Genome-wide Identification of Non-coding Transcription by RNA Polymerase V and Its Involvement in Transcriptional Gene Silencing

by

Shriya Sethuraman

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2021

Doctoral Committee:

Professor Andrzej T. Wierzbicki, Co-Chair
Associate Professor Alan P. Boyle, Co-Chair
Professor Györgyi Csankovszki
Assistant Professor Peter L. Freddolino
Professor Patricia J. Wittkopp

Shriya Sethuraman

shriyas@umich.edu

ORCID iD: 0000-0001-5033-1105

To my parents, S. Sethuraman and Sudha Sethuraman

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor, Dr. Andrzej Wierzbicki, for his constant support throughout my PhD. The reason I joined his lab was the incredible ease with which I was able to converse with him. His open-door policy, allowing me to walk into his room whenever I wanted to discuss my ideas and plans with him, definitely made life much easier throughout the 5 years in the lab and for that, I am eternally grateful. His knowledge and proficiency in molecular biology has proven to be something to aspire to and his ability to find a workable solution to most problems has been a great motivation for me. I believe that I have become a much better researcher over time and most of that credit goes to Andrzej.

I would also like to thank all the members of the Wierzbicki lab: Hafiz M Rothi, Jan Kuciński, Masayuki Tsuzuki, Gudrun Böhmdorfer, Miguel Palomar, Sho Fujii, Adriana Coke and Lilia Bouzit. From the various game and lunch parties in the lab to the many many more short talks that turned into 2-3 hour long discussions, the moments spent in the lab with these people will be something I would cherish for life. Even though we considered ourselves boring and different from other "outgoing" labs, I believe we actually had our own crazy moments, ranging from the hour long games of "Pandemic" to the debatably wonderful "apple picking" gatherings we have had. Quite frankly, I believe each of those moments were wonderful and I will definitely miss every one of you. I would specifically like to thank Hafiz, Janek and Masayuki for being there for almost the entire duration of my PhD, enlivening the lab with jokes, important long work-related discussions and planning for every project. It would have been really difficult without each of you in the lab, so thank you for everything.

I would like to also thank my co-advisor, Dr. Alan Boyle, for his constant support throughout my PhD. The monthly meetings with him helped me think critically and solve various technical problems. I have learnt many new bioinformatics tools and techniques through these meetings that have definitely helped me make great improvements in my research, and thereby, made me a better bioinformatician. Thank you for the much needed support and push.

My thesis committee members, Dr. Peter Freddolino, Dr. Györgyi Csankovszki and Dr. Patricia Wittkopp, have contributed greatly to my PhD research. The interesting discussions that took place in my committee meetings definitely gave me new directions to work on. The wide-ranging expertise of my committee; from Peter's suggestions of improved statistical tests, to Györgyi's biological insights and Trisha's bioinformatics suggestions, they have all led to my PhD projects jumping leaps and bounds ahead of what it would have been had it not been for their feedback. So, I would like to thank each one of them for their support and guidance throughout my PhD.

I would also like to thank Dr. Sundeep Kalantry and his entire lab, for giving me the opportunity to work with them. Dr. Kalantry has been very helpful and kind, and has supported me with a lot of good advise throughout my PhD life. I would also specially like to thank Dr. Emily Maclary, my graduate mentor in the Kalantry lab, for introducing me to bioinformatics and sequencing data analysis. Emily patiently taught me the basics of data analysis, which became the foundation of my informatics knowledge that I was able to build on over the many years of my PhD.

Next, I would like to thank my family starting with my parents, Sudha and Sethu, for their support through every hurdle in my life. PhD wasn't a piece of cake and I am sure I wouldn't have made it in one piece without their constant support. I would also like to thank my grandparents for their belief and trust in my abilities, that kept me motivated. I would like to specifically mention my uncle, Srinivasan Sundaresan, and aunt, Anita

Sundaresan, for providing me with a home in Michigan. I have taken a lot of advantage of them as they were just an hour away from Ann Arbor, and they have constantly indulged me and supported and helped me, for which I will remain ever grateful. I would also like to mention my cousins, Maya and Anya, who have been like my own siblings and always provided me with some fascinating stories and entertainment in the form of school soccer and swimming tournaments over the years. Thank you for everything.

I would like to thank my fiancé, Vaibhav Gogte, for being there for me and listening to my crazy rants and complaints throughout the years. It has become evident to me that when times are tense and you want to rant, you turn to those you are closest to, and Vaibhav has always stood by me, patiently listening to all my complaints. Thanks for being patient with me throughout the years and making sure I had a life outside of work and for the wonderful short trips that you would drag me to, in the midst of COVID, when I would forget to leave my home for hours or days together. Thank you for trying to find new ways to bring a smile to my face. It would have been much harder to get here, if not for you. I would also like to thank Sangita Gogta, Vinod Gogte and Aishwarya for their support and kindness during the most tedious times of my PhD.

Another important aspect to my life in Ann Arbor has been the friends I made here. Every time I needed an out from the annoying negative results and repetitive reanalysis of data, I had the support of my friends. From the Friday evening game nights to the fun trips and hikes, the "Juwaris" group has been one of the main reasons leaving Ann Arbor feels very strange. Aditi Kulkarni, Poorwa Shekhar, Sneha Joshi, Vaibhav Gogte, Kumar Aanjaneya, Bharadwaj Mantha, Niket Prakash, Kunal Garg, Harshad Dharmatti, Saurabh Mahajan, Ripudaman Singh, Manu Akula, Bedant Bhutia, Akshitha Sriraman and Keval Ramani - thank you for making Ann Arbor home, and for the fun activities and weekends we spent together with songs and games! I would also like to specially thank Aditi, Poorwa and Sneha for the fun chats and discussions we have had, and Vaibhav and Aanjaneya for welcoming me into their weekday racquetball games, teaching me the sport and challenging me everyday.

I also want to mention my various roommates, Shouchang Guo, Jyotsna Wase and Shyama Nandakumar, for being very kind and patient with me. I would specially like to thank Shouchang for listening to me patiently during the last few years of my PhD when I had the most things to complain about and Jyotsna for being there and helping me during my injuries and relocations. I would like to also thank my friends from undergrad, Amrita Meher, Kartik Palani, Naren GK, Sushrut Shitoot, Rohit Srivastava, Varsha K. Raman, Vijaya Rana and Rishi Naidu for the fun trips and meetups over the last 6 years. Special thanks to the friends I made in the department, with special mentions of Gayatri Iyer and Arushi Varshney. Some of the trips we took to Cedar Point and Frankenmuth or our thanksgiving late night shopping breaks were definitely great fun.

I would also like to thank Dr. Margit Burmeister for being a wonderful graduate advisor. She has been there for me whenever I had questions, doubts or concerns about the right next steps as a PhD student and has always given great advise keeping my best interests in mind. Last but definitely not the least, I would like to thank the Bioinformatics staff, especially Julia Eussen, for all her support and assistance throughout the years.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

RNA-mediated transcriptional gene silencing is a conserved process where non-coding RNAs target transposons and other sequences for repression by establishing repressive chromatin modifications. A central element of this process is long non-coding RNAs (lncRNAs), which in *Arabidopsis thaliana* are produced by a specialized RNA polymerase known as Pol V. These lncRNAs recruit small interfering RNAs (siRNAs) and a series of proteins that lead to the establishment of RNA-directed DNA methylation (RdDM) on transposable elements. Transposable elements extant in eukaryotic genomes pose a constant risk of disrupting the integrity of the genome via random integration events and are targeted for silencing by the RdDM machinery. The RdDM pathway results in *de novo* DNA methylation and it has been quite extensively researched, however, questions about the mechanism of recruitment of Pol V to RdDM loci and subsequent interplay between chromatin modifications and the downstream mechanism of gene silencing are still less understood.

In this work, I have utilized high-throughput molecular sequencing data to expand our understanding of the transcriptional gene silencing pathway. First, I addressed and expanded our understanding of Pol V transcription at RdDM loci. I have successfully identified and annotated Pol V transcribed RdDM loci throughout the genome. I have further shown how Pol V transcription is controlled by preexisting chromatin modifications located within the transcribed regions. I observed that Pol V transcribes into transposons in a non-strand specific manner and the DNA methylation targeted to these transposons also occur on both strands and is tightly restricted to the Pol V transcribed regions. I further show that the preferential enrichment of Pol V transcription and downstream DNA methylation

at the edges of transposons depicts a possible role of Pol V in determining heterochromatin boundaries.

Second, my research helped us better understand the mechanism of Pol V transcription. I have shown that Pol V transcription is not restricted to RdDM loci but is much more pervasive. Through my research, I show how at already established RdDM targets, Pol V and siRNA work together to maintain silencing. In contrast, some euchromatic sequences do not give rise to siRNA but are covered by low levels of Pol V transcription, which is needed to establish RdDM *de novo*, if a transposon is reactivated. Through this study, I show that Pol V surveils the genome to make it competent to silence newly activated transposons, making it essential for maintaining the integrity of the genome.

Third, I address the effect of Pol V transcription on downstream repressive chromatin modifications and gene silencing. I show that RdDM affects nucleosomes through recruitment of the SWI/SNF chromatin remodeling complex. Next, I address the relationship between the two chromatin modifications showing that despite DNA methylation being predominantly enriched at linkers, RdDM target loci show an enrichment of both nucleosomes and DNA methylation. My data further depicts that nucleosome placement by RdDM has no detectable effects on the pattern of DNA methylation. Instead, I show that DNA methylation by RdDM affects nucleosome positioning, suggesting that DNA methylation directs nucleosomes and they both coordinately bring about gene and transposon silencing at the RdDM loci.

CHAPTER I

Introduction

1.1 Gene regulation and chromatin modifications

Genome is the term used to refer to the genetic material of an organism comprised of genes and non-coding regions like transposons, pseudogenes, introns, repeat sequences and telomeres. In eukaryotes, genes are the basic physical and functional unit of heredity that code for all the proteins of the organism. They make up a very small portion of the organism's genome; approximately 1-2% in humans[1]. Non-coding regions of the genome do not code for proteins and yet make up a large part of the genome of most organisms. Despite initially being referred to as "junk DNA", it was quickly understood that even though these non-coding regions do not directly code for proteins, they play an important role in regulating gene expression[2].

Gene expression regulation includes the various cellular mechanisms and processes that control the rate and level of expression of certain genes. Regulation of gene expression can occur at many different stages of expression including: transcriptional regulation (controlling the rate of transcription of the gene into mRNAs), translational regulation (controlling the rate of translation of mRNA into proteins), and post-translational regulation (controlling the activity and stability of the proteins) [3]. The most commonly studied mechanism and most relevant to this study is the transcriptional gene regulatory mechanism. Transcriptional regulation occurs as a result of interactions between DNA and proteins that can affect the rate of transcription. The regions of the DNA that can affect transcription are called regulatory regions and they are generally present around genes and are known to interact with various proteins to control the expression of the gene[4]. Proteins that bind to these regulatory regions and assist in altering gene expression are called regulatory proteins. The concerted role of both these factors can lead to an increase or decrease in the rate of transcription of genes by altering recruitment of the DNA-dependent RNA polymerase to the transcription start site (TSS). Some examples of regulatory regions include promoters, enhancers and silencers. Promoters are regions proximal to genes and present upstream of the genes, which recruit regulatory proteins called transcription factors (TFs) which, in turn, recruit the RNA polymerase, Pol II, essential for the transcription of the gene[5, 4]. Enhancers and silencers are more distal regulatory regions of a gene that increase or decrease the binding of TFs to the promoter, thereby turning on or off the gene transcription, respectively[6, 7]. Thus, regulatory regions and proteins have the ability to directly affect the recruitment of Pol II to genes, thereby controlling gene transcription. The structural organization of the DNA is another essential factor that can have an impact on the gene expression levels.

Chromatin is a complex of DNA and proteins that is formed in eukaryotic cells and is utilized to package DNA into the nucleus of the cell. In chromatin, the DNA is highly condensed and wrapped around proteins called histones[8]. There are 5 known classes of histones, referred to as H1,H2a,H2b,H3 and H4. 2 each of histones H2a, H2b, H3 and H4 come together to form a histone octamer that makes up a nucleosome[8]. The nucleosome-free regions of the DNA are bound to histone H1 and are called linkers[9, 10]. Nucleosomes are positioned equidistantly throughout the genome and can act as repressors of gene expression by preventing the recruitment of other proteins to DNA[11, 12, 13].

Histones are characterized by the presence of N-terminal tails that can also be post-translationally modified by methylation, acetylation, phosphorylation or ubiquitinylation, which can impact the chromatin structure[14]. Histone modifiers are protein complexes that direct post-translational modifications (PTMs) to histone tails[14, 15, 16]. Most, if not all, histone PTMs are reversible. Many of these histone modifiers have been identified and their functions characterized[16]. Histone modifiers can be broadly classified into the following categories: histone acetyltransferases (HATs) and histone deacetylases (HDACs); histone methyltransferases (HMTs) and histone demethylases (HDMs). These modifiers can "write" (as in case of HATs and HMTs) and "erase" (as in case of HDACs and HDMs) modifications onto histone tails[14, 15]. The more interesting property of these modifiers is in their ability to "read" or sense the presence of specific modifications, which then directs their action at a particular locus[15, 17]. This shows that histone modifiers can not only alter the state of the chromatin by controlling the modifications at histone tails but also interact and sense the presence of other proteins and modifiers to direct modifications based on these interactions[15, 17].

PTMs of the histone tails have been shown to have a direct effect on increasing and decreasing the chromatin compaction by altering the extent of interaction with an adjacent nucleosome[14]. Another common effect of histone PTMs are in the altered recruitment of transcription factors or effector proteins that activate downstream signalling. Histone modifications can also occur downstream of TF binding, wherein a TF recruits histone modifiers to direct a specific PTM to histones. These histones have then been shown to act as a co-activator or co-repressor of transcription by altering the recruitment of the RNA polymerase to the TSS[17]. Some known histone PTMs that have been shown to positively regulate gene expression include methylation of the 4th lysine residue in histone H3 (H3K4me1,H3K4me2,H3K4me3) and modifications to the 36th lysine residue of histone H3 (H3K36me3,H3K36ac). On the other hand, methylation of H3K9 residues (H3K9me2,H3K9me3) and H3K27 residues (H3K27me2,H3K27me3) is related with transcriptional repression[18, 19]. Acetylation of histone lysines tend to activate transcription [18, 19].

Thus, histones and nucleosomes have a repressive effect on gene expression but some modifications to the histone tails of the nucleosomes could help in increasing the rate of transcription of genes, either by opening up the chromatin for access by TFs[14] or by actually attracting TFs[14] or by altering the recruitment of the RNA polymerase to the TSS along with the TFs[17]. Histone modifications are one of the well-studied chromatin modifications on a global scale that can alter gene expression.

DNA methylation is another chromatin modification that is capable of regulating gene expression[20, 21]. The DNA backbone is made up of a sequence of 4 possible nucleotides: Adenine(A), Guanine(G), Thymine(T) and Cytosine(C). DNA methylation is the addition of a methyl group (-CH_3) to the cytosine nucleotide of a DNA. DNA methylation acts as a negative regulator of gene transcription by either recruiting repressors to prevent transcription or preventing the binding of TFs necessary for transcription[22].

DNA methylation can occur in one of three possible sequence contexts, CG, CHG or CHH, where H can be A, T or C. CG and CHG contexts of methylation are called symmetric because both the complementary DNA strands, in these contexts, contain Cs which are generally methylated together[20]. Thus, once methylation occurs in the CG or CHG contexts, the epigenetic mark can be maintained through replication. Methylation in the CHH context, however, is asymmetric in nature as the cytosine in this context exists only on one strand of the DNA. At these asymmetric loci, as DNA replication occurs, only one strand maintains the methylation mark and after every replication event, the newly synthesized C needs to undergo *de novo* methylation, in order to maintain the methylation through replications[20]. In mammals, DNA methylation mainly occurs at the CG dinucleotide sequences[20].DNA methylation in other contexts of C, like CHG and CHH, have also been observed in mammalian embryos, which is believed to be a mechanism to further control transcription in actively differentiating cells[21, 23]. In plants, DNA methylation can commonly occur in all three sequencing contexts throughout their life.

DNA methylation occurs by the activity of an enzyme called DNA methyltransferase (DNMT). There exist many different DNMTs in different organisms and each of these proteins have a unique target and function. In mammals, DNMT1 is the maintenance methyltransferase which controls reestablishing methylation after replication at the CG and CHG sites[22]. The *de novo* methylation in mammals is catalyzed by DNMT3A and DNMT3B[22]. In plants, maintenance in the CG context is handled by MET1 and in the CHG context is handled by CHROMOMETHYLASE3 (CMT3)[21, 23]. The main *de novo* methyltransferases in plants is known as DOMAINS REARRANGED METHYLTRANSFERASE (DRM2), however, some locus-specific *de novo* methylation has also been seen to occur through CHROMOMETHYLASE2 (CMT2)[24].

Thus, chromatin modifications like DNA methylation, nucleosome positioning and histone modifications are all important for gene regulation. Almost all these chromatin modifications, except for certain histone PTMs, act as negative gene regulators and tend to repress the transcription of genes in their presence. This suggests that these modifications need to be both dynamic as well as immensely controlled and there needs to exist a strong interplay between these chromatin modifications. These modifications also tend to co-occur at many loci throughout the genome and they generally act together to control gene regulation. The interplay between these features has been a question for many researchers and understanding the mechanisms of interaction of these modifications has led to many interesting studies about the cross-talk between these features.

1.2 Cross-talk between chromatin modifications and remodelers

Chromatin remodeling is a dynamic process of altering the chromatin architecture making it accessible to transcriptional and translational machinery and thereby controlling gene expression[15, 16]. Chromatin remodeling can occur mainly through histone modifiers, which direct post-translational modifications to histone tails, or through ATP-dependent chromatin remodelers, which can move or restructure nucleosomes[15, 25]. DNA methyltransferases are a third class of proteins that direct DNA methylation to specific locations on the DNA, which in turn can directly affect the chromatin structure or indirectly affect the structure by recruiting remodelers to methylated loci[26]. These remodelers and modifiers act in unison, constantly remodeling the chromatin as per the needs of the cell.

DNA methylation is controlled and directed by various DNA methyltransferases (DNMTs) and histone modifications are directed by histone modifiers. Although these chromatin modifications are controlled by different enzymes, there is a constant cross-talk between the two systems that modulates the gene repression programming in every organism. There have been studies showing that histone modifications can be read and identified by DNMTs, which in turn controls the DNA methylation pattern at a locus[27, 26]. Similarly, control of histone modifications by existing DNA methylation has also been shown[27, 26].

In developing mammals, *de novo* DNA methylation has been controlled at CpG islands by preexisting histone tail modifications[28]. It has been shown that DNMT3L, a homolog of DNMT that lacks methyltransferase activity, directs other *de novo* DNMTs to CpG sites by binding to histone H3[28]. However, this modifier is unable to bind to any methylated H3K4 sites and thereby directing DNA methylation to non-PTM histones and nucleosomes. Thus, DNA methylation in young mammals is only visible at CpG islands that lack H3K4me because of the sensing ability of the DNMT3L modifier, depicting an anti-correlation between these two features[28]. Another example of a link between histone PTM and DNA methylation is at the pericentromeric satellite repeats. At the satellites, it has been shown in mammals that the SET-domain containing HMTs, SUVH39A and SUVH39B, play an important role in directing H3K9 methylation as well as recruiting essential DNA methyltransferases to target DNA methylation to the same regions leading to a self-reinforcing silencing mechanism[29]. A Dicer-mediated mechanism identifies the RNA duplexes formed at the satellite sequences, which leads to the formation of the RISC (RNA-induced silencing complex). This RISC then recruits the HMTs SUVH39A and SUVH39B leading to the establishment of H3K9 methylation at the pericentromeric regions[29]. These HMTs also play a role in recruiting DNMT3A and DNMT3B to the same loci, leading to further establishment of DNA methylation, thereby instituting a self-reinforcing repression. Studies have also shown that the DNA methylation at these centromeric regions can occur in the absence of SUVH39 HMTs as well, suggesting that despite the cross-talk between the two epigenetic features, there are mechanisms to ensure maintenance of repression at these loci[29].

