencing effects of DNA methylation are circumvented, which may be necessary

throughout the course of evolution to counteract the detrimental and complex effects of random transposon movement.

Appendix A. Construction of mRNA-seq libraries

Advancements in next-generation sequencing technology have allowed genome-wide sequencing to be more widely applied in numerous areas of biological research. Transcriptome assembly and the identification of differentially expressed genes are two common applications of next-generation sequencing and involve the profiling of the transcriptome a genomic scale. For these analyses, mRNA-seq libraries containing the expressed RNA information need to be constructed. During my Ph.D. studies, in addition to the libraries I constructed for my own projects, I generated 78 mRNA-seq libraries for our collaborators. These libraries are summarized in Table A.1, and the sequencing data can be downloaded from http://illumina.ucr.edu/ht.

Table A. 1 List of mRNA-seq libraries I constructed.

Species	Ecotype	Genotype	Project ID	Flow -cell	Lane	Index sequence	Comments
Tomato	Green	AJ1	518	173	6	ATCACG	
Tomato	Green	AJ2	518	173	6	TTAGGC	
Tomato	Green	C1	518	173	6	ACTTGA	
Tomato	Green	C2	518	173	6	GATCAG	
Tomato	Green	MS1	518	173	6	TAGCTT	
Tomato	Green	MS2	518	173	6	GGCTAC	
Tomato	Green	TV1	518	173	6	GTGGCC	
Tomato	Green	TV2	518	173	6	GTTTCG	
Tomato	Green	G1	518	173	7	ATCACG	
Tomato	Green	G2	518	173	7	TTAGGC	
Tomato	Green	G3	518	173	7	ACTTGA	
Tomato	Green	G4	518	173	7	GATCAG	
Tomato	Green	G5	518	173	7	TAGCTT	
Tomato	Green	G6	518	173	7	GGCTAC	
Tomato	Green	G7	518	173	7	GTGGCC	
Tomato	Green	G8	518	173	7	GTTTCG	
Zingiberales	Canna	Sample1	513	173	5	ATCACG	
Zingiberales	Canna	Sample1	513	173	5	TTAGGC	
Zingiberales	Canna	Sample2	513	173	5	ACTTGA	
Zingiberales	Canna	Sample2	513	173	5	GATCAG	
Zingiberales	Canna	Sample3	513	173	5	TAGCTT	
Zingiberales	Canna	Sample3	513	173	5	GGCTAC	
Zingiberales	Canna	Sample4	513	173	5	GTGGCC	
Zingiberales	Canna	Sample4	513	173	5	GTTTCG	
Zingiberales	Musa	Sample1	431	155	2	ATCACG	

Zingiberales	Musa	Sample1	431	155	2	TTAGGC	
Zingiberales	Musa	Sample2	431	155	2	ACTTGA	
Zingiberales	Musa	Sample2	431	155	2	GATCAG	
Zingiberales	Musa	Sample3	431	155	2	TAGCTT	
Zingiberales	Musa	Sample3	431	155	2	GGCTAC	
Zingiberales	Costus	Sample1	431	155	2	CGATGT	
Zingiberales	Costus	Sample1	431	155	2	TGACCA	
Zingiberales	Costus	Sample2	431	155	2	ACAGTG	
Zingiberales	Costus	Sample2	431	155	2	GCCAAT	
Zingiberales	Costus	Sample3	431	155	2	CAGATC	
Zingiberales	Costus	Sample3	431	155	2	CTTGTA	
Zingiberales	Costus	Sample4	431	155	2	AGTCAA	
Zingiberales	Costus	Sample4	431	155	2	AGTTCC	
Arabidopsis	L*er*	WT (wild-type)	445	148	7	TGACCA	
Arabidopsis	L*er*	WT	445	148	7	ACAGTG	
Arabidopsis	L*er*	*top1a*	445	148	7	GCCAAT	
Arabidopsis	L*er*	*top1a*	445	148	7	CTTGTA	
Arabidopsis	Col	WT	445	148	3	CGATGT	
Arabidopsis	Col	*nrpd1-3*	445	148	3	TGACCA	
Arabidopsis	Col	*nrpe1-1*	445	148	3	ACAGTG	
Arabidopsis	Col	*tho5*	445	148	3	GCCAAT	
Arabidopsis	Col	*top1a*	445	148	3	CAGATC	
Arabidopsis	Col	*nua-3*	445	148	3	CTTGTA	
Arabidopsis	L*er*	WT	445	148	5	CGATGT	
Arabidopsis	L*er*	*top1a*	445	148	5	TGACCA	
Arabidopsis	Col	963DMSO	445	148	5	ACAGTG	Chemical treatment

Arabidopsis	Col	963CPT	445	148	5	GCCAAT	Chemical treatment
Arabidopsis	Col	963KU	445	148	5	CAGATC	Chemical treatment
Arabidopsis	Col	963 CPTKU	445	148	5	CTTGTA	Chemical treatment
Arabidopsis	Col	*YJ* LIN	445	148	6	CGATGT	
Arabidopsis	Col	RH1 *YJ*	445	148	6	TGACCA	Over-expression
Arabidopsis	Col	RH2 *YJ*	445	148	6	ACAGTG	Over-expression
Arabidopsis	Col	SB2 *YJ*	445	148	6	GCCAAT	Over-expression
Arabidopsis	Col	SB3 *YJ*	445	148	6	CAGATC	Over-expression
Arabidopsis	Col	*pwr-2*	445	148	6	CTTGTA	Over-expression
Arabidopsis	Col	*YJ*	434	145	7	CGATGT	
Arabidopsis	Col	*hpr1 YJ*	434	145	7	TGACCA	
Arabidopsis	Col	*suvh1 YJ*	434	145	7	ACAGTG	
Arabidopsis	Col	*nua YJ*	434	145	7	GCCAAT	
Arabidopsis	Col	*hsp20 YJ*	434	145	7	CAGATC	
Arabidopsis	Col	*pwr-1*	434	145	7	CTTGTA	
Arabidopsis	Col	972	434	145	8	CTTGTA	
Arabidopsis	Col	*tex1* 972	434	145	8	ACAGTG	
Arabidopsis	Col	10-34L	434	145	8	GCCAAT	AT3G04490
Arabidopsis	Col	*taf6* 972	434	145	8	CAGATC	
Arabidopsis	Col	*mom1* 972	434	145	8	AGTCAA	
Arabidopsis	Col	*hsp20* 972	434	145	8	GTCCGC	

Arabidopsis	Col	972	434	145	1	CTTGTA	
Arabidopsis	Col	*ago4* 972	434	147	1	CGATGT	
Arabidopsis	Col	*drd1* 972	434	147	1	TGACCA	
Arabidopsis	Col	*hrp1 ago4* 972	434	147	1	CAGATC	
Arabidopsis	Col	*hpr1 drd1* 972	434	147	1	AGTCAA	
Arabidopsis	Col	*top1a ago4* 972	434	147	1	GTCCGC	

Appendix B. Gene identification through map-based cloning

As introduced in Chapter 3, a forward genetic screen was performed using ethyl methanesulfonate (EMS) on *YJ* to identify factors participating in anti-silencing processes. For the *YJ* line, Dr. Yun Ju Kim introduced a *LUCIFERASE* (*LUC*) reporter gene driven by a double *35S* promoter into the *rdr6-11* background. In the *rdr6-11* background, sense transgene-induced post-transcriptional gene silencing (S-PTGS) of transgenes is suppressed (Peragine et al. 2004a). The existence of DNA methylation and the presence of siRNAs mapping to the double *35S* promoter region indicated that the transgene in *YJ* is regulated by RdDM. Dr. Kim obtained several mutants exhibiting reduced *LUC* luminescence, and for some of these mutants, I used map-based cloning techniques to identify the genes harboring the phenotype-inducing mutations. Specifically, I identified four genes, including *SUVH1*, from five mutants with decreased *LUC* activity.

Two mutants with low *LUC* activity were found to be in the same complementation group, and through map-based cloning, *AT3G14980* (*IDM1*) was identified as the affected gene responsible for the observed phenotype (Figures B.1, B.2). *idm1-1* and *idm1-2* were found to harbor G-to-A mutations leading to stop codons in the 7th and 2nd exons, respectively. The studies of a collaborator characterized the histone acetyltransferase activity of IDM1 at loci lacking H3K4me2/H3K4me3 to promote ROS1 function (Li et al. 2012; Qian et al. 2012).

HPR1 (*AT5G09860*), which encodes a core component of the THO complex, was isolated from another mutant with decreased *LUC* activity (Figures B.1, B.2). In *hpr1*, a G-to-A

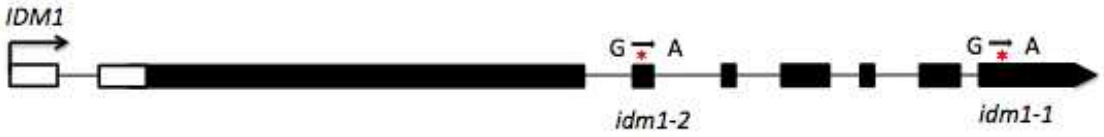
nonsense mutation in the 17th exon results in a truncated protein. The THO complex, a conserved nuclear protein complex, affects the biogenesis of mRNP and is recruited to chromatin to function at the interface between transcription and nuclear mRNA export (Rondon et al. 2010). Previous studies have shown that HPR1 participates in the biogenesis of endogenous and exogenous siRNA (Jauvion et al. 2010). The phenotype of the mutant indicated that the THO complex may function as a negative factor of RdDM. The function of HPR1 in the DNA methylation pathway was investigated by Dr. Yuanyuan Zhao.

AT1G79280 (*NUA*) was identified from another mutant with low *LUC* activity (Figure B.2). In this mutant, a G-to-A mutation at the splice junction between the 2nd and 3rd exons (Figure B.1) yields an altered transcript and protein. *NUA* encodes a nuclear pore anchor protein localized to the inner surface of the nuclear envelope and is a component in mRNA nuclear export in plants (Xu et al. 2007; Tamura et al. 2010). The identification of this gene and the THO complex as potential negative factors of RdDM indicates that nuclear RNA metabolism and RNA export help maintain RdDM homeostasis. Studies analyzing the role of *NUA* in DNA methylation-mediated transcriptional gene silencing were performed by Dr. So Youn Won.

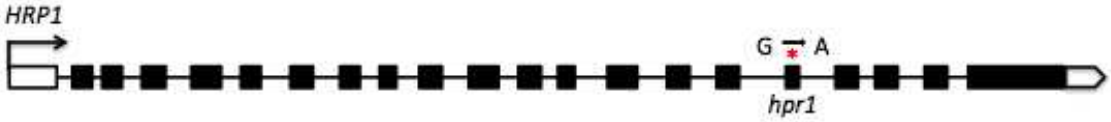
Figure B. 1 Diagrams illustrating the mutations in the isolated genes.