This cross-talk has also been studied in other organisms. In Neurospora Crassa, a mutation in dim-5 gene has been shown to affect H3K9 methylation at the heterochromatin by disrupting a SET-domain containing protein, which in turn leads to a reduction in DNA methylation at these loci. This suggests the interdependence of DNA methylation and histone modifications furthermore suggesting a unidirectional dependence where histone PTMs direct DNA methylation[30]. Similar studies exist in Arabidopsis thaliana as well, where we see KRYPTONITE (KYP), a SET-domain containing HMT that controls H3K9 methylation, can direct DNA methylation through a plant-specific DNMT, CMT3 (chromomethylase 3)[31]. A mutation of KYP has shown to not only reduce H3K9me but also lead to a loss of CpNpG (where N is C, A, T or G) DNA methylation at the same genomic locus suggesting histone PTMs acting upstream of DNA methylation[31]. However, the inverse dependence has also been shown at many loci where the loss of CMT3 has led to the loss of DNA methylation as well as a reduction in the H3K9me levels at these regions, depicting the existence of a more complicated mechanism and crosstalk in different organisms[26].

There exist other studies depicting the control of histone modifications by DNA methylation. Loss of CpG methylation in *met1* mutants (DNMT1 mutants) in *Arabidopsis* has been liked to a gain in H3K27 tri-methylation at those loci which in turn leads to the activation of chromatin and increased access by transcriptional machinery[32]. Similarly, the presence of DNA methylation has shown to lead to the gain of H3K9 methylation which leads to silenced chromatin[26, 32]. This crosstalk between histone and DNA methylation modifiers is essential in maintaining the chromatin and epigenetic state of the genome[26, 27].

Other than histone modifiers and DNA methyltransferases, we have a set of ATP-dependent chromatin remodelers that utilize energy from ATP hydrolysis to modify, move or evict nucleosomes in the chromatin. There are 4 extensively studied groups of remodelers: SWI/SNF (SWITCH/SUCROSE NONFERMENTABLE), ISWI, CHD (CHROMODOMAIN HELICASE DNA-BINDING) and INO80 remodelers[33, 34]. Each of these remodelers are specialized to carry out one of the following functions: Nucleosome assembly and organization, chromatin access and/or nucleosome editing[33].

Nucleosome assembly and organization is the process that generally follows a replication or transcription event where nucleosomes need to be assembled to reestablish the chromatin state. This is done by bringing the histones to the DNA and leading to the formation of the nucleosome as well as controlling the optimal desired spacing between adjacent nucleosomes[33, 34]. This is generally handled by the ISWI and CHD ATP-dependent chromatin remodelers, which play a role in maintaining the chromatin state after a dynamic re-positioning of nucleosomes has occurred due to transcription or replication machinery[33, 34]. Chromatin access is altered by remodelers when certain regions of the chromatin needs to be made accessible to cellular machinery by moving, altering or evicting nucleosomes. This is primarily, but not exclusively, done by the SWI/SNF remodelling complex which generally function to make various inaccessible regions of the chromatin accessible by proteins or transcriptional machinery[33, 34]. Nucleosome editing is the process of altering or changing the composition of the nucleosome. This is generally handled by the INO80 subfamily of ATP-remodelers.

There exists extensive crosstalk between these ATP-dependent nucleosome remodelers and histone PTMs and DNA methylation. The interactions between ATP-dependent remodelers and histone modifiers have been studied in many organisms[33, 34, 35]. The yeast Spt-Ada-Gcn5-acetyltransferase (SAGA) complex interacts with the SWI/SNF complex leading to displacement of nucleosomes resulting in gene expression[36]. Another great example is the involvement of the yeast and mammalian SWI/SNF complex in the

Rb/E2F pathway, which recruits SWI/SNF, HDACs and HMTs to the E2F promoter locus to repress transcription[37].

The extensive knowledge about crosstalk between ATP-dependent remodelers and DNA methylation also exists. It has been shown that the SWI2/SNF2 family of remodelers in different organisms appear to have effects on genome-wide DNA methylation. In mice, complete loss of *LSH*, an SNF2-family remodeler has led to the loss of DNA methylation on a genome-wide scale[38]. Similarly, DDM1, another SNF2-remodeler in Arabidopsis has been shown to control the DNA methylation levels throughout the genome and has thereby been named DECREASED IN DNA METHYLATION 1[39]. Similarly, mutations in ATRX, a SWI/SNF remodeler, has led to changes in the pattern of DNA methylation on a genome-wide scale in humans[40]. It has also been shown that in humans, loss of the SWI/SNF complex leads to a direct increase in the DNA methylation levels leading to aberrant gene activation[41]. Another study in *Arabidopsis* has shown the involvement of the SWI/SNF complex in the RNA-directed DNA methylation pathway which utilizes non-coding RNA to direct *de novo* DNA methylation to transposons, suggesting the possible interaction and involvement of these remodelers in DNA methylation establishment and control of gene expression[42].

These different remodeling mechanisms are acting in unison throughout every genome, trying to establish a dynamic chromatin architecture to enable the accessibility as well as maintenance of the genomic integrity of the organism. One such question of understanding the interdependence of DNA methylation and nucleosome positioning by the SWI/SNF pathway in maintaining gene silencing is the main questions addressed in Chapter IV. This chapter tries to understand the correlation between the two types of chromatin modifications throughout the genome and investigates the mechanism of action and crosstalk between the SWI/SNF chromatin remodeler and DNA methyltransferase, DRM2, in maintaining the genome integrity by repressing transposons.

1.3 Non-coding RNA

A very small portion of the genome codes for proteins. As mentioned earlier, it has been shown that only 1-2% of the human genome actually codes for proteins [1], however, about 90% of the whole human genome is transcribed [43]. Thus, a majority of the RNA produced do not code for proteins and are referred to as the non-coding RNAs (ncRNAs). ncRNAs play an important role in regulation of gene expression and have been shown to be involved in RNA splicing, chromatin modification, transcriptional and post-transcriptional regulation [3, 44]. ncRNAs have been classified into many types, however, the most relevant to this study include small RNA (smRNA) and long non-coding RNA (lncRNA).

Small RNAs are 18-30 bp long RNAs that generally play a role in gene silencing. Some of the known smRNAs include small interfering RNAs (siRNAs), microRNAs (miRNAs), piwi-interacting RNAs (piRNAs) and small nucleolar RNAs (snoRNAs). The biogenesis of each of these smRNAs is different and they each have functions in regulating gene expression in the cell at many levels like chromatin architecture, RNA silencing, RNA editing, transcription, translation and splicing[45, 46].

miRNAs are generally 21-24 nt long RNAs, which are formed from endogenous, short, hairpin precursors targeting loci with similar but not identical sequences[47, 48]. miRNAs bind complementary mRNAs causing translational repression by either cleaving, destabilizing or reducing the efficiency of translation of the mRNA[47, 49]. snoRNA are typically 60-300 nt long RNAs that are believed to be derived from introns of transcripts that do not have a protein-coding capacity. snoRNAs are capable of site-specific modifications of nucleotides in target RNAs by recruiting specific proteins to these RNAs by forming RNA duplexes[50].

siRNAs, the most important smRNAs in this study, are 21-24 nucleotides long and are produced from longer double-stranded RNAs or long hairpins[51, 52, 53]. These are mostly exogenous in origin and target homologous sequences either at the same loci or elsewhere in the genome for silencing or destruction [48, 54]. siRNAs are known to silence genes by directing DNA methylation to specific regions by sequence complementary[55]. This process of utilizing RNA to direct methylation to silence loci is called Transcriptional Gene Silencing (TGS) or RNA-directed DNA methylation (RdDM) [56, 57, 58, 59]. siRNAs are also involved in post-transcriptional gene silencing (PTGS) mechanisms, wherein siRNA target mRNA for cleavage and destruction[60]. These siRNAs are also believed to affect the chromatin structure by altering the H3K9 methylation levels[61]. Thus, siRNAs are involved in transcriptional and post-transcriptional gene silencing pathways and are capable of controlling the gene expression levels.

The other type of ncRNA, long non-coding RNA (lncRNA), are RNAs longer than 200bps that do not code for proteins but, again, have a very important role to play in controlling gene expression, genome stability and nuclear organization [62]. lncRNA is believed to affect gene regulation at various levels of expression[63]. At the pre-transcription stage, lncRNAs have been shown to affect the chromatin architecture by modifying histones, thereby altering the accessibility of the chromatin to transcription factors. lncRNAs are also capable of acting as scaffolds or guides to recruit DNA methyltransferases or histone modifiers to direct repressive chromatin modifications to regions of the chromatin to alter gene expression[64, 65]. lncRNA can control gene expression at the transcription level by either binding with the TFs [66] or RNA Pol II [67] or preventing the binding of the RNA polymerase to the promoter [68]. lncRNAs also play a role in post-transcriptional gene regulation by affecting splicing, mRNA stability or protein stability [63]. HOTAIR (Hox transcript antisense intergenic RNA) is a great example of pre-transcriptional repression by a lncRNA of the HOXD gene locus. Knockdown of HOTAIR leads to a reduction in the recruitment of PRC2 complex to the HOXD gene locus, which reduces the H3K27 methylation levels thereby leading to the transcriptional activation of the locus[69]. Another example of transcriptional gene silencing by a lncRNA is X-chromosome inactivation (XCI). XCI is brought about in mammalian females with 2 X-chromosomes by *Xist* lncRNA, where *Xist* coats the entire X chromosome and recruits the PRC2 complex to direct repressive histone modifications to prevent Pol II from accessing the inactivated X chromosome [70, 71].

Thus, ncRNAs do not code for proteins but they have been shown to be immensely important for the functioning of the cell as they control gene expression. One of the most important role of these ncRNAs is in Transcriptional gene silencing, wherein they direct repressive chromatin modifications to specific target regions on the chromatin.

1.4 RNA-directed DNA methylation (RdDM)

Transcriptional gene silencing (TGS) is an RNA-mediated process of silencing genes by repression of transcription caused by directing DNA methylation and other chromatin modifications to genes and it's regulatory regions[26]. RNA-mediated gene silencing occurs in most eukaryotes and it can be of different types in different organisms, however, the underlying mechanism remains the same. Quelling in fungi, RNA interference (RNAi), TGS or PTGS in animals and TGS or RdDM in plants, all utilize ncRNA to bring about gene silencing[72].

Transposable elements and other repetitive sequences are the regions of DNA that need to be targeted for immediate silencing to maintain the genomic integrity. Any aberrant expression of these regions could be very unfavorable for the organism. RNA-mediated transcriptional silencing, which in plants is also known as RNA-dependent DNA methylation (RdDM) (Law and Jacobsen, 2010) targets these transposons for silencing thereby preventing the deleterious impacts of transposon expression (Girard and Hannon, 2008, Faulkner et al., 2009; Zheng et al., 2012).

ncRNAs play a central role in the RdDM pathway directing chromatin modifications like DNA methylation and histone modifications to control gene expression and silence transpsons[55, 73, 74]. These ncRNAs in plants are produced by specialized RNA polymerases that are specific to plants. Animals have three RNA polymerases that transcribe all the RNAs in the organism: Pol I transcribes most of the ribosomal RNAs (rRNAs), Pol

II transcribes mRNAs, snRNAs, snoRNAs, lncRNAs and miRNAs, and Pol III transcribes tRNAs, 5S rRNAs and other smRNAs. However, plants contain two additional specialized polymerases, Pol IV and Pol V, which recently diverged from Pol II with specific roles in RNA silencing [75, 76, 74]. Pol IV and Pol V are very similar in composition to Pol II, with only few of the subunits modified[77]. The largest subunits of Pol IV and Pol V are NRPD1 and NRPE1, respectively, which are essential for their special role in producing siRNAs and lncRNAs required for the RdDM pathway[75, 76, 78]. Knockout mutants of *NRPD1* and *NRPE1* of Pol IV and Pol V have led to the complete loss of function of the complex leading to the loss or great reduction in the production of siRNAs and lncRNAs, respectively[78, 79]. The knockout mutants, thereby, do disrupt the RdDM mechanism in the organism, however, this is still not a deleterious mutation and the plant remains viable. This property of plants makes them an excellent source to study the TGS mechanism, as the viable knockouts of *NRPD1* and *NRPE1* proteins allows for a better understanding of the role of the siRNAs and lncRNAs in targeting chromatin modifications to transposons[75, 76, 74]. In other organisms lacking specialized polymerases, the ncRNAs are generally produced by Pol II, which also transcribes mRNA making it almost difficult to eradicate only the ncRNA to study their mechanism of action and downstream targets.

The RdDM pathway in plants consists of two distinct steps: i) Pol IV dependent 24nt siRNA biogenesis; ii) Pol V mediated lncRNA biogenesis and *de novo* DNA methylation [55]. In the first step of RdDM, Pol IV is believed to be recruited to silenced target loci, which are mostly transposons and repetitive elements. It is known that this recruitment requires the SAWADEE HOMEODOMAIN HOMOLOGUE 1 (SHH1) which binds to H3K9me2 and unmethylated H3K4 and in the process brings Pol IV to its target site[80]. Pol IV transcribes to produce a single-stranded RNA (ssRNA), which is then converted into double-stranded RNA (dsRNA) by RNA-dependent RNA polymerase, RDR2[81, 82, 83]. This dsRNA is then cleaved by DICER-LIKE 3 (DCL3) protein to form 24 nt siRNAs, which are then loaded onto an ARGONAUTE protein, namely AGO4 (sometimes AGO6

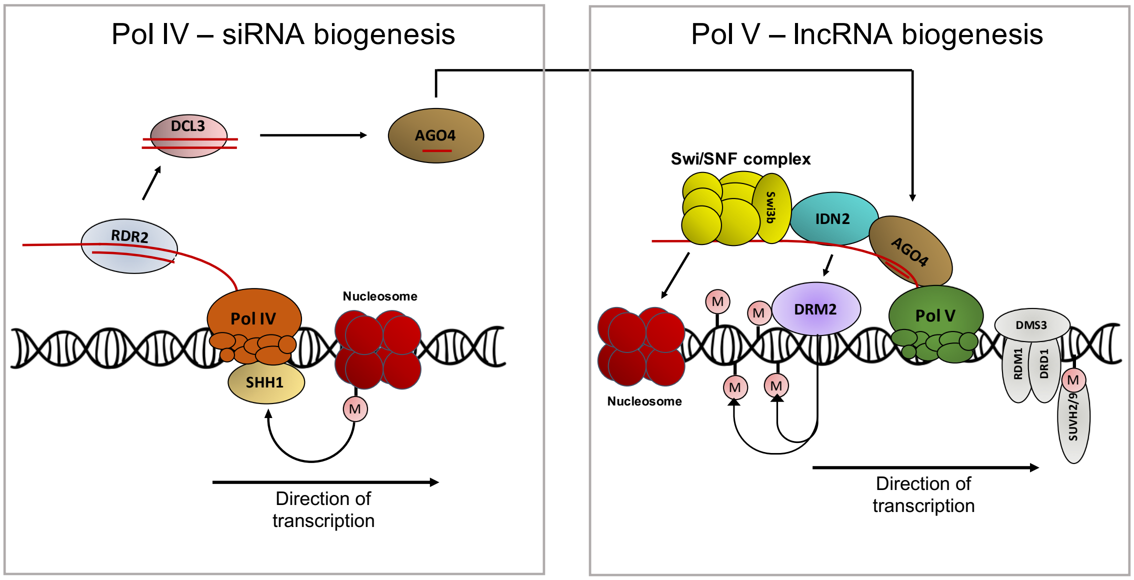

Figure 1.1: **RdDM mechanism**: In Pol IV-siRNA biogenesis (left panel), Pol IV transcribes a single-stranded RNA (ssRNA) that is copied into a double-stranded RNA (dsRNA) by RNA-DEPENDENT RNA POLYMERASE 2 (RDR2). The dsRNA is processed by DICER-LIKE 3 (DCL3) into 24-nucleotide siRNAs and incorporated into ARGONAUTE 4 (AGO4). In Pol V-lncRNA biogenesis (right panel), Pol V transcribes a scaffold lncRNA that base-pairs with AGO4-bound siRNAs. AGO4 recruitment leads to the recruitment of INVOLVED IN DE NOVO 2 (IDN2) and DOMAINS REARRANGED METHYLTRANSFERASE 2 (DRM2), which catalyses *de novo* methylation of DNA. Pol V recruitment is potentially aided by SUVH2 or SUVH9, both of which bind to methylated DNA. Nucleosome positioning is adjusted by the SWI/SNF complex, which interacts with the IDN2 (INVOLVED IN DE NOVO 2).

or AGO9, which are partially redundant) to produce the AGO4-siRNA complex[84, 55].

In the second step of RdDM, Pol V is recruited to the target loci by SU(VAR)3-9 HOMOLOGUES 2 (SUVH2) and SU(VAR)3-9 HOMOLOGUES 9 (SUVH9), which bind methylated DNA[85]. Pol V transcription is aided by the putative chromatin remodelling DDR complex, consisting of proteins DEFECTIVE IN RNA-DIRECTED DNA METHYLATION 1 (DRD1), DEFECTIVE IN MERISTEM SILENCING 3 (DMS3), and RNA-DIRECTED DNA METHYLATION 1 (RDM1)[86, 58]. Pol V transcribes lncRNA, which acts as a scaffold to recruit the AGO4-siRNA complex from the previous step through an interaction between AGO4 and NRPE1 proteins. The recruitment of the AGO4-siRNA complex is followed by recruitment of the INVOLVED IN DE NOVO 2 (IDN2) protein, which is believed to stabilize the siRNA-lncRNA pairing and leads to the recruitment of

DOMAINS REARRANGED METHYLTRANSFERASE 2 (DRM2), which is a *de novo* methyltransferase that deposits *de novo* methylation marks in all contexts of C (CG, CHG, CHH; where H is A, T or C) to specific loci [87, 88, 89, 90, 91]. This is the process by which RdDM pathway uses ncRNAs to direct DNA methylation to regions like transposons and other heterochromatin to bring about TGS.

The RdDM pathway does not only direct DNA methylation, but it has also been shown to be functionally intertwined with repositioning nucleosomes. The IDN2 protein required for recruiting the DNA methyltransferase in the second step of RdDM has also been shown to interact with SWITCH 3B (SWI3B), a subunit of the SWI/SNF chromatin remodeling complex through a protein-protein interaction suggesting the possible role of RdDM in active chromatin remodeling [42]. Subunits of the SWI/SNF complex are also known to interact with other silencing factors, including HISTONE DEACETYLASE 6 (HDA6) and MICRORCHIDIA 6 (MORC6), which are histone modifiers depicting a further possible role of RdDM in gene silencing through the SWI/SNF complex[92, 93].

Despite being a central element of the RdDM pathway, Pol V transcripts have been quite poorly understood. Unlike Pol IV transcripts, which are relatively abundant and have been characterized genome-wide[81, 94, 83], Pol V transcripts accumulate at low levels. This has made them very difficult to detect using high-through sequencing approaches like RNA-seq. Therefore, beyond the very limited loci tested in various studies, the knowledge of Pol V target loci has remained unknown. As a result of lack of our knowledge of the genome-wide Pol V target loci, identifying the different proteins that interact with these transcripts throughout the genome and how these proteins affect the transcripts is also unknown. These questions have been addressed in detail by Chapter II.

The current knowledge of the RdDM pathway also suggests the direct effect of this pathway on two of the most important chromatin modifications: DNA methylation and nucleosome remodeling. This observation has raised questions about how these two features interact with each other and how this interplay can eventually lead to gene silencing. These questions have been further investigated in Chapters IV.

1.5 Pol V regulatory elements

Pol V transcribes lncRNA and it is considered to be one of the critical factors in determining RdDM specificity. However, questions about recruitment of Pol V to its target loci is still not understood. The well-studied, canonical RNA polymerases, Pols I, II and III, are all known to recognize specific conserved promoter sequences. These promoters are recognized by transcription factors (TFs) that recruit the RNA polymerase and position it at the initiation site to begin transcription [95, 96, 97, 98]. It is believed that all polymerases should have a transcription initiation complex that directs them to the TSS. Many attempts have been made in the past to identify promoters that might be important to recruit Pol IV and Pol V [56, 86]. These promoter sequences might also serve as a good method to locate RdDM loci genome-wide. No such conserved sequences have been reported for Pol IV or Pol V at the known RdDM loci [99, 56, 86]. This could also be because of the inability to identify all the Pol V transcribed regions in the genome. Despite the existence of a knockout mutant of Pol V, a successful identification of the Pol V transcribed regions has been difficult. The RdDM loci that have been studied thus far have mostly been TEs that have been tested in a locus-specific manner. Attempts of discovery of promoters at these few loci and around known RdDM targets, including transposons, has been in vain. Other studies have suggested that Pol IV and V activity occurs at regions with pre-existing repressive chromatin modifications suggesting the involvement of repressive chromatin modifications being the "promoter" for these polymerases [85, 100, 101].

In *Arabidopsis*, Pol IV recruitment to its target loci has been shown to be affected by Pol IV accessory proteins including CLASSY SNF2-related putative chromatin remodeler family (CLSY) [102] and the SAWADEE HOMEODOMAIN HOMOLOG 1 (SHH1), which binds repressive H3K9 methylation [82, 103]. The H3K9 methylation in *Arabidopsis* requires a family of SET domain histone methyltransferases (SUVH4, SUVH5, and

SUVH6) which can also bind to methylated DNA[104]. These factors, together, generate a self-reinforcing loop of DNA and histone methylation, wherein DNA methylation leads to the recruitment of SUVH4/5/6 to the loci, leading to the deposition H3K9 methylation that is bound by SHH1. SHH1, in turn, leads to the recruitment of Pol IV which ultimately establishes DNA methylation to complete the loop[104].

On the Pol V recruitment front, the DDR protein complex and the SU(VAR)3-9 homologs (SUVH and related SUVR proteins), specifically SUVH2 and SUVH9, also known as the DNA-methylation reader proteins, have been shown to affect Pol V chromatin occupancy [58, 86, 100, 101]. The DDR complex is composed of 3 important proteins, DEFECTIVE IN RNA-DIRECTED DNA METHYLATION 1 (DRD1), a putative chromatin remodeling protein [105], DEFECTIVE IN MERISTEM SILENCING 3 (DMS3) or INVOLVED IN DE NOVO1 (IDN1) [106], and RNA-DIRECTED DNA METHYLATION 1 (RDM1) [58, 107]. This complex is believed to be essential for remodeling the chromatin to recruit Pol V. The SUVH2 and SUVH9 proteins bind methylated DNA and their absence shows a reduction in Pol V accumulation, suggesting that preexisting DNA methylation and chromatin remodeling might be important for Pol V recruitment and transcription[100, 101]. This has led to a Pol V - DNA methylation based self-reinforcing loop where DNA methylation deposited by DRM2 methyltransferase at RdDM loci is required for Pol V localization and establishment of further DNA methylation at the same loci.

The existence of this self-reinforcing mechanism of Pol IV and V transcription and chromatin modifications could explain the mechanism of maintenance of RdDM at locations where RdDM has already been established and needs to be maintained and continued through various replication events. However, Pol V has been previously shown to be required not only for RdDM maintenance, but also for *de novo* RdDM establishment at regions that do not have preexisting chromatin modifications[108, 109, 110]. The existing understanding of the field has not been able to yield a reasonable explanation as to how

Pol V is recruited to these *de novo* loci to bring about RdDM. Chapter III identifies a new mechanism of transcription of Pol V suggesting that Pol V is not recruited to specific loci but is more broadly present throughout most of the genome.

CHAPTER II

Long Non-coding RNA Produced by RNA Polymerase V Determines Boundaries of Heterochromatin

Gudrun Böhmdorfer, ***Shriya Sethuraman****, M Jordan Rowley, Michal Krzyszton, M Hafiz Rothi, Lilia Bouzit, Andrzej T Wierzbicki*

2.1 Abstract

RNA-mediated transcriptional gene silencing is a conserved process where small RNAs target transposons and other sequences for repression by establishing chromatin modifications. A central element of this process are long non-coding RNAs (lncRNA), which in Arabidopsis thaliana are produced by a specialized RNA polymerase known as Pol V. Here we show that non-coding transcription by Pol V is controlled by preexisting chromatin modifications located within the transcribed regions. Most Pol V transcripts are associated with AGO4 but are not sliced by AGO4. Pol V-dependent DNA methylation is established on both strands of DNA and is tightly restricted to Pol V-transcribed regions. This indicates that chromatin modifications are established in close proximity to Pol V. Finally, Pol V transcription is preferentially enriched on edges of silenced transposable elements, where Pol V transcribes into TEs. We propose that Pol V may play an important role in the determination of heterochromatin boundaries.

2.2 Introduction

RNA-mediated transcriptional gene silencing, in plants known as RNA-directed DNA methylation (RdDM), takes place in most eukaryotic organisms and results in heterochromatin formation by the deposition of DNA methylation and repressive histone modifications [111]. This process relies on small RNAs, which usually are generated by the activities of an RNA-dependent RNA polymerase and Dicer. Small RNAs are subsequently incorporated into Argonaute and direct repressive chromatin modifications to complementary genomic regions [111]. Recognition of target sequences by small RNAs requires ongoing non-coding transcription of the targets. This non-coding transcription gives rise to long non-coding RNA (lncRNA) which has been proposed to serve as a scaffold for Argonaute binding to chromatin, where incorporated small RNAs base pair with lncRNA [111].

In *Arabidopsis thaliana* this lncRNA is produced by a specialized DNA-dependent RNA polymerase, known as Pol V [76, 112]. Activity of Pol V is required for DNA methylation but not for the biosynthesis of the vast majority of small interfering RNAs (siRNAs) [78, 113, 79, 112]. This implicates lncRNAs produced by Pol V (referred to as Pol V transcripts) as a factor required for recognition of target loci by siRNAs. Pol V transcripts are believed to be capped or triphosphorylated on their 5' ends and not polyadenylated on their 3' ends [112]. They associate with several RNA binding proteins, including ARGONAUTE 4 (AGO4) [88]. It has been proposed that siRNAs incorporated into AGO4 base pair with Pol V transcripts and recruit AGO4 to specific loci in the genome. Binding of AGO4 is followed by the binding of INVOLVED IN DE NOVO 2 (IDN2) which interacts with a subunit of the SWI/SNF ATP-dependent chromatin remodeling complex [90, 42]. Finally, Pol V transcripts, AGO4, and/or other associated factors recruit DOMAINS REARRANGED METHYLTRANSFERASE 2 (DRM2), which is a *de novo* DNA methyltransferase [90, 107, 91]. DNA methylation is then responsible for repression of Pol II transcription on silenced loci [114]. While Pol V is involved in the late stages of the RdDM pathway, siRNA biogenesis starts with the activity of another specialized RNA polymerase, Pol IV [115, 116]. Pol IV transcripts are substrates for RNA-DEPENDENT RNA POLYMERASE 2 (RDR2) and DICER-LIKE 3 (DCL3), which produce 24nt siRNAs [117, 94]. Despite being a central element of the RdDM pathway, Pol V transcripts are poorly understood. Unlike Pol IV transcripts, which are relatively abundant and have been characterized genome-wide [81, 94, 83], Pol V transcripts accumulate at low levels. This makes them difficult to detect using high-through sequencing approaches like RNA-seq. Therefore, it remains unknown if Pol V produces any RNAs beyond the very limited number of loci tested so far. It is also unclear what defines a Pol V promoter beyond published work suggesting that both Pol IV and Pol V are recruited by preexisting repressive chromatin modifications [100, 56, 86]. It is further unknown which proteins interact with Pol V transcripts throughout the genome and how these proteins affect the transcripts. Additionally, the role of Pol V transcripts in forming the RdDM effector complex remains mysterious with several key mechanistic aspects being mostly based on speculation. These include the distance between the progressing polymerase and proteins binding to lncRNAs and the identity of nucleic acids base pairing with siRNAs [118, 119]. Finally, it is unknown if and how the specificity of Pol V recruitment to chromatin targets RdDM to individual genomic regions.

2.3 Results

2.3.1 Genome-wide identification of transcripts associated with Pol V

Current knowledge of the *in vivo* functions of Pol V and RNAs produced by this polymerase is based on a very limited number of loci [88, 57, 42, 90, 112]. To overcome this limitation, we designed an experimental approach to identify Pol V transcripts throughout the *Arabidopsis* genome. We first enriched Pol V-associated RNAs using RNA immunoprecipitation with an antibody against NRPE1, the largest subunit of Pol V [79, 88], and then subjected the samples to high-throughput sequencing (Pol V RIP-seq). We performed

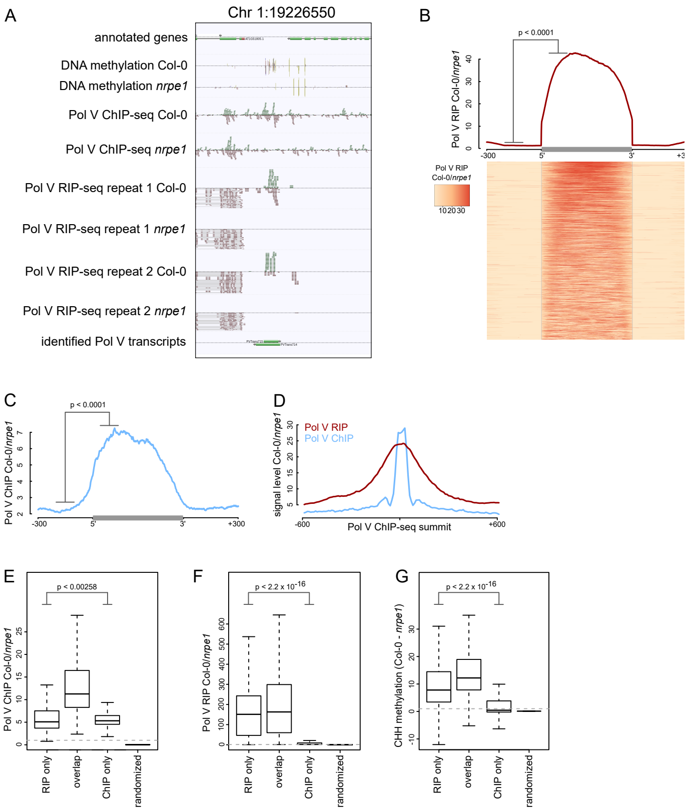

Figure 2.1: **Genome-wide identification of RNA produced by Pol V: (A)** Genomic region giving rise to Pol V transcripts. The screenshot shows sequencing reads from both repeats of Pol V RIP-seq as well as Pol V ChIP-seq [56], DNA methylation [120], and annotations of genes and Pol V transcripts. **(B)** Pol V RIP signal is largely limited to identified Pol V transcripts. All annotated Pol V transcripts were scaled to uniform lengths and average Pol V RIP signal from both biological repeats combined (Col-0/*nrpe1*, [RPM]) was plotted. The heatmap below shows Pol V RIP signal on individual transcripts sorted by length. The p value was calculated using the permutation test by comparing 100 nt long regions starting 200 nt upstream and 50 nt downstream of 5' ends of the annotated transcripts. **(C)** Pol V binding to chromatin is enriched on Pol V transcripts. Profile of average Pol V ChIP-seq signal (Col-0/*nrpe1* [RPM]) on scaled Pol V transcripts ± 300 bp. The p value was calculated using the permutation test by comparing 100 nt long regions starting 200 nt upstream and 50 nt downstream of 5' ends of the annotated transcripts. **(D)** Pol V RIP-seq signal is enriched on regions where Pol V binds chromatin. Profiles of average Pol V ChIP-seq signal (Col-0/*nrpe1*) and Pol V RIP signal (Col-0/*nrpe1*) on Pol V ChIP-seq peaks [56] aligned with their summits +/- 600 bp (10 bp resolution). **(E–G)** Loci generating Pol V transcripts are bound by Pol V and are targets of RdDM. Boxplots show regions producing Pol V transcripts but not overlapping ChIP-seq peaks and vice versa (RIP only and ChIP only, respectively) and on Pol V transcript regions overlapping ChIP peaks (overlap). Significance has been tested using the Wilcoxon test. **(E)** Pol V ChIP-seq (Col-0/*nrpe1* [RPM]), **(F)** Pol V RIP-seq (Col-0/*nrpe1* [RPM]) and **(G)** CHH DNA methylation (Col-0 - *nrpe1*).

these experiments in Col-0 wild-type and in the *nrpe1* mutant. This assay allowed the identification of genomic regions, where sequencing reads accumulated in Col-0 wild-type but not in *nrpe1* (Figure 2.1A, Figure 2.2). Hence, these reads originate from RNAs which are specifically associated with Pol V. Given that Pol V is a DNA-dependent RNA polymerase *in vitro* [117], these reads most likely stem from transcripts generated by Pol V. We also detected a considerable amount of signal over annotated genes, however, these remained unchanged in *nrpe1* (Figure 2.1A) and are therefore unlikely to be associated with Pol V. This signal was mostly present on active genes and indicates transcription by Pol I, II, III, and/or IV.

We used the RIP-seq data to annotate Pol V-associated RNAs genome-wide and identified 4502 individual high confidence Pol V-associated transcripts. Transcript calling used data from two independent biological replicates of RIP-seq. Data from both repeats were first combined to determine the ends of Pol V-associated RNAs. Then, read counts from both repeats were considered separately to filter the transcript list applying a combination of arbitrary criteria and statistical testing using the negative binomial test. The filtering criteria included a minimum of 8 reads in Col-0, a minimum four-fold enrichment between Col-0 and *nrpe1*, a p value of 0.05 and an FDR of 0.05. Details of the transcript calling strategy are described in the Materials and methods.

To visualize Pol V-associated RNAs, we plotted the average Pol V-RIP (Col-0/*nrpe1*) signal combined from both biological repeats on identified RNAs and their flanking regions (Figure 2.1B).

Individual transcripts were scaled to allow visualization over the entire lengths of the transcripts. We observed high levels of Pol V-associated transcription throughout the predicted transcripts in both replicates and only trace amounts of Pol V-associated transcription outside the annotations (Figure 2.1B). This was true for the vast majority of Pol V-associated transcripts as shown on the corresponding heatmap where every row represents an individual transcript (Figure 2.1B). This was also true when we analyzed both biologi-

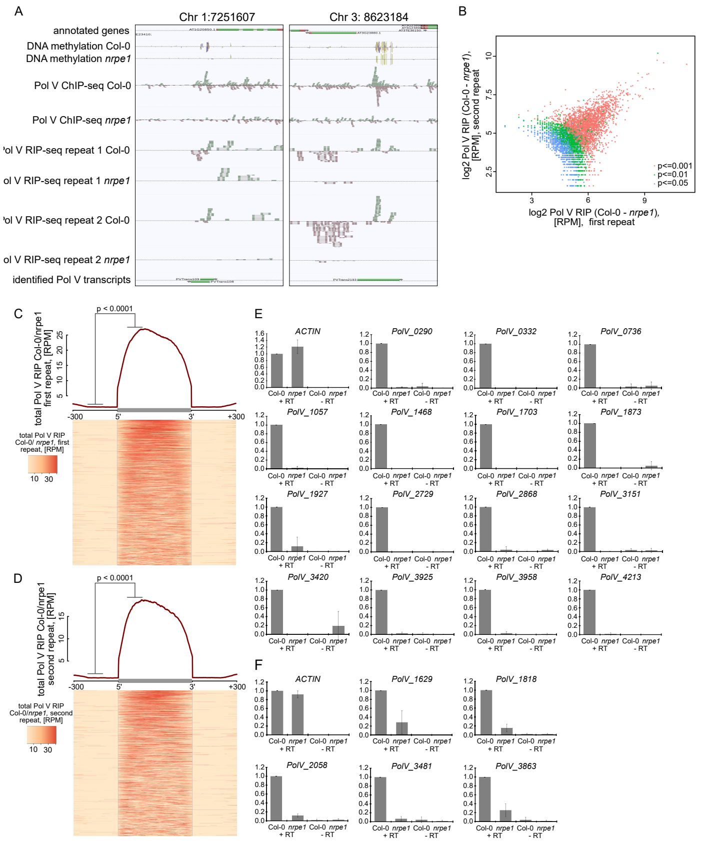

Figure 2.2: **Genome-wide identification of RNA produced by Pol V (Supplementary): (A)** Genomic regions giving rise to Pol V transcripts. The screenshots show sequencing reads from both repeats of Pol V RIP-seq as well as Pol V ChIP-seq [56], DNA methylation [120], annotations of genes (TAIR10), and annotation of Pol V transcripts obtained in this study. **(B)** Correlation between both biological repeats of RIP-seq. Scatterplot shows total Pol V RIP signal obtained from the first and the second repeat on annotated Pol V transcripts. Colors correspond to p values obtained using the negative binomial test included in the transcript calling protocol. **(C)** Pol V RIP signal is largely limited to identified Pol V transcripts – first biological repeat only. All annotated Pol V transcripts were scaled to uniform lengths and average Pol V RIP signal from the first biological repeat (Col-0/*nrpe1*, [RPM]) was plotted. The heatmap below shows Pol V RIP signal on individual transcripts sorted by length. Gray box on the x-axis indicates the position of the Pol V transcripts. In the heatmap every row represents an individual Pol V transcript sorted by size. The p value was calculated using the permutation test by comparing 100 nt long regions starting 200 nt upstream and 50 nt downstream of 5' ends of the annotated transcripts. **(D)** Pol V RIP signal is largely limited to identified Pol V transcripts – second biological repeat only. All annotated Pol V transcripts were scaled to uniform lengths and average Pol V RIP signal from the second biological repeat (Col-0/*nrpe1*, [RPM]) was plotted. The heatmap below shows Pol V RIP signal on individual transcripts sorted by length. Gray box on the x-axis indicates the position of the Pol V transcripts. In the heatmap every row represents an individual Pol V transcript sorted by size. **(E,F)** Transcripts associated with Pol V are Pol V-dependent. RT-qPCR for specific Pol V transcripts in Col-0 and *nrpe1*. Average signal levels relative to wild type and standard deviations from three biological replicates are shown.

cal repeats separately (Figure 2.2). Reproducibility between biological repeats was further tested by comparing Pol V RIP-seq signal intensities on annotated Pol V-associated RNAs in both repeats (Figure 2.2). Signal strengths measured as differences between RPM normalized read counts in Col-0 and *nrpe1* were significantly correlated (Pearson correlation r = 0.719, p<2.2*10–16), which further increases confidence in the quality of our transcript calling. Taken together, we developed a strategy which allows the sensitive and reproducible identification of Pol V-associated RNAs throughout the genome.

2.3.2 Pol V-associated RNAs are produced by Pol V

Transcripts identified using RIP-seq are bound by Pol V and are expected to be the products of Pol V based on its DNA-dependent RNA polymerase activity [117]. However, they could also be produced by another RNA polymerase and bind to Pol V posttranscriptionally. To distinguish between these possibilities, we first checked if accumulation of those transcripts required Pol V. We tested several newly identified Pol V-associated transcripts using locus-specific RT-qPCR and identified 20 loci which were suitable for primer design and had a strong reduction of RNA accumulation in *nrpe1* in RT-qPCR (Figure 2.2). Therefore, transcripts obtained by RIP-seq are not only associated with Pol V but their accumulation also depends on Pol V, which indicates that these transcripts are products of Pol V.

Next, we tested if Pol V-associated transcripts are produced from regions where Pol V is bound to DNA. We compared our RIP-seq with previously published Pol V ChIP-seq obtained using the same antibody [56] and found that Pol V binds chromatin at regions where we detected Pol V-associated transcripts (Figure 2.1C). Pol V transcription was also enriched on many genomic regions bound by Pol V [56] (Figure 2.1D). Most regions producing Pol V-associated transcripts also displayed Pol V binding to chromatin and Pol V-dependent CHH methylation [120] which is a hallmark of RdDM (Figure 2.1E–G). Regions identified only by ChIP-seq but not in RIP-seq had very low levels of Pol V-associated transcripts and low levels of Pol V-dependent CHH methylation [120] (Figure 2.1E–G). In contrast, regions giving rise to Pol V-associated transcripts, which do not overlap Pol V ChIP-seq peaks, still displayed reduction of DNA methylation in *nrpe1* (Figure 2.1G), suggesting that RIP-seq is a much more sensitive approach for detecting genomic regions transcribed by Pol V. Overall, this analysis indicates that Pol V-associated transcripts are produced from regions bound by Pol V. Together with published *in vitro* data [117], these results suggest that Pol V-associated RNAs identified using RIP-seq are produced by Pol V transcribing a genomic DNA template. Therefore, these RNAs are likely to be *bona fide* Pol V transcripts.

2.3.3 Pol V regulatory elements

RIP-seq identifies Pol V transcripts with a higher resolution than ChIP-seq and should facilitate discovery of the promoter of Pol V. RNA polymerases I, II, and III all use conserved sequence elements as their core promoters and, thus, Pol V may as well. Our attempts to identify conserved sequence motifs upstream of Pol V transcripts yielded no conclusive results. Although *de novo* discovery of promoter elements in plant genomes is not trivial [121], it is possible that Pol V may be directed to specific genomic loci by factors other than conserved sequence motifs.