Exons and introns are represented by rectangles and lines, respectively. UTRs and 3' end regions are designated with unfilled rectangles and triangles, respectively, and the asterisks denote the positions of the mutations. A, Schematic diagram of the *IDM1* gene showing the G-to-A mutations in *idm1-1* (7th exon) and *idm1-2* (2nd exon). Both mutations lead to a stop codon in the respective exons. B, Schematic diagram of *HPR1* showing the G-to-A mutation in the 17th exon, which leads to a stop codon. C, Schematic diagram of the *NUA* gene showing the G-to-A mutation at the splice junction between the 2nd and 3rd exons.

A



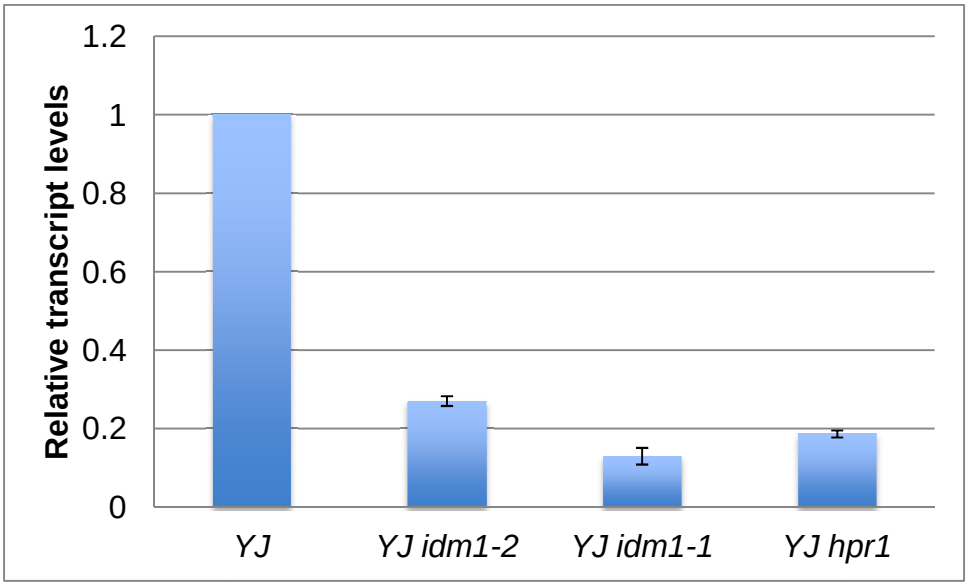
B



C

Figure B. 2 Reduced *LUC* expression in mutant determined by real-time PCR.

Similar results were obtained for three biological replicates.



Appendix C. Gene identification based on whole-genome sequencing

Forward genetics mutagenesis screens are widely used in the field of molecular biology and involve the identification of mutants exhibiting a phenotype of interest and the subsequent identification of the genes harboring the relevant mutations. Ultimately, this method of gene identification is aimed at improving the understanding of the mechanism underlying the biological process of interest. Prior to the development of next-generation sequencing, the identification of phenotype-inducing mutations was largely accomplished through map-based cloning (Jander et al. 2002). However, the decreasing cost of deep-sequencing technology has permitted sequencing-based gene identification, as reported in a number of studies (Zuryn et al. 2010; Schneeberger and Weigel 2011; Zhu et al. 2012). In addition to significantly reducing the labor cost associated with traditional map-based cloning, next-generation sequencing may also be applied in circumstances where the traditional method does not work. Challenges associated with map-based cloning include the production of the mapping population, which requires a second ecotype that does not compromise the phenotype of the isolated mutant. In some cases, these conditions cannot be met. A second challenge arises when the distance between a parental insertion and the mutation is too small to be resolvable by map-based cloning. In my Ph.D. studies, I employed sequencing-based gene identification on several occasions and helped streamline the methodology for our lab and collaborators.

Identification of phenotype-inducing mutations in a genetic screen for DNA methylation factors

In our lab, both EMS and T-DNA mutagenesis screens using the reporter lines *YJ* and *LUCH* (Won et al. 2012) were initiated with the goal of identifying novel factors involved in DNA methylation. Many mutants exhibiting either high or low *LUC* luminescence were subsequently isolated (as introduced in Chapter 3 and Appendix B). In several cases, the affected genes proved difficult to identify by map-based cloning despite intensive labor input, and genome sequencing and profiling were subsequently attempted.

The mutation in one of the mutants with low *LUC* activity (10-34L) isolated from the EMS-treated *LUCH* screen had already been mapped to a region between 1M and 2M on Chromosome 3. A DNA-seq library was constructed using the DNA from a single mutant individual and submitted for sequencing on an Illumina HiSeq 2000. After analyzing the SNPs in the aforementioned region, a C-to-T nonsense mutation was found in the 29th exon of *AT3G04490* (Figure C.1A). Downstream genotyping was performed to confirm that the mutation was responsible for the low *LUC* expression phenotype. Little is known about AT3G04490 function, but it is homologous to XPO4 in higher eukaryotes, which functions as a mediator of a novel nuclear export protein (Lipowsky et al. 2000; Bollman et al. 2003). The phenotype of the mutant suggests that AT3G04490 may function as a negative factor of RdDM.

Another mutant, named 9-60H, was isolated from the T-DNA mutagenesis of *LUCH* and exhibited increased *LUC* expression. Linkage analysis showed that the increased *LUC* expression phenotype was not associated with the T-DNA insertion; thus, the phenotype was attributable to a mutation accompanied by the T-DNA insertion. The result of map-based cloning identified a linked position on Chromosome 3, but the insertion of the *LUC* transgene in the same area made it impossible to further narrow the region. To isolate the gene, DNA was extracted from the F2 mapping population and used to construct the DNA-seq library. After SNP analysis, the mutation was linked to a region centered at 8M on Chromosome 3. After checking all mutation types (insertions, deletions and point mutations), a C deletion was detected in the 4th exon of *AT3G27380* (*NRPD2*), which encodes the second largest subunit of Pol IV/Pol V (Onodera et al. 2005). Pol IV and Pol V are key factors in the RdDM pathway: Pol IV is responsible for the biogenesis of siRNA (Zhang et al. 2007; Mosher et al. 2008), while Pol V is responsible for generating scaffold transcripts that recruit the AGO4-siRNA complex (Wierzbicki et al. 2008; Wierzbicki et al. 2009). The isolation of a Pol IV/Pol V mutant further confirmed that *LUCH* is under RdDM regulation.

The YY1170 mutant isolated from the T-DNA mutagenesis of *YJ* exhibited low *LUC* expression. Linkage analysis revealed that the decreased *LUC* expression phenotype was associated with, and thus caused by, the T-DNA insertion. A DNA-seq library was constructed, and mosaic reads containing partial T-DNA sequences and partial sequences of *AT3G06290* from the *Arabidopsis* genome were found, indicating that disrupted

AT3G06290 function caused the decreased *LUC* expression. *AT3G06290* encodes a homolog of SAC3, a core component of the conserved TREX-2 complex (Tamura et al. 2010). The functional studies of *AtSAC3* were performed by Dr. Yuanyuan Zhao.

Identification of phenotype-inducing mutations from a genetic screen for factors involved in flower development

AGAMOUS (AG) is an important transcription factor controlling the termination of the floral stem cells. In *ag* null mutants, both floral stem cell termination and floral organ identity specification are disrupted (Bowman et al. 1989). In contrast, a weak allele known as *ag-10* exhibits normal floral organ identity specification and only mild defects in floral stem cell termination (Ji et al. 2011; Liu et al. 2011b). Taking advantage of the weak *ag-10* phenotype for the identification of genes potentially involved in the temporal regulation of floral stem cells, a forward genetics EMS mutagenesis screen was performed in *ag-10.* Four mutants with enhanced *ag-10* phenotypes are discussed below. In a previous round of screening, another *ag* allele, which contained a second SNP in the *AG* coding region, was identified and named *ag-11*. EMS mutagenesis was also carried out on *ag-11*, and a mutant with a suppressed *ag-11* phenotype was isolated. To identify the genes affected in this mutant and the four aforementioned mutants from the *ag-10* screen, whole-genome sequencing was performed using the F2 populations of backcrossed plants and the *ag-10* and *ag-11* parental lines. None of the five mutants were found to be in the same complementation group, indicating that the relevant SNPs occur in distinct genes. Following SNP identification for each library, the candidate SNPs were

further characterized by mutation type and the resulting amino acid changes. Several candidate genes were identified for each of the mutants (summarized in Tables C.1-C.5), but further experiments are required to confirm which genes are responsible for the enhanced or suppressed phenotypes. These analyses will include complementation testing, phenotypic assessments of different alleles and genotyping of the segregating populations.

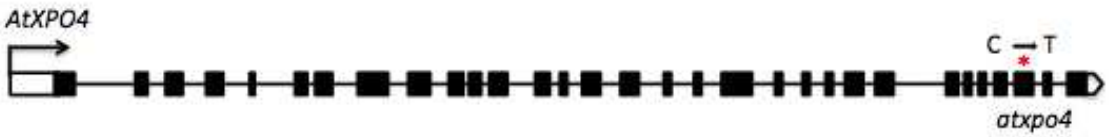
Identification of factors participating in the antiviral defense pathway

To identify novel factors involved in antiviral defense, Dr. Shou-wei Ding's group screened a population of homozygous T-DNA insertion lines of *Arabidopsis* for mutants exhibiting altered resistance to viral infection. Among the lines identified from the screen, some had phenotypes linked to the T-DNA insertion, while others did not. Lines 049 and 149 were found to be in the same complementation group, and the phenotype of interest was not linked to the T-DNA insertion. To identify the phenotype-inducing mutations, whole-genome DNA sequencing libraries were constructed using two groups of F2 backcrossed populations for both lines, and the group with altered resistance (mutant) was compared to the corresponding control group with no change in resistance. Bioinformatics analysis revealed a large chromosome deletion in both line 049 and line 149 but not in the control (Figure C.2). Because these large chromosome deletions resulted in the absence of many genes, further experiments are required to identify the genes of interest.

Figure C. 1 Diagrams showing the mutations in the genes isolated from the *YJ* and *LUCH* screens.

Exons and introns are represented by rectangles and lines, respectively. UTRs and 3' end regions are designated by unfilled rectangles and triangles, respectively, and asterisks denote the positions of the mutations. A, Schematic diagram of *AtXPO4* (*AT3G04490*) showing the nonsense G-to-A mutation in the 29th exon in 10-34L. B, Schematic diagram of *NRPD2* showing the *nrpd2* mutation. The C deletion in the 4th exon is a frameshift mutation. C, Schematic diagram of *AtSAC3B* (*AT3G06290*) showing the T-DNA insertion in the 17th exon in the *atsac3b* mutant.

A


AtXPO4
C → T
atxpo4

B


NRPD2
C deletion
nrpd2

C

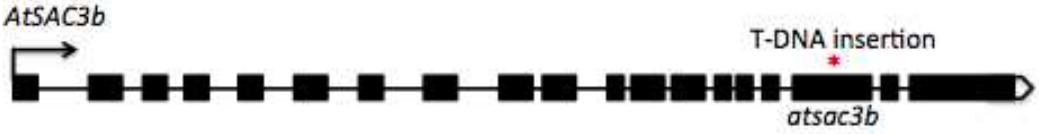

AtSAC3b
T-DNA insertion
atsac3b

Figure C. 2 A genome browser view of the aligned reads showing the big deletion in the two mutants.