To identify features that may guide Pol V, we first determined which categories of loci are transcribed by Pol V. Consistent with previously published data [57, 86], Pol V transcripts originated from pericentromeric regions and from euchromatic chromosome arms (Figure 2.3A). Pol V transcripts were preferentially produced from intergenic regions, gene promoters, and all transposon families except long terminal repeat (LTR) transposons (Figure 2.3B). This distribution is consistent with previous reports suggesting that preexisting repressive chromatin modifications are necessary for Pol V activity [85, 100, 101].

CG methylation is required for efficient Pol V binding to chromatin [100]. To test if Pol V transcription overlaps CG methylation, we analyzed published whole genome bisulfite

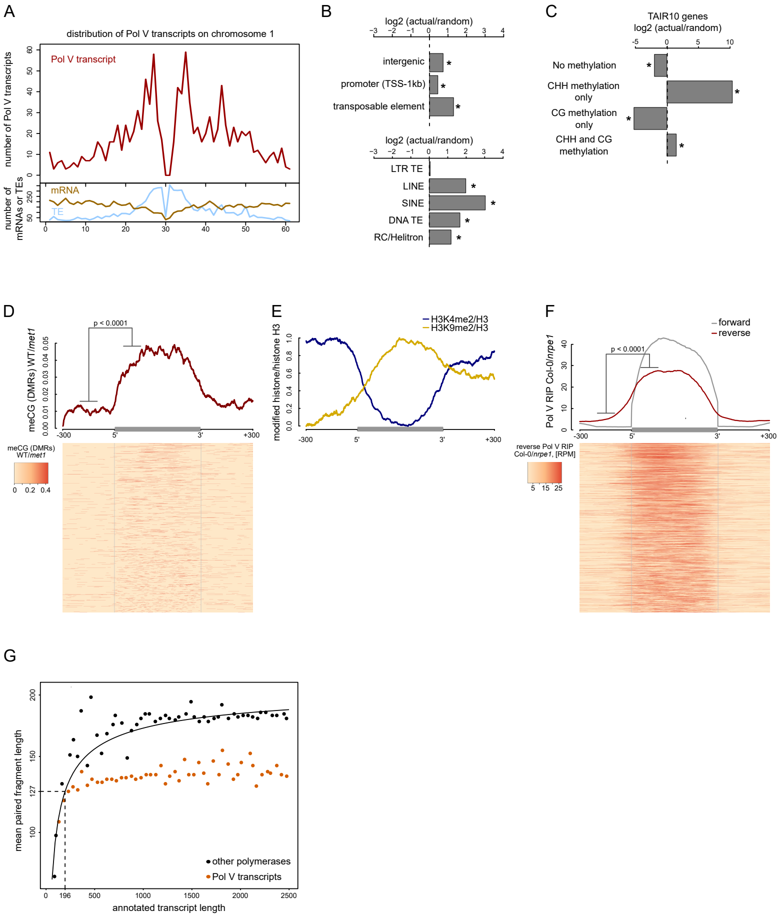

Figure 2.3: **Pol V regulatory elements:** **(A)** Pol V transcripts are produced from both pericentromeric regions and chromosome arms. The number of mRNAs, transposons (TAIR10) or Pol V transcripts was plotted on chromosome 1 in 500 kb windows. **(B)** Pol V transcripts are significantly enriched on promoters, intergenic sequences, and transposons of all families except LTR transposons. Plots show ratios of features overlapping Pol V transcripts to those overlapping randomized genomic regions. Stars denote significant differences based on permutations ($p<0.001$). **(C)** CG methylation is not sufficient to mediate Pol V transcription. Genes annotated in TAIR10 were split into four categories based on the presence of CHH methylation (greater than 2%) and CG methylation (greater than 10%). Enrichment of annotated Pol V transcripts on those categories of genes was calculated by comparing the actual overlap with overlaps of random genomic loci. Stars denote $p<0.004$. **(D)** MET1-dependent CG methylation is enriched within Pol V-transcribed regions. Average CG methylation levels [120] within differentially methylated regions (DMRs) were plotted on scaled Pol V transcripts. **(E)** A repressive histone modification is enriched on Pol V transcribed regions. Profiles of average enrichment of the modified histone (H3K9me2 and H3K4me2) over histone H3 were plotted on scaled Pol V transcripts. Enrichment of H3K9me2 and depletion of H3K4me2 were statistically significant ($p<0.0066$and $p<0.0001$, respectively; permutation test). **(F)** Pol V transcribes bidirectionally. Profiles of averaged Pol V RIP-seq signal (Col-0/*nrpe1*) in forward (grey) or reverse orientation (red) on scaled Pol V transcripts. Forward strand refers to annotated transcripts, reverse strand refers to the strand opposite to the annotated transcripts. **(G)** Annotated Pol V transcripts are composed of multiple shorter RNAs. Lengths of paired-end RNA fragments sequenced in RIP-seq mapping to nuclear and organellar genes (TAIR10) or to Pol V transcripts were compared to sizes of full length RNAs derived from annotations (TAIR10).

sequencing datasets [120]. DNA methylation in the CG context was increased throughout the genomic regions transcribed by Pol V but not outside of those regions (Figure 2.3D). Even though most Pol V transcripts were enriched in MET1-dependent CG methylation, the levels of CG methylation were not sufficient to predict the levels of Pol V transcription (Figure 2.4). This observation shows that CG methylation and Pol V transcription overlap and provides support for preexisting CG methylation being an important factor in guiding Pol V to specific loci. However, it also indicates that CG methylation does not regulate the level of Pol V transcription and that CG methylation is not needed upstream of transcription initiation sites. Instead it may be required within the transcribed regions.

We also tested if Pol V transcription overlaps with various posttranslational histone modifications [122, 124, 123] and found that H3K9me2 overlapped Pol V- transcribed regions in a way similar to CG methylation, while H3K4me2 was depleted on Pol V-transcribed regions (Figure 2.3E). H3K4me3, H3K36me3, and H3K9ac also appeared depleted but due to a higher noise level this depletion was not significant (Figure 2.4). Although it is unknown which histone modifications are controlling Pol V transcription and which are established in a Pol V-dependent manner, this is consistent with Pol V being guided to its genomic targets by repressive chromatin modifications present within the transcribed regions.

The overlap between CG methylation and Pol V transcription suggests that CG methylation may be required for Pol V transcription. To test this possibility, we assayed the accumulation of six individual Pol V transcripts in the *met1* mutant and *suvh4/5/6* triple mutant (Figure 2.4). Five loci showed a significant reduction in the accumulation of Pol V transcripts in the met1 mutant, which is consistent with a requirement of MET1 for Pol V transcription and with previous reports [100]. The *suvh4/5/6* mutant had a reduced accumulation of Pol V transcripts at two loci (Figure 2.4), which indicates that SUVH4, 5, and 6 (and H3K9me2 they presumably establish) may have a more subtle or locus-specific effect on Pol V transcription. One of the tested loci showed a significant increase in RNA

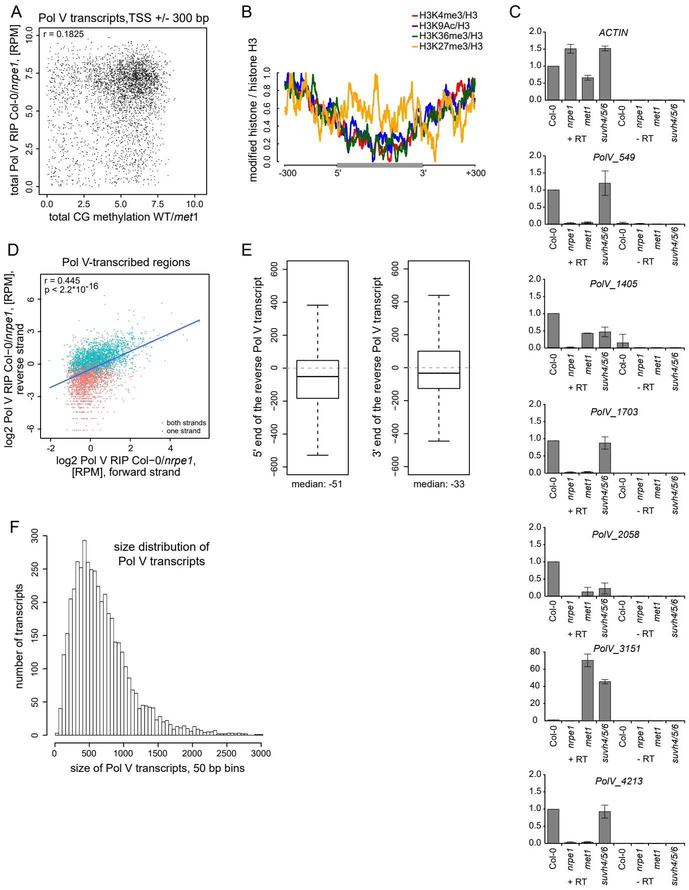

Figure 2.4: **Pol V regulatory elements (Supplementary): (A)** CG DNA methylation levels do not correlate with the levels of Pol V transcripts. Scatterplot shows CG methylation levels (wild-type/met1) [120] and Pol V RIP signal (total Col-0/*nrpe1*, [RPM]) around the 5' ends (+/- 300 bp) of Pol V transcripts. **(B)** Histone modifications on Pol V transcribed regions. Profiles of average enrichment of modified histones (H3K4me3, H3K9ac, H3K36me3, and H3K27me3 [122, 123] over histone H3 [122] were plotted on scaled Pol V transcripts +/- 300 bp. (C) Accumulation of Pol V transcripts is partially affected in *met1* and *suvh4/5/6* mutants. RT-qPCR for selected Pol V transcripts was performed in Col-0, *nrpe1*, *met1*, and *suvh4/5/6*. Averages signal levels relative to wild type and standard deviations from three biological replicates are shown. (D) Most regions are transcribed by Pol V on both strands. Scatterplot shows Pol V RIP-seq signal (Col-0-*nrpe1*, [RPM]) from the forward (annotated) and reverse (opposite to annotated) strand on Pol V-transcribed regions. Regions containing annotated transcripts on one or two strands are shown with different colors. Trend line (blue), correlation coefficient and p value have been calculated using linear regression. (E) The start and end sites of forward (annotated) and reverse (opposite to annotated) Pol V transcripts are shifted. Boxplot (left) showing the distance between the 5'-end of the reverse to the 3'-end of the forward Pol V transcript from the same genomic region (median distance -51 bp). Boxplot (right) showing the distance between the 3'-end of the reverse to the 5'-end of the forward Pol V transcript from the same genomic region (median distance -33 bp). (F) Size distribution of Pol V transcripts in 50 bp bins. The median size of Pol V transcripts is 689 bp

accumulation in both mutants (Figure 2.4), which may be attributed to the loss of Pol II silencing. These results are consistent with CG methylation being required for Pol V transcription.

The requirement of CG methylation for Pol V binding to chromatin and transcription may be interpreted as evidence of CG methylation recruiting Pol V to specific loci in the genome. This would predict that CG methylation should be sufficient for Pol V transcription. Alternatively, CG methylation may be one of many factors working together to determine the specificity of Pol V. To distinguish between these possibilities, we analyzed protein-coding genes, which show gene body CG methylation [125]. Pol V transcripts were significantly depleted on body-methylated genes, identified by high levels of CG methylation and low levels of CHH methylation (Figure 2.3C). In contrast, Pol V transcripts were enriched on genes having high levels of CHH methylation (Figure 2.3C), which are likely caused by intronic transposons [126]. This suggests that CG methylation is not sufficient for Pol V transcription and consequently, CG methylation is one of many factors involved in guiding Pol V to specific genomic loci.

The possibility that preexisting repressive chromatin modifications guide Pol V predicts that this polymerase should transcribe both strands of DNA. To test this prediction, we plotted forward and reverse Pol V-RIP signal on aligned and scaled Pol V transcripts. We found that, indeed, Pol V transcribed bidirectionally on annotated Pol V transcripts (Figure 2.3F) and that transcription levels on both strands were somewhat correlated (Figure 2.4). Furthermore, transcription on both strands was shifted with annotated Pol V transcripts on the reverse strand starting 51 bp before annotated transcripts on the forward strand end (Figure 2.4). These results support the idea that internal chromatin modifications may be important for Pol V transcription. Furthermore, they also suggest that transcription on one strand and, possibly, subsequently deposited chromatin modifications may be important for the initiation of transcription on the other strand.

If internal repressive chromatin modifications control Pol V transcription, one might predict that Pol V could have multiple transcription initiation sites within one transcribed region. This would manifest itself in the presence of several shorter RNAs within most annotated Pol V transcripts. Alternatively, if Pol V had an external promoter, we would expect the presence of one predominant RNA with a transcription start site close to the beginning of the annotated transcript. Our RIP-seq protocol included sonication and random priming steps, which preclude us from directly capturing the ends of intact RNAs. However, we performed paired-end RNA sequencing, which provides an alternative way to distinguish between these two possibilities. To do so, we mapped the paired-end reads to all *Arabidopsis* transcripts annotated in TAIR10 as well as to Pol V transcripts. We then plotted the relationship between the transcript length and the mean size of paired-end sequenced RNA fragments mapping to this transcript (Figure 2.3G). As the RIP-seq datasets include significant amounts of background reads originating from polymerases other than Pol V, we were able to determine the relationship between the size of each transcript known from TAIR10 annotations and the mean length of mapped read-pairs (Figure 2.3G). Small Pol V transcripts also followed this relationship, however, longer Pol V transcripts did not produce longer sequenced fragments (Figure 2.3G). This indicates that actual RNAs produced from annotated Pol V transcripts are shorter than the size of these annotations. If the relationship between transcript length and sequenced RNA fragments is the same or at least similar for Pol V as it is for other DNA-dependent RNA polymerases, we would predict an RNA length of 196 nt based on the median paired end fragment obtained from the first repeat of Pol V RIP-seq (Figure 2.3G). A similar analysis of the second biological repeat of Pol V RIP-seq predicts a median RNA length of 205 nt. Considering that the median length of annotated Pol V transcripts is 689 nt (Figure 2.4), this indicates that annotated Pol V transcripts contain more than one transcription initiation and/or termination site. This is not only consistent with Pol V being controlled by internal promoters but also demonstrates that Pol V transcripts annotated in our study are not individual continuous transcriptional units but rather regions of Pol V transcriptional activity.

Overall, our analysis suggests that Pol V is controlled by internal promoters, similar to what has been reported for a subset of Pol III transcripts [127]. Although any involvement of DNA sequence elements cannot be excluded at this time, our data are consistent with repressive chromatin modifications being at least important for Pol V recruitment and possibly working as a functional equivalent of a promoter.

2.3.4 AGO4 binds most Pol V transcripts

Pol V is required for AGO4 binding to chromatin genome-wide [57], however, it is unknown if AGO4 associates with Pol V transcripts beyond the handful of loci tested so far [88, 90]. To test if the association with Pol V transcripts is a general feature of AGO4, we performed RIP-seq using an antibody against AGO4 [88] in Col-0 wild type, *ago4*, and *nrpe1*. Because the library prep method we used does not efficiently amplify siRNAs, this approach should specifically detect long RNAs associated with AGO4. AGO4 RIP-seq signal was significantly enriched on Pol V transcripts (Figure 2.5A,B) and this binding was dependent on Pol V (Figures 2.5B, Figure 3—figure supplement 1). The presence of AGO4 RIP-seq signal on most transcripts shown on the heatmap (Figure 2.5A) indicates that AGO4 associates with most if not all Pol V transcripts. This is further supported by a significant correlation between Pol V and AGO4 RIP-seq signals (Figure 3—figure supplement 1). Additionally, annotations based on AGO4 RIP-seq yielded transcripts with start and end sites similar to Pol V transcripts and overlapped regions with hallmarks of RdDM [128, 120, 57] (Figure 2.5C, Figure 3—figure supplement 1). It should be noted that the RIP assay includes formaldehyde crosslinking, which may preserve indirect interactions. Therefore, the association we observed may reflect direct physical interactions or indirect interactions with other proteins or nucleic acids in between AGO4 and lncRNAs.

AGO4 was shown to interact with Pol II on a limited number of loci where Pol II transcripts have been suggested to fulfill a role similar to Pol V transcripts [129]. To test if AGO4 binds to RNAs produced by polymerases other than Pol V, we identified RNAs

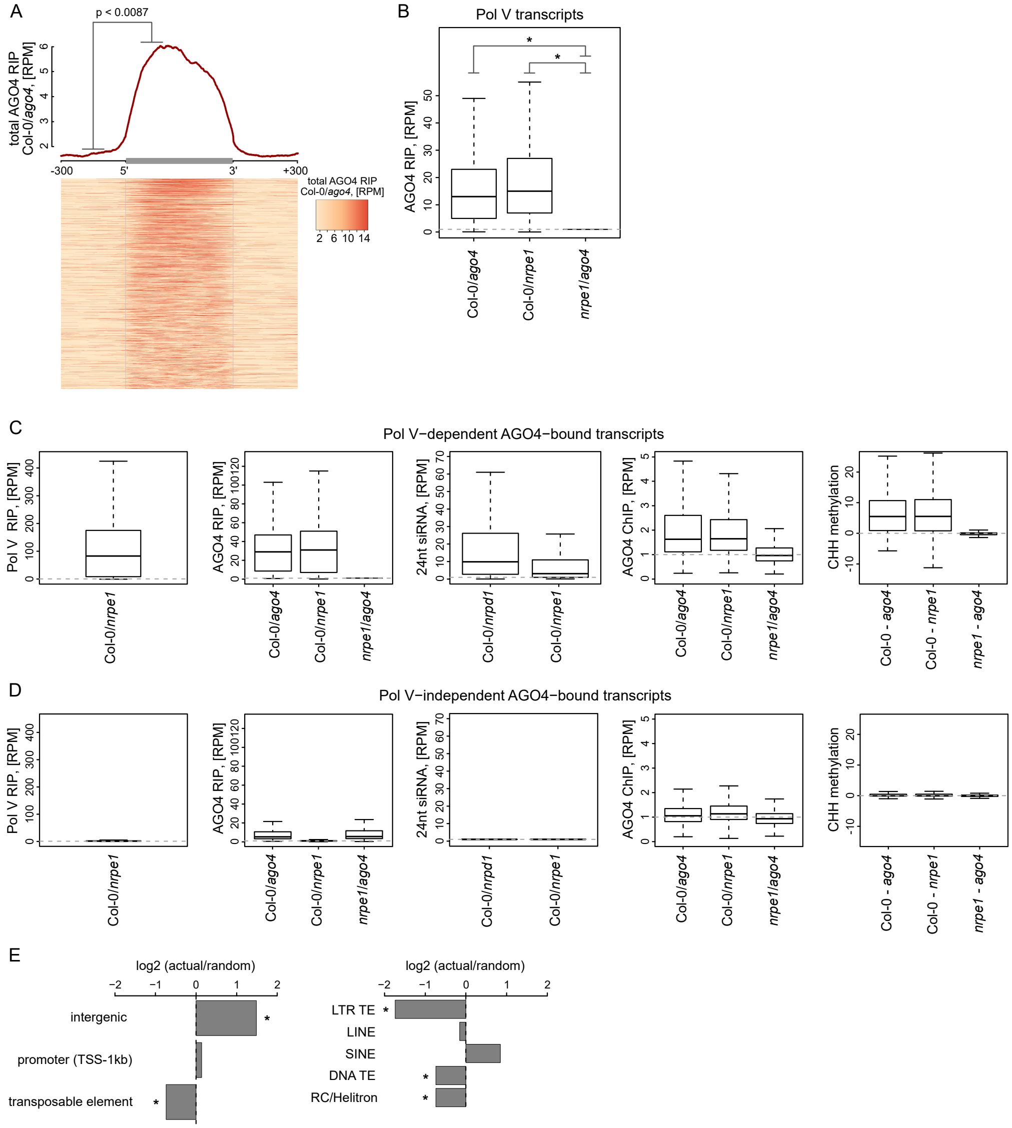

Figure 2.5: **AGO4 binds most Pol V transcripts:** **(A)** AGO4 RIP-seq signal (Col-0/*ago4*) is enriched on the majority of Pol V transcripts. Total AGO4 RIP signal was plotted on scaled Pol V transcripts. The p value was calculated using the permutation test by comparing 100 nt long regions starting 200 nt upstream and 50 nt downstream of 5' ends of the annotated transcripts. **(B)** Binding of AGO4 to Pol V transcripts depends on Pol V. The box plot shows AGO4 RIP-seq signal on Pol V transcripts. Stars denote p<2.2 * 10–16 (Wilcoxon test). **(C)** Pol V-dependent association of AGO4 with RNA is correlated with RdDM. Boxplots show signal levels for Pol V RIP-seq, AGO4 RIP- seq, 24nt siRNA, AGO4 ChIP-seq, and CHH methylation. Transcript were called using AGO4 RIP-seq Col-0/*ago4* and considered Pol V-dependent if AGO4 RIP-seq Col-0/*nrpe1* >= 4. **(D)** Pol V-independent association of AGO4 with RNA is not correlated with RdDM. Box plots show signal levels for Pol V RIP-seq, AGO4 RIP-seq, 24nt siRNA, AGO4 ChIP-seq, and CHH methylation. Transcripts were called using AGO4 RIP-seq Col-0/*ago4* and considered Pol V-independent if AGO4 RIP-seq *nrpe1*/*ago4* >= 4. **(E)** Pol V-independent transcripts bound by AGO4 are enriched on intergenic sequences but are depleted on all transposons except SINEs. Plots show ratios of features overlapping transcripts to those overlapping randomized transcripts. Stars denote p<0.001 (permutation test).