The aligned reads from the two mutant libraries (lines 049 and 149) are absent at the shown region, while aligned reads are present in the respective wild-type control libraries.

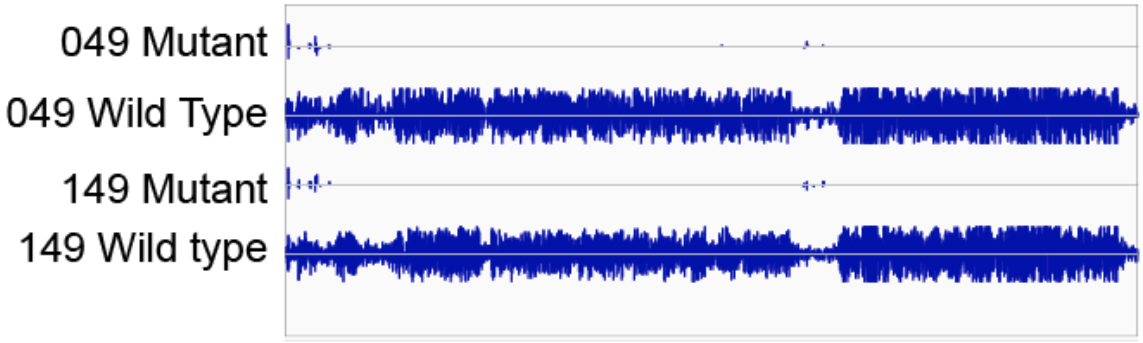


Table C. 1 Candidate genes for *m7* (an *ag-10* enhancer).

Chromo-some	Position	SNPs	Amino acid change	Gene ID	Gene name and annotation
4	2489095	C-T	A-V	AT4G04970	GSL1
4	4996162	G-A	H-N	AT4G08180	ORP1C
4	8904603	G-A	G-R	AT4G16280	FCA
4	10846397	G-A	R-K	AT4G20910	HEN1
4	13950547	G-A	G-S	AT4G20010	Plastid transcriptionally active 9

Table C. 2 Candidate genes for *m317* (an *ag-10* enhancer).

Asterisks denote a stop codon.

Chromo-some	Position	SNPs	Amino acid change	Gene ID	Gene name and annotation
2	5790732	G-A	G-R	AT2G14120	A dynamin related protein
2	7070719	C-T	P-S	AT2G16850	Plasma membrane intrinsic protein 2;8
2	8733631	C-T	R-W	A T2G20950	Arabidopsis phospholipase-like protein family
2	8792768	C-T	P-S	AT2G21150	XAP5 family protein
2	9667606	C-T	P-L	AT2G23380	CLF
2	9886374	C-T	A-V	AT2G23900	Pectin lyase-like superfamily protein
2	10446194	G-A	V-I	AT2G25220	Protein kinase superfamily protein
2	12102532	C-T	S-F	AT2G28890	A protein phosphatase 2C like gene
2	14901073	G-A	G-E	AT2G36350	Protein kinase superfamily protein
2	14982039	G-A	G-E	AT2G36500	CBS
2	16731708	C-T	P-L	AT2G40930	Ubiquitin-specific protease
2	17011418	G-A	P-L	AT2G41630	TFIIB1
2	18182421	G-A	D-N	AT2G44930	Plant protein of unknown function (DUF247)
3	19018713	G-A	W-*	AT3G52250	POWERDRESS
3	19476395	C-T	L-F	AT3G53510	ABC-2 type transporter family protein
5	2517384	G-A	M-I	AT5G07950	Unknown protein
5	6285761	G-A	E-K	AT5G18980	ARM repeat superfamily protein

Table C. 3 Candidate genes for *m140* (an *ag-10* enhancer).

Chromo -some	Position	SNPs	Amino acid change	Gene ID	Gene name and annotation
4	17219396	G-A	E-K	AT4G37370	Member of CYP81D

Table C. 4 Candidate genes for *m446* (an *ag-10* enhancer).

Asterisks denote a stop codon.

Chromo-some	Position	SNPs	Amino acid change	Gene ID	Gene name and annotation
1	27253628	C-T	L-F	AT1G73600	S-adenosyl-L-methionine-dependent methyltransferases superfamily protein
3	7216197	C-T	H-Y	AT3G20720	Unknown protein
3	7350140	C-T	P-L	AT3G21060	A structural core component of a COMPASS-like H3K4 histone methylation complex
3	7793253	C-T	W-*	AT3G2170	Monomeric G protein
3	8385051	G-A	W-*	AT3G23510	Cyclopropane-fatty-acyl-phospholipid synthase
3	10391234	G-A	P-S	AT3G28070	Nodulin MtN21-like transporter family protein
3	10699414	G-A	H-Y	AT3G28715	TPase, V0/A0 complex, subunit C/D
5	3655256	G-A	R-Q	AT5G11510	MYB3R-4
5	5358501	G-A	D-N	AT5G16510	RGP5
5	6482613	G-A	E-K	AT5G19390	A Rho GTPase activating protein
5	645202	G-A	M-I	AT5G02870	Ribosomal protein L4/L1 family
5	1315632	G-A	D-N	AT5G04640	AGAMOUS-like 99
5	1814236	G-A	E-K	AT5G06090	Putative sn-glycerol-3-phosphate 2-O-acyltransferase

Table C. 5 Candidate genes for *m40* (an *ag-11* suppressor).

Chromo-some	Position	SNPs	Amino acid change	Gene ID	Gene name and annotation
1	18440823	G-A	G-R	AT1G50575	Putative lysine decarboxylase family protein
1	22255520	C-T	R-C	AT1G61310	LRR and NB-ARC domains-containing disease resistance protein
1	22279929	G-A	E-K	AT1G61360	S-locus lectin protein kinase family protein
1	23730678	C-T	T-I	AT1G64960	ARM repeat superfamily protein
1	24612808	G-A	R-K	AT1G67040	TON1 RECRUITING MOTIF 22
1	25787610	G-A	D-N	AT1G69670	ATCUL3B
1	26097726	G-A	V-M	AT1G70370	PG2
1	26533215	G-A	G-D	AT1G71691	GDSL-like Lipase/Acylhydrolase superfamily protein
1	27157127	G-A	A-T	AT1G73350	Unknown
1	28625268	G-A	R-K	AT1G77300	ASHH2

Appendix D. Genome-wide profiling of nuclear transcripts dependent on Pol II and Pol V

In recent years, widespread intergenic and antisense transcripts have been identified in fungi, animals and plants using tiling arrays and high-throughput sequencing methods. While these studies have drastically altered our view of the transcriptional landscape of the genome, the functions of these non-coding RNAs (ncRNAs) are only beginning to be uncovered. The findings for several ncRNAs suggest that one of the prominent roles of long ncRNAs is to act in the nucleus to recruit chromatin-modifying factors. For example, Xist, HOTAIR and COLDAIR have been reported to recruit the PRC2 complex to alter the chromatin status of the corresponding locus (Plath et al. 2003; Rinn et al. 2007; Heo and Sung 2011). In addition to the function of ncRNAs, their biogenesis and metabolism also remain unclear. In *Arabidopsis*, both Pol II and Pol V have been shown to generate non-coding transcripts from a few siRNA-generating loci (Wierzbicki et al. 2008; Zheng et al. 2009). However, the genomic scale of Pol II-dependent and Pol V-dependent non-coding transcripts has not yet been studied.

Because long ncRNAs appear to act predominantly in the nucleus, we extracted nuclear RNA to enrich functional long ncRNAs. We subsequently employed high-throughput sequencing techniques to identify transcripts with polyA tails dependent on Pol II and Pol V by profiling and comparing the sequences obtained from wild-type, *nrpb2-3* (a Pol II mutant) and *nrpe1-1* (a Pol V mutant) nuclear RNA. After analyzing the mRNA-seq

libraries (Table D. 1), the differentially expressed transcripts were obtained using DESeq (Table D.2) (Anders and Huber 2010).

Table D. 1 The nuclear mRNA-seq library information of Pol II and Pol V.

Genotype	Project ID	Index	Flowcell	Lane
WT	254	ATCACGA	192	2
WT	254	GATCAGA	192	2
nrpb2-3	254	TTAGGCA	192	2
nrpb2-3	254	TAGCTTA	192	2
nrpe1-1	254	ACTTGAA	192	2
nrpe1-1	254	GGCTACA	192	2

Table D. 2 The number of differentially expressed nuclear transcripts in *nrpb2-3* and *nrpe1-1* compared to WT.

The cut off used here are p-value < 0.05 and fold change > 2.

Changes in *nrpb2-3* compared to WT	Transcripts located at genic regions	Increased	229
		Decreased	384
	Transcripts located at intergenic regions	Increased	24
		Decreased	17
	Transcripts located at intronic regions	Increased	24
		Decreased	5
Changes in *nrpe1-1* compared to WT	Transcripts located at genic regions	Increased	28
		Decreased	4
	Transcripts located at intergenic regions	Increased	54
		Decreased	1
	Transcripts located at intronic regions	Increased	74
		Decreased	3

Appendix E. Genome-wide profiling of transcripts with different 5’ end structures

RNAs transcribed by different polymerases possess different structures. The transcripts of Pol II and Pol V have a 7-methylguanosine cap at the 5’ initiating nucleotide, those of Pol I and Pol III have a triphosphate group and the transcripts of Pol IV have a monophosphate group. To profile RNAs with different structures at the genomic scale, RNAs with a caps or triphosphate group were obtained and used to construct libraries.

The RNAs with a 5’ cap (5’ CAP) were obtained through the following procedure. 30 ug DNA-free RNAs extracted with TRIzol were fragmented using Fragmentation Reagents (Ambion, AM8740). The 160-300 nt RNAs were purified from a denaturing polyacrylamide gel and treated with Terminator Exonuclease (Epicenter, TER51020) at 30°C for an hour to get rid of RNAs with a 5’ monophosphate. After phenol-chloroform purification and ethanol precipitation, the RNAs were treated with CIP (NEB, M0290S) to get rid of RNAs with a 5’ phosphate group. After another round of phenol-chloroform purification and ethanol precipitation, the RNAs were treated with Tobacco Acid Pyrophosphatase (TAP, Epicenter T19250) to hydrolyze the 7-methylguanosine cap to a monophosphate group. The RNAs purified with phenol-chloroform were used to build RNA-seq libraries using the True-seq small RNA preparation kit (Illumina, RS-200-0012).

The RNAs with a 5’ triphosphate (5’ PPP) were obtained following a similar procedure. 30 ug DNA-free RNAs extracted with TRIzol were fragmented using Fragmentation Reagents (Ambion, AM8740). The 160-300 nt RNAs were purified from a denaturing polyacrylamide gel and treated with Terminator Exonuclease (Epicenter,

TER51020) at 30°C for an hour to get rid of RNAs with a 5’ monophosphate. After phenol-chloroform purification and ethanol precipitation, the RNAs were treated with RNA 5’ Polyphosphatase (Epicenter, RP8092H) to convert the 5´-triphosphate to 5´-monophosphate. The RNAs purified with phenol-chloroform were used to build RNA-seq libraries using the True-seq small RNA preparation kit (Illumina, RS-200-0012).