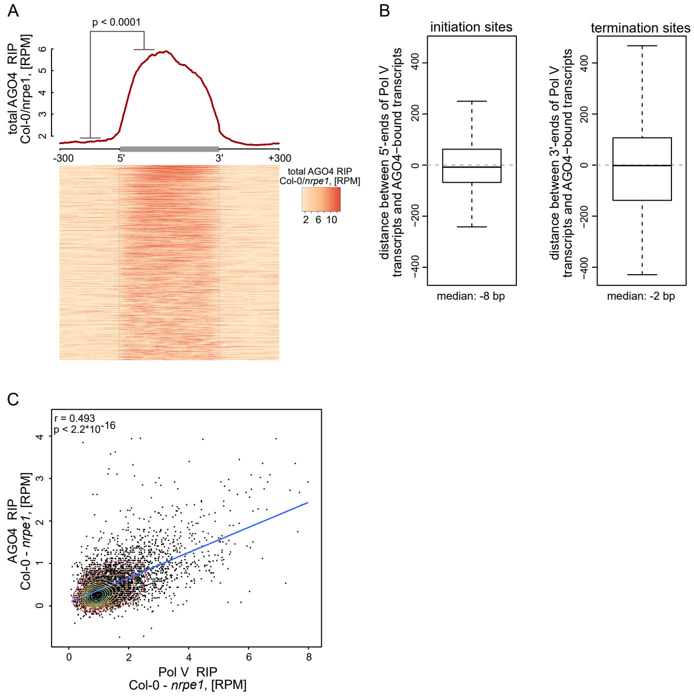

Figure 2.6: **AGO4 binds most Pol V transcripts (Supplementary): (A)** AGO4 binds Pol V transcripts in a Pol V-dependent manner. Profile and heatmap for scaled Pol V transcripts +/- 300 bp. The profile represents averaged total AGO4-RIP Col-0/*nrpe1* signal [RPM]. In the heatmap every row represents an individual Pol V transcript (sorted by size). The p-value was calculated using the permutation test by comparing 100 nt long regions starting 200 nt upstream and 50 nt downstream of 5' ends of the annotated transcripts. (B) Close overlap between start and end sites of transcripts annotated in AGO4 and Pol V RIP-seq. Left boxplot shows the distance between the initiation sites of overlapping transcripts with the same orientation annotated based on Pol V RIP-seq and AGO4 RIP-seq, respectively (median distance: -8 bp). Right boxplot shows the distance between the 3'-ends of overlapping transcripts with the same orientation called in Pol V RIP-seq and AGO4 RIP-seq, respectively (median distance: -2 bp). (C) Intensity of AGO4 RIP-seq signal is correlated with the intensity of Pol V RIP-seq signal. Scatterplot shows AGO4 RIP-seq signal and Pol V RIP-seq signal (repeat 1) on annotated Pol V transcripts. The plot shows a trend line calculated using linear regression (blue) as well as Spearman correlation coefficient and its p-value.

whose association with AGO4 did not depend on Pol V. They were characterized by RIP signal present in Col-0 wild type and *nrpe1* but not in *ago4* (Figure 2.5D) and were depleted on transposons and enriched on intergenic sequences and promoters (Figure 2.5E). These RNAs were also enriched on Pol III-transcribed SINE elements (Figure 2.5E), which suggests that AGO4 may be binding Pol III transcripts. Regions generating those RNAs were not only not transcribed by Pol V but also failed to show any of the hallmarks of

RdDM i.e. CHH methylation [120], presence of siRNAs [128] or AGO4 binding to chromatin [57] (Figure 2.5D). This suggests that association of AGO4 with RNAs produced by a polymerase other than Pol V does not lead to RdDM. Therefore, this interaction is either non-specific or reflects functions of AGO4 independent of RdDM. We conclude that AGO4 associates with most Pol V transcripts and may have an additional role not related to RdDM.

2.3.5 AGO4 and IDN2 enhance the accumulation of Pol V transcripts

Widespread association of AGO4 with Pol V transcripts suggests that Pol V transcripts may be a substrate for AGO4 slicer activity [130]. Alternatively, AGO4 could function without slicing Pol V transcripts. To distinguish between these possibilities, we performed RIP-seq with the anti-NRPE1 antibody in the *ago4* mutant. Presence of slicing by AGO4 would predict longer and more abundant transcripts in the *ago4* mutant. We observed none of those effects (Figure 2.7A,B, Figure 2.8). In contrast, accumulation of Pol V transcripts was decreased in the *ago4* mutant compared to Col-0 (Figure 2.7A,B, Figure 2.8). Moreover, analysis of the lengths of paired-end sequencing reads in the *ago4* mutant predicted an average transcript length of 200 nt, which is very similar to the size predicted for Col-0 wt (Figure 2.3G). These results indicate that AGO4 does not slice Pol V transcripts. Instead, AGO4 seems to enhance Pol V transcription or to stabilize Pol V transcripts. This is consistent with slicing activity being dispensable for Ago recruitment to chromatin in *S. pombe* [131]. Alternatively, AGO4 slicing products originating from Pol V transcripts could be undetectable in our assay due to their size or loss of association with the Pol V complex.

Another protein shown to associate with Pol V transcripts is IDN2 [90, 42]. Although the biochemical function of IDN2 remains unknown, it could potentially affect the stability of Pol V transcripts. We tested this possibility by performing RIP-seq with the anti-NRPE1 antibody in the *idn2* mutant. Accumulation of Pol V transcripts was also reduced in *idn2*

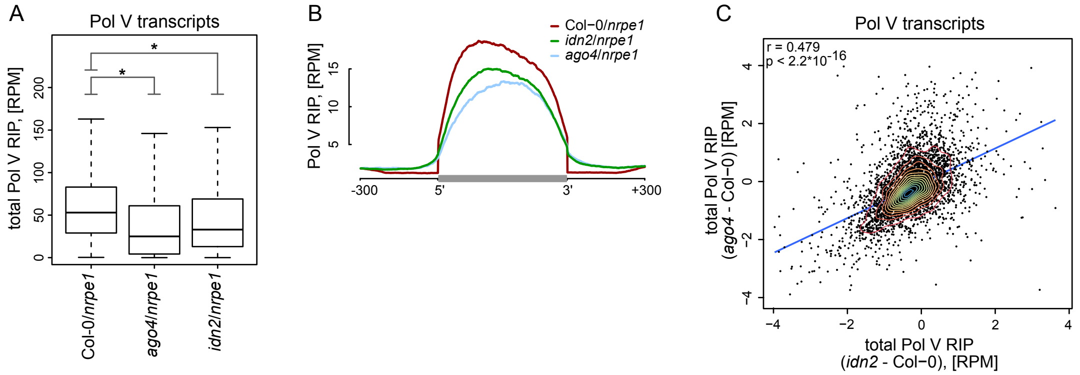

Figure 2.7: **AGO4 and IDN2 enhance the accumulation of Pol V transcripts:** **(A)** Accumulation of Pol V transcripts is reduced in *ago4* and *idn2*. Box plots show ratios of Pol V RIP-seq signals on Pol V transcripts in various genotypes. Stars denote p<2.2 * 10–16 (Wilcoxon test). **(B)** Accumulation of Pol V transcripts is reduced in *ago4* and *idn2* over the entire lengths of Pol V transcripts. Average Pol V RIP-seq enrichment was plotted on scaled Pol V transcripts. Differences between Col-0 and *ago4* as well as Col-0 and *idn2* are significant when measured between positions 50 nt and 150 nt downstream of 5' ends of Pol V transcripts (p<0.0001, permutation test). **(C)** On most Pol V transcripts, Pol V-transcription is affected in a similar way in *ago4* and *idn2*. Scatterplot of total Pol V RIP signal in *ago4* - Col-0 vs. *idn2* - Col-0. The plot shows a trend line calculated using linear regression (blue) as well as Pearson correlation coefficient and its p value.

(Figure 2.7A, Figure 2.8) over the entire lengths of the annotated transcripts (Figure 2.7B). This indicates that, like AGO4, IDN2 also enhances Pol V transcription or increases the stability of Pol V transcripts on both strands (Figure 2.8). Effects observed in *ago4* and *idn2* were somewhat correlated (Figure 2.7C) and could not be explained by an overall reduction of RNA levels in *ago4* and *idn2* as we did not observe a similar decrease on mRNAs (Figure 2.8). This suggests that mutations in AGO4 and IDN2 could affect the stability of Pol V transcripts. Alternatively, these mutations could indirectly affect Pol V transcription by causing a general reduction in repressive chromatin modifications on Pol V transcribed regions, which, in turn, would reduce the rate of Pol V transcription. Overall, our results show that AGO4 and IDN2 are unlikely to contribute to slicing or other forms of degradation of Pol V transcripts.

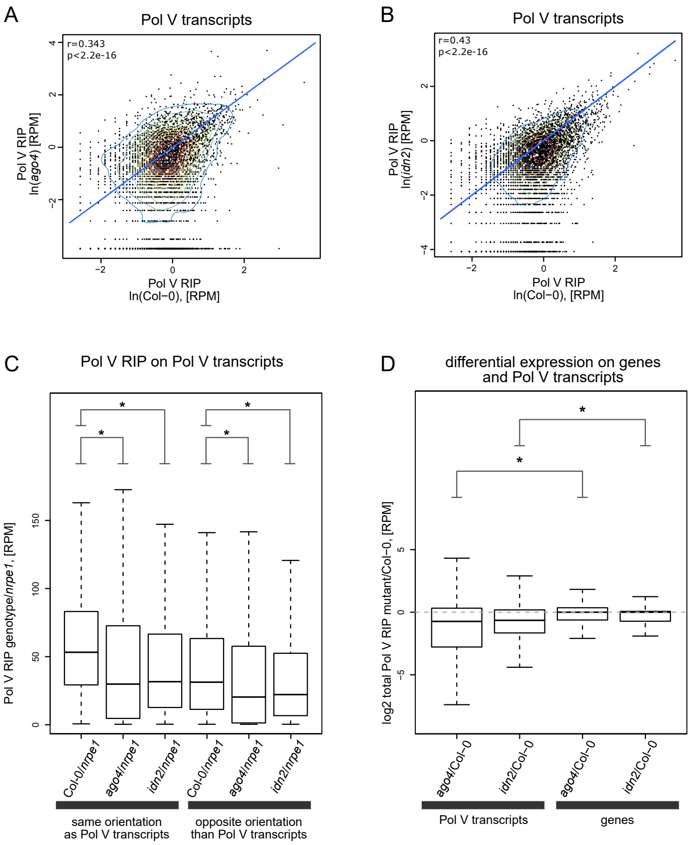

Figure 2.8: **AGO4 and IDN2 enhance the accumulation of Pol V transcripts (Supplementary): (A)** Accumulation of Pol V transcripts is reduced in the *ago4* mutant. Scatterplot shows read counts from Pol V RIP-seq in Col-0 wild type and *ago4*. The plot shows a line where read counts are equal in both genotypes (blue). Pearson correlation coefficient and its p value are calculated using linear regression. (B) Accumulation of Pol V transcripts is reduced in the *idn2* mutant. Scatterplot shows read counts from Pol V RIP-seq in Col-0 wild type and *idn2*. The plot shows a line where read counts are equal in both genotypes (blue). Pearson correlation coefficient and its p-value are calculated using linear regression. (C) Forward and reverse transcripts are reduced in *ago4* and *idn2* on Pol V-transcribed regions, suggesting that AGO4 and IDN2 are important to stabilize transcripts coming from both strands. Box plots show Pol V RIP signal calculated using reads with the same or opposite orientation than Pol V transcripts. Stars denote p<2.2 * 10–16 (Wilcoxon test). (D) Pol V transcript levels are reduced in *ago4* and *idn2*. mRNA levels are on average not reduced in Pol V RIP samples obtained in *ago4* and *idn2* compared to Col-0, suggesting that the reduction of Pol V transcripts in the mutant backgrounds is not an artefact. Box plots show Pol V RIP signal (mutant/Col-0) on Pol V transcripts and genes. Stars denote p<2.2 * 10–16 (Wilcoxon test).

2.3.6 RdDM is restricted to Pol V-transcribed regions

The current models of RdDM show that AGO4, IDN2, and other RNA-binding proteins interact with Pol V transcripts at some distance from the transcribing core Pol V complex.

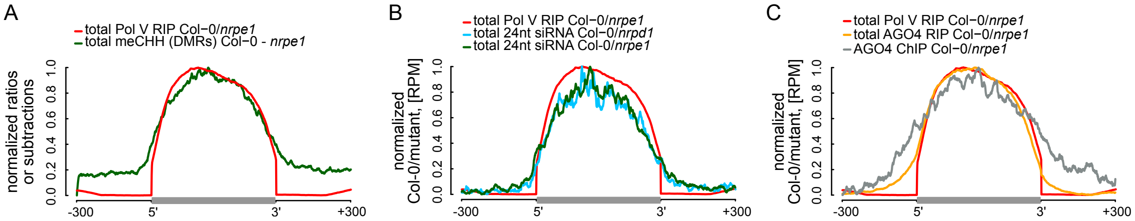

Figure 2.9: **RdDM is restricted to Pol V-transcribed regions:** **(A)** CHH methylation dependent on Pol V closely overlaps Pol V transcription. Average CHH methylation levels within differentially methylated regions (DMRs) were plotted on scaled Pol V transcripts ± 300 bp. Average Pol V RIP-seq signal (Figure 2.1B) was plotted as a reference. **(B)** siRNAs closely overlap Pol V transcription. Average enrichment of 24nt siRNA (Col-0/*nrpd1* and Col-0/*nrpe1*) was plotted on scaled Pol V transcripts. Average Pol V RIP-seq signal (Figure 2.1B) was plotted as a reference. **(C)** AGO4 binds to Pol V transcripts and corresponding DNA over the entire regions transcribed by Pol V. Average signals of AGO4 RIP-seq and AGO4 ChIP-seq Col-0/*nrpe1* were plotted on scaled Pol V transcripts. Average Pol V RIP-seq signal (Figure 2.1B) is shown as a reference.

This distance is allowed by the lengths of lncRNA and the C-terminal domain of Pol V which interacts with AGO4 [132, 133]. Although this spatial separation between Pol V and downstream factors has not been addressed experimentally, it predicts that the DNA methylation machinery may have some level of spatial flexibility especially over densely packed chromatin relative to the progressing position of the core Pol V polymerase complex. This flexibility could result in DNA methylation being established outside of the regions transcribed by Pol V. To test this possibility, we plotted Pol V-dependent CHH methylation [120] on DNA sequences corresponding to Pol V transcripts. CHH methylation was significantly enriched within these sequences (Figure 2.9A, Figure 2.10). However, only trace levels of CHH methylation were observed outside (Figure 2.9A, Figure 2.10). This result shows that at least in genomic regions in close proximity to the annotated Pol V transcripts, the RdDM pathway is only able to deposit DNA methylation within regions transcribed by Pol V.

We further tested if other components of RdDM are present outside of DNA sequences corresponding to Pol V transcripts. Analysis of previously published smRNA datasets [128] indicated that Pol IV- and Pol V-dependent 24nt siRNAs mostly accumulated within regions transcribed by Pol V (Figure 2.9B, Figure 2.10), indicating that Pol IV and Pol V likely

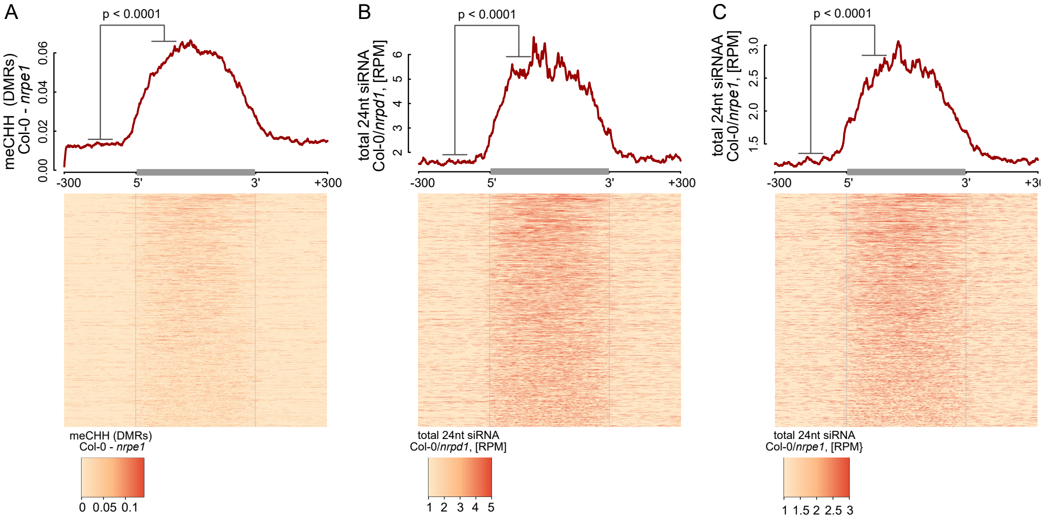

Figure 2.10: **RdDM is restricted to Pol V-transcribed regions (Supplementary): (A-C)** Pol IV- and Pol V-dependent 24 nt siRNAs as well as CHH methylation span the entire lengths of Pol V transcripts. Average signals of Pol V-dependent CHH methylation (*nrpe1* DMRs) **(A)** as well as Pol IV-dependent **(B)** and Pol V-dependent **(C)** 24 nt siRNA were plotted on scaled Pol V transcripts. Heatmaps show individual transcripts sorted by length. The p-values were calculated using the permutation test by comparing 100 nt long regions starting 200 nt upstream and 50 nt downstream of 5' ends of the annotated transcripts.

transcribe the same genomic regions. Similarly, AGO4 associated with chromatin [57] mostly within Pol V transcribed regions (Figure 2.9C). These results show that, at least in genomic regions in close proximity to the annotated Pol V transcripts, several features of RdDM are predominantly restricted between the start and the end of Pol V transcription. Another and not mutually exclusive explanation of these results is that Pol V transcription is tightly limited to regions with preexisting DNA methylation. Both interpretations would be inconsistent with models assuming that the flexibility of Pol V transcripts could allow chromatin modifying enzymes to reach outside of the transcribed regions. Overall, these results are consistent with Pol V transcripts working in *cis* and mediating repressive chromatin modifications exclusively within transcribed regions.