Table E. 1 Information of 5’ end RNA-seq libraries.

Genotype	Structure	Project ID	Index	Flowcell	Lane
WT	5’ CAP	653	GGCTAC	208	5
dcl234	5’ CAP	213	CGATGT	213	2
dcl234 nrpd1	5’ CAP	213	TGACCA	213	2
dcl234 rdr2	5’ CAP	213	ACAGTG	213	2
dcl234	5’ PPP	213	GCCAAT	213	2
dcl234 nrpd1	5’ PPP	213	CAGATC	213	2
dcl234 rdr2	5’ PPP	213	CTTGTA	213	2

Appendix F. A discussion about DNA methylation independent of RDR2

Two companion papers published in Molecular Cell in October 2012 described a new DNA methylation pathway dependent on SDE3 or NERD and involving PTGS components (Garcia et al. 2012; Pontier et al. 2012). The studies have two major conclusions. One is that SDE3 and NERD regulate DNA methylation through AGO2 and RDR1/6, but not thruogh AGO4 and RDR2. The other is that NERD-dependent DNA methylation is dependent on 21 nt siRNAs but not 24 nt siRNAs.

For the first conclusion, I agree with the authors on the fact that the increased expression at psORF is accompanied by decreased DNA methylation in *nerd, sde3, ago2* and *rdr6*. However, I have some questions about the role of NERD or SDE3 on DNA methylation. First, only two NERD-dependent loci (psORF and AT1E93275) and one SDE3-dependent locus (psORF) shown in the paper exhibit the correlation between increased expression and decreased DNA methylation. Second, the numbers of NERD-dependent hypomethylated or hypermethylated loci are very small (Table F. 1) when comparing to the tens of thousands of loci dependent on other DNA methylation pathway factors, for example MET1, CMT3, DRM2, CMT2 etc, (data not shown). Third, for the DNA methylation loci validated in the papers, most have significant CHH methylation. Considering the fact that the CHH methylation loci only represent a small portion of the NERD-dependent hyper/hypo methylated loci (Table F.1), we conclude that the validated loci in the papers are not representative of NERD-regulated DNA methylation loci. In summary, the number of NERD-regulated DNA methylation loci is very small, which

makes us wonder whether NERD regulates DNA methylation or the Differentially Methylated Regions (DMRs) between wild type and *nerd* are just a random occurrence.

For the second conclusion that NERD regulates DNA methylation through 21 nt siRNAs but on 24 nt siRNAs, I agree with the author that this is a good explanation for the fact that increased expression and decreased DNA methylation are observed in *rdr6* but on in *rdr2*. However, more experimental data are needed to support this point. First, the 21 nt siRNAs need to be shown to decrease in *rdr6* but not in *rdr2*. However, decreased levels of 21 nt siRNAs were only detected in the reporter line *SucSUL* in *nerd*, where the DNA methylation status was not provided. Second, there should be data showing that 21 nt siRNAs are produced at NERD-regulated DNA methylation loci at the genomic scale, which is absent in the papers.

In summary, the studies show that, at psORF, the loss of DNA methylation leads to increased DNA methylation with many PTGS proteins mutation, however, more loci need to be provided to support the general roles of PTGS proteins in TGS. In addition, more experiment data are needed to provide direct evidence for the role of 21 nt siRNAs in the NERD/SDE3-regulated DNA methylation pathway.

Table F. 1 Overlap between Hyper/Hypo methylated loci in nerd with total DNA methylation loci.

	Hypomethylated in *nerd*	Hypermethylated in *nerd*
Total	651	146
CG methylated loci[1]	340	114
CHG methylated loci[2]	48	27
CHH methylated loci[3]	51	26

[1,2,3] The loci were obtained through previous methylome data (Stroud et al. 2013).

[1] CG methylated loci are 100-bp windows with methylation levels higher than 0.4 in wild-type plants.

[2] CHG methylated loci are 100-bp windows with methylation levels higher than 0.2 in wild-type plants.

[3] CHH methylated loci are 100-bp windows with methylation levels higher than 0.1 in wild-type plants.

References

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome biology* **11**(10): R106.

Bollman KM, Aukerman MJ, Park MY, Hunter C, Berardini TZ, Poethig RS. 2003. HASTY, the Arabidopsis ortholog of exportin 5/MSN5, regulates phase change and morphogenesis. *Development* **130**(8): 1493-1504.

Bowman JL, Smyth DR, Meyerowitz EM. 1989. Genes directing flower development in Arabidopsis. *The Plant cell* **1**(1): 37-52.

Garcia D, Garcia S, Pontier D, Marchais A, Renou JP, Lagrange T, Voinnet O. 2012. Ago hook and RNA helicase motifs underpin dual roles for SDE3 in antiviral defense and silencing of nonconserved intergenic regions. *Molecular cell* **48**(1): 109-120.

Heo JB, Sung S. 2011. Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA. *Science* **331**(6013): 76-79.

Jander G, Norris SR, Rounsley SD, Bush DF, Levin IM, Last RL. 2002. Arabidopsis map-based cloning in the post-genome era. *Plant physiology* **129**(2): 440-450.

Jauvion V, Elmayan T, Vaucheret H. 2010. The conserved RNA trafficking proteins HPR1 and TEX1 are involved in the production of endogenous and exogenous small interfering RNA in Arabidopsis. *The Plant cell* **22**(8): 2697-2709.