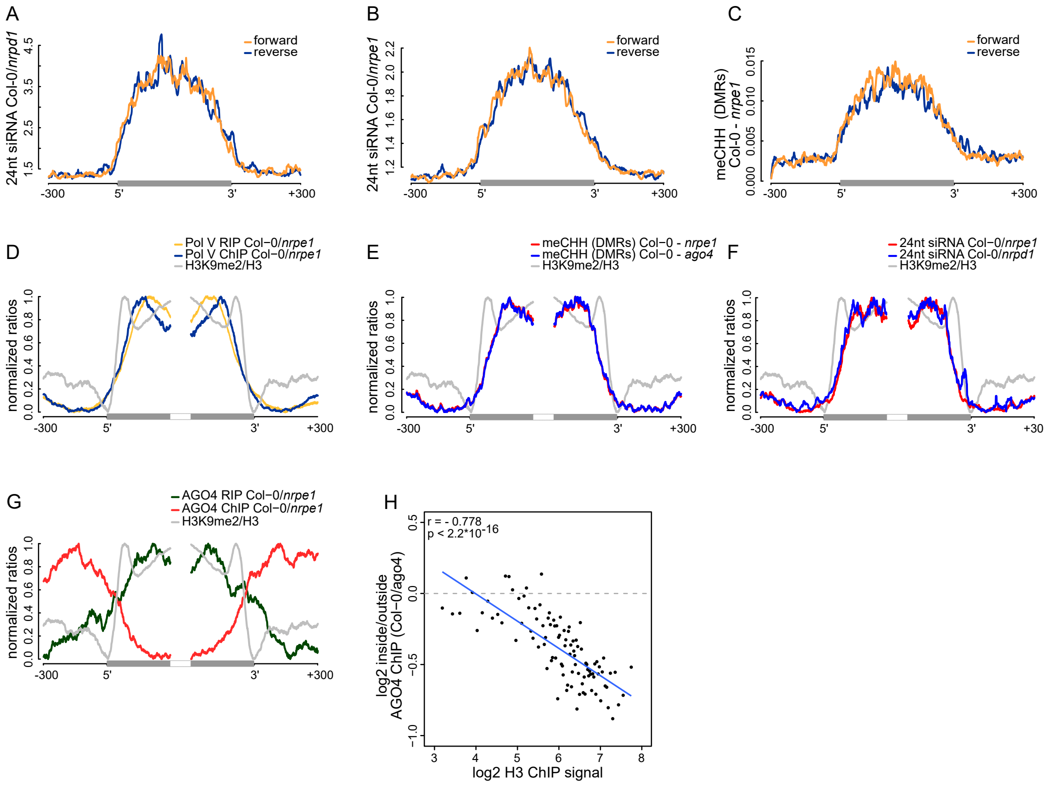

Figure 2.11: **Strand bias of RdDM and importance of AGO4 binding to Pol V transcripts: (A)** Pol IV-dependent 24nt siRNAs do not show a strand bias on Pol V transcripts. Average signal (Col-0/*nrpd1*, [RPM]) for reads with the same or the opposite orientation as Pol V transcripts was plotted on scaled Pol V transcripts ± 300 bp. **(B)** Pol V-dependent 24nt siRNAs do not show a strand bias on Pol V transcripts. Average signal (Col-0/*nrpe1*, [RPM]) for reads with the same or the opposite orientation as Pol V transcripts was plotted on scaled Pol V transcripts. **(C)** CHH methylation does not show a strand preference on Pol V transcripts. Average signal of called differentially methylated regions (DMRs) for CHH methylation (Col-0 - *nrpe1*) from the same or opposite strand as the Pol V transcript was plotted on scaled Pol V transcripts. **(D–G)** CHH methylation follows AGO4 interactions with Pol V transcripts on the edges of heterochromatic domains. **(D)** Pol V binds and transcribes DNA at the edges of heterochromatic domains. **(E)** CHH methylation is deposited on regions transcribed by Pol V. **(F)** 24nt siRNAs overlap Pol V transcripts. **(G)** AGO4 associates with RNA on regions transcribed by Pol V but association of AGO4 with DNA detectable by ChIP-seq is present outside of the heterochromatic domains. Profiles represent normalized average signals on heterochromatic domains (with H3K9me2) +/- 300 bp, aligned at the ends. In each panel, gray bars on the x-axis (H3K9me2 region) and gray profiles (H3K9me2/H3) are shown. **(H)** High nucleosome density prevents AGO4 from binding to DNA within heterochromatic domains. Scatterplot compares H3 ChIP-seq signal to AGO4 ChIP-seq signal outside or inside of heterochromatic domains. Heterochromatic domains were combined in 100 groups based on their H3 levels and plotted against the log2 value of AGO4 ChIP-seq inside/outside the H3K9me2 region. 'Outside' was defined as the 50 to 250 bp upstream of the left end of the heterochromatic domain, while 'inside' corresponds to 50 to 250 bp inside the heterochromatic domain. The plot shows a trend line calculated using linear regression (blue) as well as Pearson correlation coefficient and its p-value.

2.3.7 Strand bias of RdDM

The observed high spatial resolution of RdDM could be accompanied by a correlation between strand preference of Pol V transcripts and CHH methylation. Alternatively, Pol V transcripts may mediate the establishment of CHH methylation on both strands of DNA. Although Pol V tends to transcribe both strands of DNA (Figure 2.3F), these two scenarios can be distinguished because the levels of Pol V transcription are often not equal between both strands (Figure 2.4). To test if RdDM displays a strand preference on Pol V-transcribed regions, we separately plotted both strands of siRNAs [128] (Figure 2.11A,B) and DNA methylation [120] (Figure 2.11C) on Pol V-transcribed sequences. We observed no differences in mean signal strengths on either strand throughout the transcribed regions. We also found no correlations in strand preference between Pol V transcription, CHH methylation [120], and siRNAs [128] on Pol V transcribed regions (Figure 2.12). This indicates that there is no strand preference of DNA methylation relative to siRNAs or Pol V transcripts, which indicates that Pol V transcripts mediate the establishment of CHH methylation on both strands of DNA.

2.3.8 Importance of AGO4 binding to Pol V transcripts

Several key RdDM factors have been shown to interact with both DNA and RNA on silenced loci [90, 88]. It remains unknown, which interaction is more important for RdDM. To answer this question, we focused our analysis on heterochromatic domains where protein binding to DNA may be constrained by a high density of nucleosomes but Pol V can still transcribe. We identified heterochromatic domains with high density of H3K9me2 [122] and plotted various features of RdDM over their edges. Both Pol V binding to chromatin reported by ChIP-seq [56] and Pol V transcription observed by RIP-seq overlapped the heterochromatic domains, which is consistent with Pol V transcribing silenced genomic regions (Figure 2.11D). CHH methylation [120] was also enriched on the heterochromatic domains (Figure 2.11E). Similarly, AGO4 binding to Pol V transcripts ob-

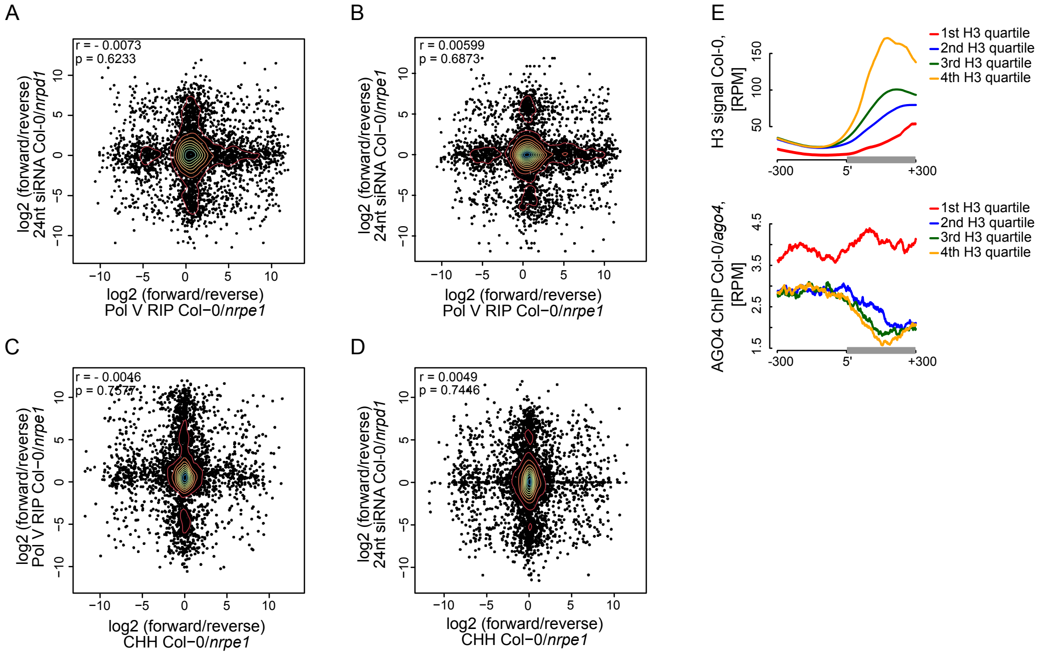

Figure 2.12: **Strand bias of RdDM and importance of AGO4 binding to Pol V transcripts (Supplementary): (A-D)** Strand preference of Pol V RIP on Pol V-transcribed regions is not correlated with strand-preference of 24nt siRNA or CHH methylation. Scatterplots show ratios (forward/reverse) of sequencing reads (Pol V RIP-seq and 24nt siRNA [128]) or CHH methylation [120] observed on Pol V transcripts. (A) Comparison of Pol V-transcription and Pol IV-dependent siRNAs. (B) Comparison of Pol V-transcription and Pol V-dependent siRNAs. (C) Comparison of Pol V-transcription and Pol V-dependent CHH methylation. (D) Comparison of Pol V-dependent CHH methylation and Pol IV-dependent siRNAs. (E) High nucleosome density prevents AGO4 from binding to DNA within heterochromatic domains. Heterochromatic domains were split into quartiles according to the strength of their internal H3 signal [122] (first 200 bp from the end). Profiles show the average H3 or AGO4 ChIP-seq (Col-0/*ago4*) [57] signal for ends of heterochromatic domains.

served by RIP-seq overlapped H3K9me2 [122], Pol V transcription, 24nt siRNAs [128], and CHH methylation [120] (Figure 2.11D–G). Interestingly AGO4 binding to DNA observed by ChIP-seq [57] was strongly enriched on chromatin flanking these heterochromatic domains (Figure 2.11G). This indicates that CHH methylation more closely follows AGO4 binding to Pol V transcripts than to DNA, suggesting that AGO4 interaction with Pol V transcripts may be the primary event directing downstream factors of RdDM. Binding of AGO4 to chromatin outside of heterochromatic domains could be explained by exclusion of protein binding to DNA by nucleosomes. To test this possibility, we grouped heterochromatic domains by strength of H3 ChIP-seq signal [122], which should indicate nucleosome density. Indeed, AGO4 binding [57] outside of the heterochromatic regions was correlated with nucleosome density (Figure 2.11H, Figure 2.12). Overall, these results show the central role of AGO4 interactions with Pol V transcripts in RdDM. Although the exact positions of various RdDM components within the silencing machinery and potential effects of formaldehyde crosslinking remain unknown, our data are consistent with a speculative model where downstream components of RdDM interact with Pol V transcripts in direct proximity to the Pol V complex and siRNAs base pair with RNA exiting the Pol V complex.

2.3.9 Pol V determines the edges of transposons

RdDM targets edges of transposons while interior regions of large transposons are silenced by other epigenetic mechanisms [134]. However, current mechanistic understanding of RdDM does not explain this preference towards the edges of transposons. One possibility is that Pol V preferentially transcribes the edges of transposons. Alternatively, Pol V could transcribe the entire lengths of transposons but siRNAs could only be produced on the edges. To distinguish between these possibilities, we plotted Pol V RIP-seq data on all transposons which overlap annotated Pol V transcripts. While short transposons were entirely transcribed by Pol V, longer TEs had a strong enrichment of Pol V transcription on their edges (Figure 2.13A). We further analyzed euchromatic transposons longer than 4 kb, similar to those studied by [134]. They also displayed a strong enrichment of Pol V transcription and Pol V-dependent CHH methylation [120] on their edges (Figure 2.13B, Figure 2.14). In contrast, regions inside the transposons appeared to be depleted in Pol V transcription compared to regions outside (Figure 2.13B). Pol IV transcripts [81] were also enriched on edges of both categories of transposons, however, they were not depleted inside the transposons (Figure 2.14). These results indicate that transcription by both Pol IV and Pol V are involved in targeting RdDM to the edges of transposons. Preferential

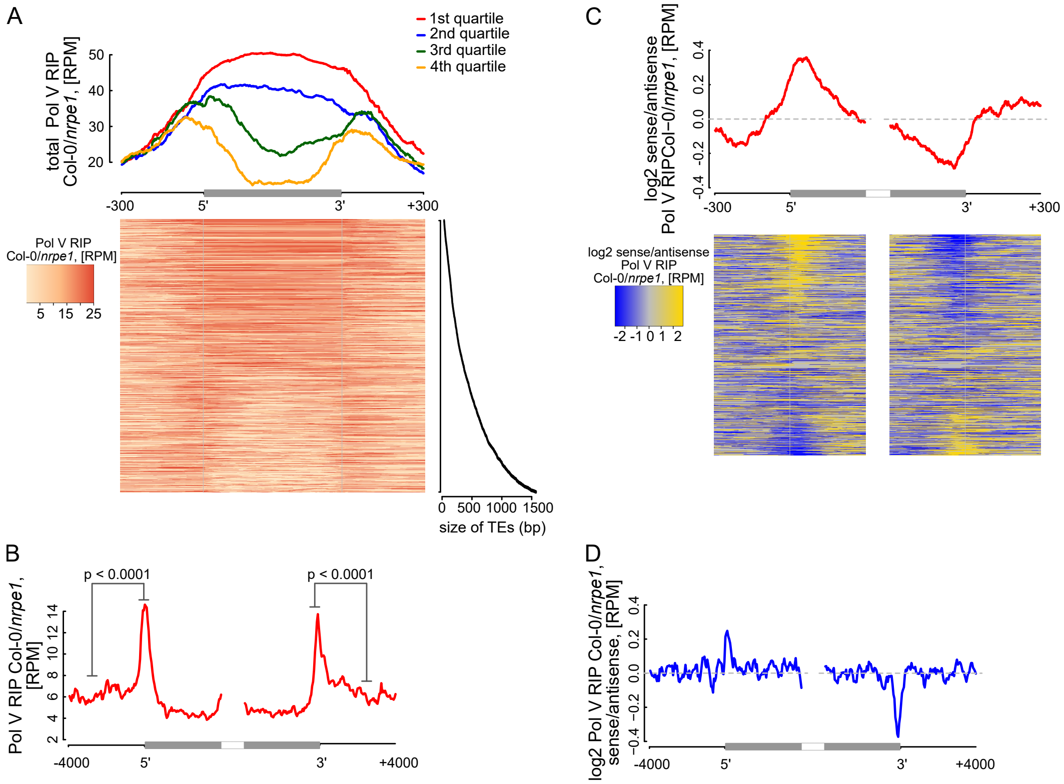

Figure 2.13: **Pol V determines the edges of transposons:** **(A)** Pol V transcripts are produced over the entire lengths of small transposons but are enriched at the edges of larger transposons. Annotated transposons (TAIR10) overlapping Pol V transcripts were split into quartiles according to their size (smallest to largest), scaled and the Pol V RIP-seq signal Col-0/*nrpe1* was plotted. Heatmap shows individual transposons sorted by size (sizes shown on the adjacent plot). **(B)** Pol V transcription is enriched on the edges of large transposons. Annotated euchromatic transposons greater than 4 kb were aligned by their 5'- and 3'-ends and average Pol V RIP-seq signal was plotted in 50 bp windows. The p value was calculated using the permutation test by comparing 500 nt long regions starting 1000 nt outside and 250 nt within the TEs. **(C)** Pol V transcribes into transposons. Transposons used in Figure 2.13A were aligned by their 5'- and 3'-ends and the average ratio of sense to antisense Pol V RIP-seq signal was plotted. Heatmaps show individual transposons sorted by the strength of transcription into the TEs. **(D)** Pol V transcribes into transposons. Transposons used in Figure 2.13B were aligned with their ends and the average ratio of sense to antisense Pol V RIP-seq signal was plotted.

transcription of transposon edges by Pol IV and Pol V suggests that these polymerases may be involved in determining the borders of silenced regions. Little is known about the mechanisms determining chromatin boundaries in plants, however, transcription is inherently directional and therefore could be involved in this process [135]. Although Pol V tends to transcribe both strands of DNA (Figure 2.3F), a subset of Pol V transcripts is enriched on

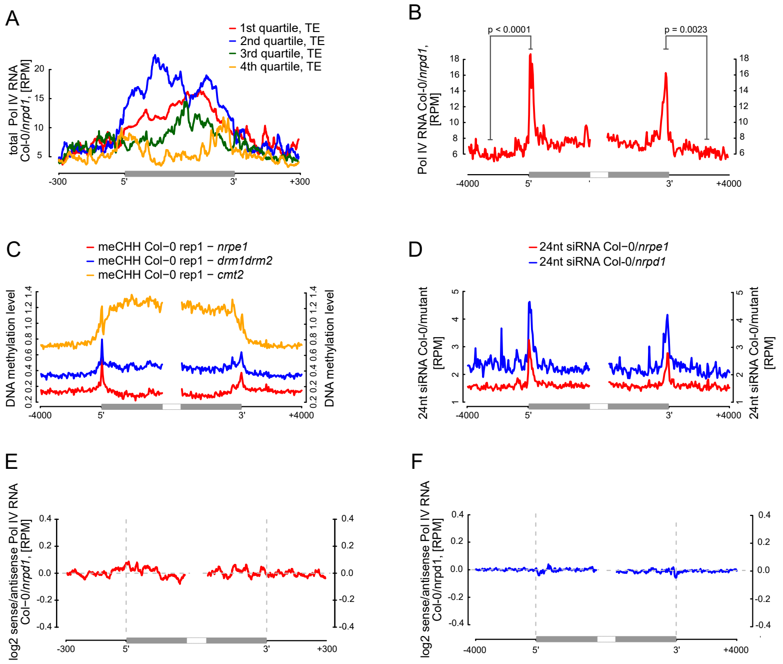

Figure 2.14: **Pol V determines the edges of transposons (Supplementary): (A)** Pol IV transcription on transposons. Annotated transposons (TAIR10) overlapping Pol V transcripts were split into quartiles according to their size (smallest to largest), scaled and the Pol IV RNA-seq signal Col-0/*nrpd1* [81] was plotted. (B) Pol IV transcription is enriched on the edges of large transposons. Annotated euchromatic transposons greater than 4 kb were aligned by their 5'- and 3'-ends and average Pol IV RNA-seq signal [81] was plotted in 50 bp windows. The p values were calculated using the permutation test by comparing 500 nt long regions starting 1000 nt outside and 250 nt within the TEs. (C) Pol V- and DRM1/DRM2-dependent CHH methylation are restricted to the edges of large TEs (>4 kb), while CMT2 is more important for CHH methylation inside of TEs. Average CHH methylation levels [120] were plotted in 50 bp windows on the 5'- and 3'-ends (+/- 4 kb) of large transposons. (D) Pol V-dependent 24 nt siRNAs and Pol IV-dependent 24 nt siRNAs are limited to the edges of large transposons (>4 kb). Plots represent average signal on the 5'- and 3'-ends (+/- 4 kb) of large transposons. Ratios of Pol IV- or Pol V-dependent 24 nt siRNAs [128] were plotted. (E) Pol IV transcription shows no strand preference on transposon edges. Transposons used in Figure 2.13A were aligned by their 5'- and 3'-ends and the average ratio of sense to antisense Pol IV RNA-seq signal [81] was plotted. (F) Pol IV transcription shows no strand preference on transposon edges. Transposons used in Figure 2.13B were aligned with their ends and the average ratio of sense to antisense Pol IV RNA-seq signal [81] was plotted.

one strand (Figure 2.12) indicating that limited strand preference of Pol V may be involved in determining boundaries of heterochromatin. To test this hypothesis, we determined the ratio of strand preference of Pol V transcription relative to the orientation of transposons. We first analyzed transposons selected for the presence of overlaps with Pol V transcripts.

Pol V transcription on the 5'-ends of transposons was enriched on the sense strand while Pol V transcription on the 3'-ends of transposons was enriched on the antisense strand. This indicates that Pol V transcripts showed an enrichment of Pol V transcription into the transposons at both ends (Figure 2.13C). Similarly, euchromatic transposons longer than 4 kb also showed enrichment of Pol V transcription into the transposons at both ends (Figure 2.13D). In contrast, datasets of Pol IV transcripts [81] did not show any evidence of strand preference (Figure 2.14). This is consistent with the tight physical and functional association of Pol IV with RDR2 [117, 80], which promptly converts Pol IV transcripts into double stranded RNA. Because accumulation of Pol IV transcripts requires RDR2 [81], the existence of strand preference of Pol IV transcription remains unknown. These results indicate that Pol V preferentially transcribes into transposons. Therefore, we propose that Pol V may play a key role in determining the boundaries of heterochromatin by transcribing into silenced regions.