Ji L, Liu X, Yan J, Wang W, Yumul RE, Kim YJ, Dinh TT, 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 genetics* **7**(3): e1001358.

Li X, Qian W, Zhao Y, Wang C, Shen J, Zhu JK, Gong Z. 2012. Antisilencing role of the RNA-directed DNA methylation pathway and a histone acetyltransferase in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America* **109**(28): 11425-11430.

Lipowsky G, Bischoff FR, Schwarzmaier P, Kraft R, Kostka S, Hartmann E, Kutay U, Gorlich D. 2000. Exportin 4: a mediator of a novel nuclear export pathway in higher eukaryotes. *The EMBO journal* **19**(16): 4362-4371.

Liu X, Kim YJ, Muller R, Yumul RE, Liu C, Pan Y, Cao X, Goodrich J, Chen X. 2011. AGAMOUS terminates floral stem cell maintenance in Arabidopsis by directly

repressing WUSCHEL through recruitment of Polycomb Group proteins. *The Plant cell* **23**(10): 3654-3670.

Mosher RA, Schwach F, Studholme D, Baulcombe DC. 2008. PolIVb influences RNA-directed DNA methylation independently of its role in siRNA biogenesis. *Proceedings of the National Academy of Sciences of the United States of America* **105**(8): 3145-3150.

Nuthikattu S, McCue AD, Panda K, Fultz D, DeFraia C, Thomas EN, Slotkin RK. 2013. The initiation of epigenetic silencing of active transposable elements is triggered by RDR6 and 21-22 nucleotide small interfering RNAs. *Plant physiology* **162**(1): 116-131.

Onodera Y, Haag JR, Ream T, Costa Nunes P, Pontes O, Pikaard CS. 2005. Plant nuclear RNA polymerase IV mediates siRNA and DNA methylation-dependent heterochromatin formation. *Cell* **120**(5): 613-622.

Peragine A, Yoshikawa M, Wu G, Albrecht HL, Poethig RS. 2004. SGS3 and SGS2/SDE1/RDR6 are required for juvenile development and the production of trans-acting siRNAs in Arabidopsis. *Genes & development* **18**(19): 2368-2379.

Plath K, Fang J, Mlynarczyk-Evans SK, Cao R, Worringer KA, Wang H, de la Cruz CC, Otte AP, Panning B, Zhang Y. 2003. Role of histone H3 lysine 27 methylation in X inactivation. *Science* **300**(5616): 131-135.

Pontier D, Picart C, Roudier F, Garcia D, Lahmy S, Azevedo J, Alart E, Laudie M, Karlowski WM, Cooke R et al. 2012. NERD, a plant-specific GW protein, defines an additional RNAi-dependent chromatin-based pathway in Arabidopsis. *Molecular cell* **48**(1): 121-132.

Qian W, Miki D, Zhang H, Liu Y, Zhang X, Tang K, Kan Y, La H, Li X, Li S et al. 2012. A histone acetyltransferase regulates active DNA demethylation in Arabidopsis. *Science* **336**(6087): 1445-1448.

Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, Farnham PJ, Segal E et al. 2007. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129**(7): 1311-1323.

Rondon AG, Jimeno S, Aguilera A. 2010. The interface between transcription and mRNP export: from THO to THSC/TREX-2. *Biochimica et biophysica acta* **1799**(8): 533-538.

Schneeberger K, Weigel D. 2011. Fast-forward genetics enabled by new sequencing technologies. *Trends in plant science* **16**(5): 282-288.

Tamura K, Fukao Y, Iwamoto M, Haraguchi T, Hara-Nishimura I. 2010. Identification and characterization of nuclear pore complex components in Arabidopsis thaliana. *The Plant cell* **22**(12): 4084-4097.

Wierzbicki AT, Haag JR, Pikaard CS. 2008. Noncoding transcription by RNA polymerase Pol IVb/Pol V mediates transcriptional silencing of overlapping and adjacent genes. *Cell* **135**(4): 635-648.

Wierzbicki AT, Ream TS, Haag JR, Pikaard CS. 2009. RNA polymerase V transcription guides ARGONAUTE4 to chromatin. *Nature genetics* **41**(5): 630-634.

Won SY, Li S, Zheng B, Zhao Y, Li D, Zhao X, Yi H, Gao L, Dinh TT, Chen X. 2012. Development of a luciferase-based reporter of transcriptional gene silencing that enables bidirectional mutant screening in Arabidopsis thaliana. *Silence* **3**(1): 6.

Xu XM, Rose A, Muthuswamy S, Jeong SY, Venkatakrishnan S, Zhao Q, Meier I. 2007. NUCLEAR PORE ANCHOR, the Arabidopsis homolog of Tpr/Mlp1/Mlp2/megator, is involved in mRNA export and SUMO homeostasis and affects diverse aspects of plant development. *The Plant cell* **19**(5): 1537-1548.

Zhang X, Henderson IR, Lu C, Green PJ, Jacobsen SE. 2007. Role of RNA polymerase IV in plant small RNA metabolism. *Proceedings of the National Academy of Sciences of the United States of America* **104**(11): 4536-4541.

Zheng B, Wang Z, Li S, Yu B, Liu JY, 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**(24): 2850-2860.

Zhu Y, Mang HG, Sun Q, Qian J, Hipps A, Hua J. 2012. Gene discovery using mutagen-induced polymorphisms and deep sequencing: application to plant disease resistance. *Genetics* **192**(1): 139-146.

Zuryn S, Le Gras S, Jamet K, Jarriault S. 2010. A strategy for direct mapping and identification of mutations by whole-genome sequencing. *Genetics* **186**(1): 427-430.