2.4 Discussion

2.4.1 Regulation of Pol V transcription

Pol V is one of the critical factors determining the specificity of RdDM. Therefore, the promoter of Pol V and mechanisms regulating Pol V transcription may determine which regions of the genome are targeted by RdDM. Previous studies using ChIP-seq failed to identify conserved sequence elements which could be the Pol V promoter [56, 86]. Despite much higher resolution of RIP-seq we also did not identify any conserved external or internal sequence elements. Although *de novo* discovery of promoter elements is difficult even on Pol II genes [121] and requirement of short or variable sequence elements for Pol V transcription cannot be excluded, this is consistent with recent work showing the requirement of methyl-binding proteins SUVH2 and SUVH9 for Pol V binding to chromatin and transcription [100, 101]. Since SUVH2 and SUVH9 bind methylated DNA through their

SRA domains [136, 100], this indicates that pre-existing DNA methylation may serve as an equivalent of the Pol V promoter. We provide additional support for that possibility by showing a strong overlap between CG methylation and Pol V transcription. At the same time the accumulation of Pol V transcripts is not correlated with the levels of CG methylation and CG methylation is not sufficient to recruit Pol V. Moreover, Pol V transcription is limited to the edges of transposons, while DNA methylation usually spans the entire lengths of the transposable elements. This indicates that other unknown factors are involved in the determination and regulation of Pol V transcription. RNA polymerases I and II primarily use promoters located upstream of the transcription start sites, while Pol III mostly but not exclusively uses internal promoters [127]. Pol V seems to behave more like Pol III in being controlled by features present within the transcribed regions. This means that, although Pol V is derived from Pol II [137], it possibly uses a different transcription initiation machinery.

2.4.2 The RdDM effector complex

Our data also provide insights into the interplay between Pol V, Pol V transcripts, and associated proteins, which we refer to as the RdDM effector complex. First, our data demonstrate on a genome-wide scale that RdDM of endogenous loci is mostly restricted to regions transcribed by Pol V. This is consistent with data obtained in transgene or virus-induced silencing [138, 139, 140] and could be interpreted as evidence that siRNAs base pair with DNA and that Pol V only facilitates this interaction. Although this possibility cannot be excluded based on currently available data [119], we provide additional indirect support of base pairing between siRNAs and nascent Pol V transcripts. One reason why siRNAs-lncRNA base pairing seems to be a more likely scenario is lack of detectable strand preference between siRNAs, Pol V transcripts, and DNA methylation on Pol V-transcribed loci. Although such a preference has been reported on other categories of loci [114, 91], this discrepancy may be explained by the potential involvement of other RNA polymerases or other factors. Since we show that both strands of DNA are equally likely to be methylated even if siRNAs or Pol V transcripts show a strong strand preference, this could indicate that double stranded DNA is a substrate for DRM2. The simplest explanation for this observation would be base pairing between siRNAs and lncRNAs. Another result favoring siRNAs-RNA base pairing is our analysis of regions with high density of nucleosomes with H3K9me2. Although this analysis does not specifically take into account the activities of at least two putative ATP-dependent chromatin remodeling complexes involved in RdDM [105, 42], it indicates that CHH methylation more closely follows AGO4 interactions with RNA than with DNA. The exact architecture of the RdDM effector complex remains mysterious mostly because formaldehyde crosslinking does not allow distinguishing direct from indirect interactions. Our data show that this effector complex includes AGO4 throughout the genome, however AGO4 does not slice Pol V transcripts. Previous work indicates the involvement of IDN2 [90, 42], which also seems to be involved in enhancing the accumulation of Pol V transcripts. If, as argued above, siRNAs incorporated into AGO4 base pair with Pol V transcripts, restriction of DNA methylation to Pol V-transcribed regions suggests that this base pairing is likely to occur in close physical proximity to the Pol V complex. This is consistent with observations that AGO4 physically interacts with the C-terminal domain of Pol V [133]. Another important question about the RdDM effector complex is the involvement of RNA polymerases other than Pol V, which has been suggested by genetic evidence [129]. Our data do not provide evidence of Pol I, II or III functionally substituting for Pol IV in the *nrpe1* mutant. However, it remains possible that these other polymerases may be involved in the initiation of transcriptional silencing or work with Argonaute proteins other than AGO4 [141].

2.4.3 Determination of heterochromatin boundaries

Although RdDM has been implicated in various biological processes [55], its functions remain to some extent mysterious. It is especially true for its presumably main role of silencing transposons. Transposons are stably silenced by the combined action of three silencing pathways. Two of them are maintenance pathways which rely on CG methylation maintained by MET1 [114] and non-CG methylation maintained by CMT2 and CMT3 [24, 134]. RdDM on the other hand is capable of establishing DNA methylation *de novo*. Recently, yet another pathway has been implicated in directing silencing to active transposons [142]. Our data show that Pol IV and Pol V direct RdDM to edges of transposons. Moreover, Pol V preferentially transcribes into transposons, which indicates that Pol V may be involved in determining the boundaries of heterochromatin on transposable elements. This is reminiscent of the BORDERLINE lncRNA in *S. pombe* [143], however, unlike BORDERLINE, Pol V transcripts are more likely to maintain heterochromatin over the edges of transposons rather than to prevent the spreading of heterochromatin. The role of RdDM in the determination of heterochromatin boundaries has been recently studied in maize, where mutations in RdDM components MOP1, MOP2, and MOP3 enhance spreading of euchromatin from genes into nearby transposons [144]. Although this supports a role for RdDM in heterochromatin boundaries, the causal role of Pol V transcription in this process remains to be directly demonstrated. The mechanism responsible for preferential Pol V transcription into transposable elements remains unknown. One possibility is that the transition from euchromatin to heterochromatin facilitates Pol V transcription. This is however inconsistent with our data showing that Pol V transcription is controlled by multiple internal promoter-like features. Another explanation is the presence of more than one Pol V transcription initiation mechanism, with one mechanism responsible for unidirectional transcription on TE boundaries and the other mechanism mediating bidirectional transcription on all Pol V-transcribed loci.

2.5 Materials and Methods

2.5.1 Plant Material

Col-0 wild type, *nrpe1* (*nrpd1b-11*, [145], *ago4-1* (introgressed into the Col-0 background, [88, 87], *idn2-1* [89], *met1-3* [146], and *suvh4R203/suvh5-2/suvh6-1* [147, 148] plants were grown in soil in long-day conditions. For all experiments, approximately 2.5 weeks old seedlings were used.

2.5.2 Antibodies

The antibodies against the largest subunit of Pol V (NRPE1) or against AGO4 were described previously [79, 88].

2.5.3 RT-qPCR and RIP-seq

RT-qPCR experiments were performed in biological triplicates as reported in [149]. Oligonucleotides used for PCR are provided in 2.1. Fixation (0.5% formaldehyde) and RIP were performed according to the previously published protocol [149] up to step 37, using optimized amounts of protein A agarose beads coated with salmon sperm DNA in case of RIPs performed with the alpha-NRPE1 antibody [79] or Dynabeads protein A in case of alpha-AGO4 [88], respectively. Next, 1/10th vol. 3M NaOAc (pH 5.3), 2.5 vol. 96% ethanol and 1 ml NF-Pellet Paint were added to the samples and the inputs, precipitated overnight at -80C, and washed as described in steps 40 to 43 of the published protocol [149]. After resuspension in 10 ml milliQ water, the samples were digested with 5.8 u Turbo DNase I in the presence of 24 u of Ribolock RNase inhibitor at 25C for 30 min. The reaction was stopped by adding 3 ml 25 mM EDTA (pH 8.0) and incubating at 65C for 10 min. To ensure that the DNase I digest and the immunoprecipitation had been successful, 1/10th of the reaction was tested by RT-qPCR for the presence Pol V-transcripts and absence of genomic DNA as described in [149]. Minus-RT controls were included in all RT-qPCR assays and mock controls were included during protocol optimization experiments. The remaining 9/10th were precipitated and resuspended in 5 ml milliQ water for library generation.

Finally, library preparation was performed for mutant and wild-type samples by the University of Michigan Sequencing Core using the Illumina TruSeq Stranded mRNA Sample Prep Kit, replacing the heat fragmentation with an incubation step on ice for 5 min. No mRNA or rRNA depletion steps were performed. Libraries were sequenced by 50 bp paired-end sequencing.

2.5.4 Mapping

Reads were mapped to the Arabidopsis genome (TAIR10) using SOAPsplice 1.10 [150], allowing a maximum gap size within a two segment alignment of 10300 bp (corresponding to the longest intron in Arabidopsis), choosing as output format SAM, and otherwise using default conditions. These conditions correspond to a maximum of 3 mismatches and 2 indels allowed and non-unique reads being mapped only once. Separately mapped reads belonging to the same pair were joined after mapping and only kept if reads from the same pair had mapped to the same chromosome and were at most 3 kb apart.

2.5.5 RIP-seq transcript calling

We combined unique, non-genic (outside of TAIR10-annotated genes) sequencing reads from both biological repeats of RIP-seq. Regions with more than 8 reads positioned no further than 200 bp apart were identified as potential transcripts. We filtered the transcripts to keep only those with at least 1 read per 100 bp and with more than four fold enrichment (Col-0/*nrpe1*). The four fold enrichment test was then repeated with all sequencing reads, including not unique reads. Transcripts containing genes annotated in TAIR10 have been removed. We then performed the negative binomial test using NBPSeq R package[151]. Only transcripts with p<0.05 and FDR <0.05 were kept. We further filtered transcripts to

Transcript	Primer orientation	Sequence (5’-3’)
ACTIN	Forward	GAGAGATTCAGATGCCCAGAAGTC
	Reverse	TGGATTCCAGCAGCTTCCA
PolV_0290	Forward	AGCGGCCTAAATGAACATAATCCAGC
	Reverse	TCGTTGCTGGTTGTTCAAAACTGAC
PolV_0332	Forward	ATGTTTCATCTTGTTGTGGCCAAGG
	Reverse	GTTTCGACAAGGTCTTCCAAACTAAAG
PolV_0736, PolV_0737	Forward	ACCATATCCATTAATTTCGGGTTGG
	Reverse	AGTTCTGGGCACAAATATGGAACC
PolV_1057	Forward	AATTTGGTGTTGGTACATCTCAACTG
	Reverse	TTTTCACCTTCCCTTTCGAGGTGG
PolV_1468	Forward	AAAGCGATTTAGGCGGTCGACTAGG
	Reverse	ACAGTTGTCTATACGTCGCGTGAGC
PolV_1629	Forward	ATCATATCTTGCACCTCGGAAT
	Reverse	CGGGAATTTTTGCCACTAAA
PolV_1702, PolV_1703	Forward	TACCCTTGCCCTTTGTATCTTCTCC
	Reverse	GTGAGTGCCAATTTCTGCATCAAG
PolV_1818, PolV_1819	Forward	CGAAGGACGAAACTTTTTGG
	Reverse	GGTTTAAACGCAGCCAATGT
PolV_1873, PolV_1874	Forward	ATGGCCGAAATGTTGATAATGTGTAATC
	Reverse	CATGTTATGCTCAACCGGCGAC
PolV_1927, PolV_1926	Forward	GACCCATCTGCGATTCTGCGTTATG
	Reverse	GCGGATGACAGAGGGAGAACCAATC
PolV_2058	Forward	GGGCTTCCCTCTGAGTGTTT
	Reverse	CCGAAGCCCAACTAATATCG
PolV_2729	Forward	TGGCCCTTTCTCCTTCGACAACAAC
	Reverse	TTTTACATTGCAACGCACCCGTCC
PolV_2868	Forward	GACGGCACGGTTTCCTTGAATTCTC
	Reverse	GTCAAGTGGGAATGTGACACTGCGG
PolV_3151	Forward	CCTCACTCAAAGAAACGAGTTCCGAG
	Reverse	AGTGAAAGGGAGAGGAGTTGTTTGTG
PolV_3420	Forward	TTATTTTCAGGCCATAAAGAACCCAC
	Reverse	TTGTTGTAACTTGTAACTCGGACAAAG
PolV_3481	Forward	TTGTGGTCCAATTTGCTACG
	Reverse	GGCAGCAGGATATTCGGTTA
PolV_3863, PolV_3862	Forward	TCTTAATCGAACGCATGTGG
	Reverse	TGCAGCATCTGATCAACAAA
PolV_3926, PolV_3925	Forward	CGTGTCTGGTTGAGACCAAATTAGC
	Reverse	ATTAAACTCTGGAATCCGCGAGAAG
PolV_3958	Forward	TACCAACGCATCTCAAAATTGAACC
	Reverse	AAAATATTAAAGGGCGCGCTATTCG
PolV_4213	Forward	TTTGGAACAGACAATAAACCGACGC
	Reverse	AGTCTTCGACGGACTAACTACGGAC

Table 2.1: Oligonucleotides used in this study

keep only those with more than four fold enrichment (Col-0/*nrpe1*) in each repeat counted separately and with more than 2 reads in Col-0 in each of the biological repeats.

We called AGO4-bound transcripts as described for Pol V-transcripts, using the data obtained in the AGO4 RIP-seq for Col-0 and *ago4*, except that we filtered to 4 reads in Col-0, six-fold enrichment in Col-0/*ago4* and did not perform the negative binomial test (since we performed one biological repeat of AGO4 RIP-seq). Transcripts were considered Pol V-dependent if the enrichment in the AGO4-RIP (Col-0/*nrpe1*, [RPM]) was at least four-fold. AGO4-bound transcripts were considered to be Pol V-independent if the AGO4-RIP *nrpe1*/*ago4* signal was at least four-fold.

2.5.6 Heterochromatic regions

The genome was divided into 100 bp windows with 50 bp overlaps and H3K9me2 ChIP-seq[122] reads were counted. Those windows containing a greater number of reads than the median were kept and combined if sequentially located. Regions greater than 1 kb were then used in the analysis. Pol V and AGO4 ChIP-seq[56, 57], RIP-seq, siRNA[128], and CHH methylation[120] data were plotted as averages on aligned ends of heterochromatic regions +/- 300 bp.

For comparing AGO4-ChIP[57] and H3-ChIP[122] intensity(Figure 2.11F,Figure 2.12), heterochromatic regions were grouped into 100 or 4 groups according to the strength of the H3 signal (left end + 50 to 250 bp). The log2 value of the ratio of AGO4-ChIP signal inside (first 50 to 250 bp from the left end of the heterochromatic region) and outside (region 250 to 50 bp upstream of the left end of the heterochromatic region) was calculated and plotted against the median H3-ChIP signal. To visualize the ranking and the binding of AGO4 next to the H3K9me2 regions, the ranked heterochromatic regions were split into quartiles and the average H3 ChIP-seq signal[122] or AGO4 ChIP-seq Col-0/ago4 enrichment[57] was plotted as profiles at the 5'-end +/- 300 bp of the called heterochromatic regions.

2.5.7 Data visualization on heatmaps and profiles

Reads were counted using BEDTools 2.15.0 on Pol V-transcripts +/- 300 bp and RPM normalized. Transcripts were scaled to a uniform length and ratios of wild type and mutant signal was plotted on scaled individual transcripts (heatmap) or as an average (profile). To allow the visualization of individual transcripts, heatmaps show every other transcript sorted by size. For DNA methylation [120], subtraction of mutant methylation levels from wild-type were used instead of ratios and average methylation levels on differentially methylated regions (DMRs) were plotted.

Transposons annotated in TAIR10 were filtered for the presence of any overlaps with annotated Pol V transcripts. Alternatively, transposons larger than 4 kb from genomic regions with more genes than transposons were used. For strand preference analysis, ratios of the Pol V RIP-seq signal with the same or the opposite orientation than transposons were plotted on scaled transposon overlapping Pol V transcripts (+/- 300 bp) or, for transposons larger than 4 kb, at the 5'- and 3'-ends +/- 4 kb.

Significance of differences observed on profiles of average signal strengths was tested using the permutation test with 10,000 permutations. Averages from all nucleotides for specified regions were calculated for each transcript/TE without scaling to uniform transcript lengths.

2.5.8 Comparison of ends of Pol V transcripts

Differences between the positions of 5'- and 3'-ends of Pol V transcripts produced at the same locus but from opposite strands were plotted as boxplots. For comparing Pol V transcripts and AGO4-bound transcripts, the distances between ends were calculated for transcripts with the same orientation and a minimum overlap of 50%.

2.5.9 Comparison of Pol V RIP-seq and Pol V ChIP-seq

Overlapping Pol V-transcripts were combined to Pol V-transcribed regions and compared to a previously published list of Pol V ChIP-seq peaks[56] obtained using the same anti-NRPE1 antibody. Boxplots show enrichment on ChIP peaks or Pol V-transcripts or regions where both overlap. For randomization, 1000 permutations of the overlap of Pol V transcribed regions and Pol V ChIP peaks were performed.

2.5.10 Prediction of transcript size

Pol V RIP-seq paired end sequencing reads were mapped to all Arabidopsis transcripts (TAIR10) and Pol V transcripts. Mean lengths of sequenced fragments were calculated for transcripts with distinct origins and annotated lengths. Regression analysis using transcripts annotated in TAIR10 was applied to predict the length of Pol V transcripts based on the median length of sequenced fragments.

2.6 Previously Published Sequencing Datasets

Arabidopsis genome annotations (TAIR10) were obtained from TAIR. Pol V ChIP-seq data (SRA054962) and peak list, as well as, the AGO4 ChIP-seq data (GSE35381) were published previously [56, 57]. DNA methylation data (GSE39901) were used from Stroud et al.[120]. ChIP-seq data for histone modifications (GSE37644, GSE49090, and GSE28398) were published previously[124, 122, 123]. Pol V ChIP-seq dataset in the met1 mutant (GSE52041) was reported by [100]. siRNA (GSE36424) were reported previously [128]. Pol IV transcription data (SRP059814) were used from [81].

2.6.1 Data access

The sequencing data from this study have been submitted to the NCBI Gene Expression Omnibus under accession number GSE70290.

2.7 Author Contributions

2.7.1 Authors

Gudrun Böhmdorfer [G.B.]: Department of Molecular, Cellular, and Developmental Biology, University of Michigan, Ann Arbor, United States

Shriya Sethuraman [S.S.]: Bioinformatics Graduate Program, University of Michigan, Ann Arbor, United States

M Jordan Rowley [M.J.R.]: Department of Molecular, Cellular, and Developmental Biology, University of Michigan, Ann Arbor, United States

Michal Krzyszton [M.K.]: Faculty of Biology, Institute of Genetics and Biotechnology, University of Warsaw, Warsaw, Poland

M Hafiz Rothi [M.H.R.]: Department of Molecular, Cellular, and Developmental Biology, University of Michigan, Ann Arbor, United States

Lilia Bouzit [L.B.]: Department of Molecular, Cellular, and Developmental Biology, University of Michigan, Ann Arbor, United States

Andrzej T Wierzbicki [A.T.W.]: Department of Molecular, Cellular, and Developmental Biology, University of Michigan, Ann Arbor, United States

2.7.2 Individual author contributions

G.B., and A.T.W. designed research.

G.B., M.K., M.H.R., and L.B. performed research.

G.B., S.S., M.J.R., and A.T.W. analyzed data.

G.B., S.S., M.J.R., and A.T.W. wrote the paper.

My contribution (S.S.) in this project involved data analysis, which included Pol V transcribed region identification followed by step-wise data interpretation.

2.8 Publication

The work described in this chapter has been published in eLife journal and can be accessed at https://elifesciences.org/articles/19092 [152].

CHAPTER III

Broad Non-coding Transcription Suggests Genome Surveillance by RNA Polymerase V

Masayuki Tsuzuki[1], ***Shriya Sethuraman***[1], *Adriana N. Coke, M. Hafiz Rothi, Alan P. Boyle, and Andrzej T. Wierzbicki*

3.1 Abstract

Eukaryotic genomes are pervasively transcribed, yet most transcribed sequences lack conservation or known biological functions. In *Arabidopsis thaliana* RNA polymerase V (Pol V) produces non-coding transcripts, which base-pair with small interfering RNA (siRNA) and allow specific establishment of RNA-directed DNA methylation (RdDM) on transposable elements. Here, we show that Pol V transcribes much more broadly than previously expected, including subsets of both heterochromatic and euchromatic regions. At already established RdDM targets Pol V and siRNA work together to maintain silencing. In contrast, some euchromatic sequences do not give rise to siRNA but are covered by low levels of Pol V transcription, which is needed to establish RdDM *de novo* if a transposon is reactivated. We propose a model where Pol V surveils the genome to make it competent to silence newly activated or integrated transposons. This indicates that pervasive transcription of non-conserved sequences may serve an essential role in maintenance of genome integrity.

[1] co-first authors

3.2 Significance Statement

Eukaryotic genomes are pervasively transcribed, yet most transcribed sequences lack conservation or known biological functions. We show that a specialized plant-specific RNA Polymerase V broadly transcribes the Arabidopsis genome. We propose a model where Pol V transcription surveils the genome and is required to recognize and repress newly inserted or reactivated transposons. Our results indicate that pervasive transcription of non-conserved sequences may serve an essential role in maintenance of genome integrity.

3.3 Introduction

Eukaryotic genomes confront a variety of threats to their integrity. Transposable elements (TEs) are prevalent in most eukaryotic genomes and their activity is repressed by a variety of gene silencing mechanisms. One of those processes is RNA-mediated transcriptional gene silencing. In plants it is known as RNA-directed DNA methylation (RdDM) and represses TEs, repeats and other potentially harmful genetic elements by establishing repressive chromatin marks [153]. This process relies on small interfering RNAs (siRNAs), which in plants are most commonly produced from precursors generated by a specialized RNA polymerase, Pol IV. RdDM also requires the presence of lncRNA scaffolds, which are needed for the recognition of complementary target sequences by siRNA [55]. While fungi and possibly also animals use RNA Polymerase II (Pol II) to produce scaffold transcripts [153], plants use another specialized RNA Polymerase, Pol V [112, 88]. Several lines of evidence suggest that siRNA-AGO4 complexes recognize target loci by base-pairing with nascent scaffold transcripts [152, 154, 88], although siRNA-DNA base pairing is also possible [155]. Interaction of Pol V and its transcripts with siRNA-AGO4 leads to the recruitment of de novo DNA methyltransferase DRM2, which establishes DNA methylation [90, 91].

A key feature of RdDM is sequence-specificity, which assures efficient targeting of TEs and prevents inadvertent silencing of endogenous genes. This specificity is determined by the recruitment of both Pol IV and Pol V to TEs. However, unlike most DNA-dependent RNA polymerases, Pol IV and Pol V do not rely on sequence-encoded promoters [99]. Instead, they are recruited by pre-existing repressive chromatin modifications [100, 82, 101, 103]. This explains how silencing of already repressed TEs is specifically maintained by a positive feedback of RdDM.

An important unresolved question about RdDM is the mechanism responsible for the initial silencing of newly inserted or reactivated TEs. This process requires the activity of Pol V [108, 142, 109, 110, 156](Gallego-Bartolomé et al., 2019); however, it is unknown how Pol V is specifically recruited to TEs in the absence of pre-existing DNA methylation. Another open question about RdDM is the functional relationship between Pol II, Pol IV and Pol V, which are all involved in this process. This especially applies to Pol IV and Pol V, which are recruited by H3K9me2 and DNA methylation, respectively [157, 82, 101, 103]. Because these repressive chromatin modifications are closely functionally related, Pol IV and Pol V are expected to be recruited to the same loci. This negates the need for two specialized polymerases and raises the question why one polymerase cannot produce both siRNA precursors and scaffold transcripts, as is the case in fission yeast [153].

Here, we show that Pol V transcription is not limited to transposable elements and other known RdDM targets. Instead, Pol V transcribes broadly and pervasively. Specificity of silencing is not restricted by the recruitment of Pol V; instead, siRNA production is likely its primary determinant. Pol V is needed to facilitate silencing of reactivated transposable elements, presumably by making them competent to receive the silencing signal. This explains how RdDM may be established de novo and explains the functional relationship between Pol II, Pol IV and Pol V. This also demonstrates that pervasive transcription of non-conserved genomic regions serves an important role in maintaining genome integrity.

3.4 Results

3.4.1 Pol V transcribes broadly

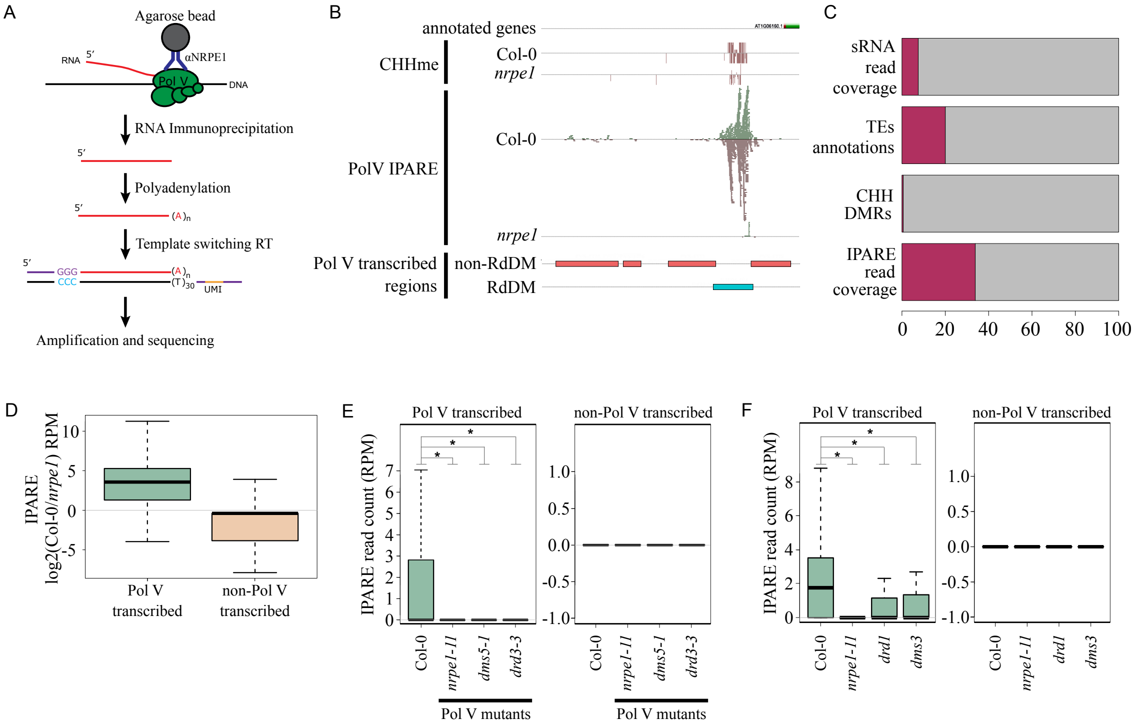

Figure 3.1: **Pol V transcribes more broadly than expected: (A)** IPARE method of detecting Pol V transcription. **(B)** Genome browser screenshot of a region transcribed by Pol V. TAIR10 genome annotation, DNA methylation in CHH contexts and IPARE data are shown. **(C)** IPARE reads from Col-0 wild type cover a greater proportion of the genome than known features of RdDM including siRNA [102], annotated transposable elements [110] and CHH DMRs. **(D)** HMM identifies Pol V-transcribed regions of the genome. Boxplot shows IPARE signal using combined data from three biological replicates comparing bins identified as Pol V transcribed (States 0 and 1) or non-Pol V transcribed (States 2 and 3). **(E)** Pol V IPARE signal depends on the enzymatic activity of Pol V. Boxplots show RPM-normalized IPARE signal levels at Pol V transcribed and non-Pol V transcribed regions in Col-0, *nrpe1* (null allele), *dms5-1* (early termination allele of *NRPE1*) and *drd3-3* (catalytic active site point mutant of *NRPE1*). Stars indicate Wilcoxon test p < 2.2e-16. **(F)** Pol V IPARE signal depends on the activity of the DDR complex. Boxplots show RPM-normalized IPARE signal levels at Pol V transcribed and non-Pol V transcribed regions in Col-0 and *nrpe1* as well as mutants in DDR subunits *drd1* and *dms3*. Stars indicate Wilcoxon test p < 2.2e-16.

The current model of RdDM predicts that Pol V transcribes only *bona fide* RdDM targets. To test this prediction, we developed a high sensitivity method to detect Pol V transcription: Immunoprecipitation followed by Analysis of RNA Ends (IPARE) (Figure 3.1A).

It combines immunoprecipitation of the Pol V complex using an antibody specific towards NRPE1, the largest subunit of Pol V[152], with a modified nanoPARE RNA-seq protocol[158]. This method achieved a greatly improved sensitivity in detecting Pol V transcripts (Figure 3.2A) but not mRNA (Figure 3.2B) compared to previously used RNA Immunoprecipitation [152] or GRO-seq [154].

Analysis of the IPARE results confirmed the presence of the anticipated Pol V transcription signal on known RdDM targets (Figure 3.2A)[152, 154, 112], which were previously shown to have high levels of CHH methylation and 24nt siRNA[152]. One locus is shown on a genome browser screenshot in Figure 3.1B and Figure 3.2C. Surprisingly, IPARE sequencing reads from Col-0 wild type covered a relatively large proportion of the genome, which could be observed in three biological replicates of IPARE (Figure 3.2D) and in a merged dataset, where as much as 31.2% of the genome is covered by sequencing reads in Col-0 wild type (Figure 3.2D with thinning and Figure 3.1C without thinning). This is substantially more than expected based on the extent of siRNA accumulation, CHH methylation or transposable element (TE) annotations (Figure 3.1C). This indicates that Pol V transcribes more broadly than the estimates of RdDM prevalence.

Detection of Pol V transcription by IPARE relies on the availability of a negative control, the null *nrpe1-11* mutant, which does not contain the epitope for IP and has no strong developmental or physiological phenotypes [152, 78, 79, 88]. The *nrpe1* mutant had 8.1% of the genome covered by IPARE sequencing reads, compared to 22.8% in the Col-0 wild type dataset thinned to the same coverage as *nrpe1* (Figure 3.2D). To eliminate background signal originating from Pol I, II and III we filtered all sequencing reads based on known properties of different RNA polymerases[152, 154] (Figure 3.2E). This eliminated the vast majority of reads present in *nrpe1* (1.9% genome coverage remaining) but only a small subset of reads from Col-0 wild type (23.3% genome coverage remaining without thinning) (Figure 3.2F). This confirms that the IPARE signal originates from *bona fide* Pol V transcripts and Pol V transcribes more broadly than the estimates of RdDM prevalence.

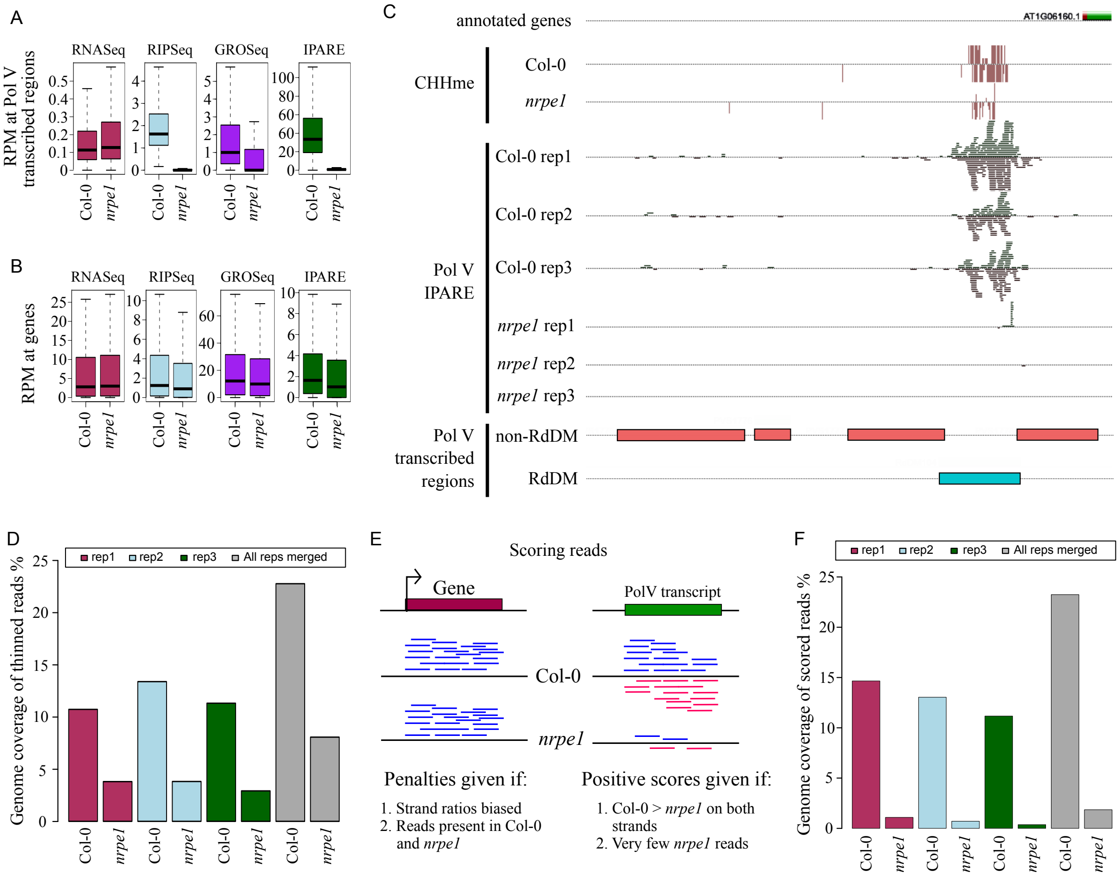

Figure 3.2: **Pol V transcribes more broadly than expected:** **(A)** High sensitivity of Pol V IPARE compared to previously published datasets. Boxplot shows RPM-normalized signal levels of RNA-Seq [42], Pol V RIP-Seq [152], GRO-Seq [154] and Pol V IPARE (this study) counted on Pol V transcripts identified previously [152].Differences in signal levels are indicated by different y-axis scales. Wilcoxon test p < 2.2* 10-16 for Pol V RIP, GRO-seq and IPARE. **(B)** IPARE signal is not enriched at genes. Boxplot shows RPM-normalized signal levels of RNA-Seq [42], Pol V RIP-Seq [152], GRO-Seq [154] and Pol V IPARE (this study) counted on genes annotated in TAIR10. **(C)** Genome browser screenshot of the same region transcribed by Pol V that is shown in Fig 3.1B. TAIR10 genome annotation, DNA methylation in CHH contexts, three independent biological replicates of Pol V IPARE and annotated Pol V transcribed regions data are shown. **(D)** Pol V IPARE reads cover a substantial proportion of the Arabidopsis genome. Bar plot shows percentages of the genome covered by reads from three individual biological replicates and combined reads from all replicates of Col-0 and *nrpe1*. The numbers of reads in both genotypes were randomly thinned to obtain equal sequencing depths. **(E)** Diagram of scoring of Pol V IPARE reads to eliminate background originating from Pol I, II and III. **(F)** Filtered Pol V IPARE reads cover a substantial proportion of the Arabidopsis genome. Bar plot shows percentages of the genome covered by reads from three individual biological replicates and combined reads from all replicates of Col-0 and *nrpe1*.

To obtain a reliable and unbiased way to identify Pol V transcribed genomic regions we split the genome into 200 bp long bins and used a Hidden Markov Model (HMM) to identify bins with evidence of Pol V transcription (Figure 3.3A). Using raw RPM normalized Pol V IPARE sequencing data combined from three biological replicates, this approach split the genome into Pol V-transcribed (42.4%) and non-Pol V-transcribed (57.6%) bins (Figure 3.1D). IPARE signal levels were not caused by stochastic mapping of sequencing reads (Figure 3.3BC) and were significantly correlated between three independent biological replicates (Figure 3.3D). Non-Pol V transcribed bins include loci with no detectable transcription, transcription by other RNA polymerases and a smaller subset of *bona fide* RdDM targets where the loss of RdDM in *nrpe1* leads to increased transcription by other RNA polymerases. Together, these results further confirm that Pol V transcribes more broadly than previously expected.

To determine if IPARE specifically detects Pol V transcription, we performed this assay using the *drd3-3* mutant[78], which is an allele of NRPE1 with a point mutation in the catalytic active site[113]. This mutant is expected to contain the epitope for IP but no Pol V transcripts[112]. The IPARE signal in *drd3-3* was significantly lower than in Col-0 wild type (Figure 3.1E), which confirms that IPARE specifically detects Pol V transcripts. We obtained a similar result with *dms5-1*, another allele of NRPE1, which contains a premature stop codon[106] (Figure 3.1E). To further test the specificity of IPARE we used *drd1* and *dms3* mutants, which lack subunits of the DDR complex and are expected to disrupt Pol V transcription without affecting the accumulation of Pol V[58, 112, 88, 86]. IPARE signal was slightly higher in *drd1* and *dms3* than in *nrpe1* (Figure 3.1F), which is consistent with the epitope (NRPE1) being present in those mutants but could also be explained by the presence of another mechanism recruiting Pol V. It was, however, significantly lower than in Col-0 wild type (Figure 3.1F), which further confirms that IPARE specifically captures Pol V transcripts. Although pervasive transcription is inherently difficult to demonstrate, these results indicate that alternative explanations of the broad IPARE signal are unlikely and confirm that Pol V transcribes more broadly than previously expected.

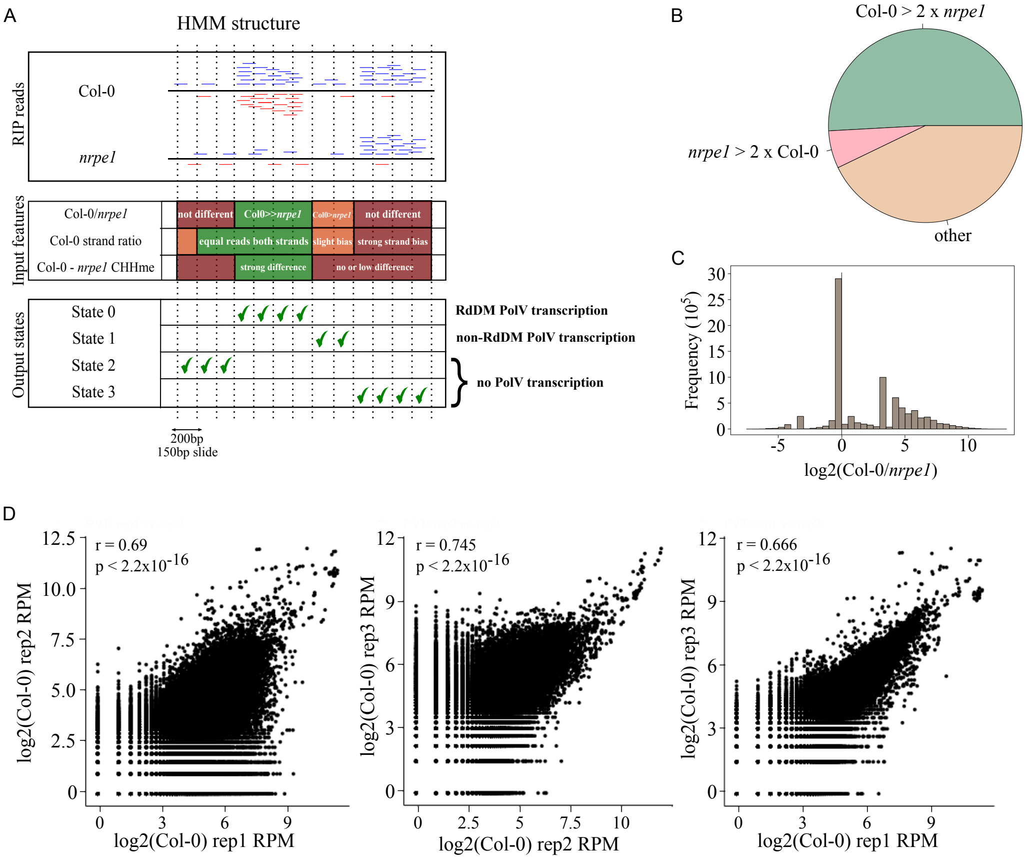

Figure 3.3: **Identification of Pol V transcribed regions using HMM: (A)** Diagram of the HMM workflow depicting the inputs and predicted output states.**(B)** Pol V IPARE signal is not the product of noise. The chart shows genome coverage percentages of 200bp bins with 2 times more reads in Col-0 over *nrpe1*, 2 times more reads in *nrpe1* over Col-0 and other bins.**(C)** Genome-wide enrichment of IPARE reads in Col-0 wild type vs. *nrpe1*. Histogram shows the distribution of the log2(Col-0/*nrpe1*) RPM values in 200bp genomic bins throughout the entire genome.**(D)** Three biological replicates of Pol V IPARE are highly correlated. Scatterplots show comparisons of Pol V IPARE signals on Pol V-transcribed regions between individual biological replicates.

3.4.2 Pol V is not the primary determinant of silencing specificity

Broad presence of Pol V throughout the genome suggests that it may not be the primary determinant of the specificity of RdDM. To test this prediction, we designed HMM-based identification of Pol V-transcribed regions in a way that allows distinguishing RdDM targets from sequences not targeted by RdDM. This determination was possible based on whole genome bisulfite sequencing data of CHH methylation. Among four identified HMM states, two (State 0 and State 1) included Pol V-transcribed regions (Figure 3.1D, Figure 3.4A,

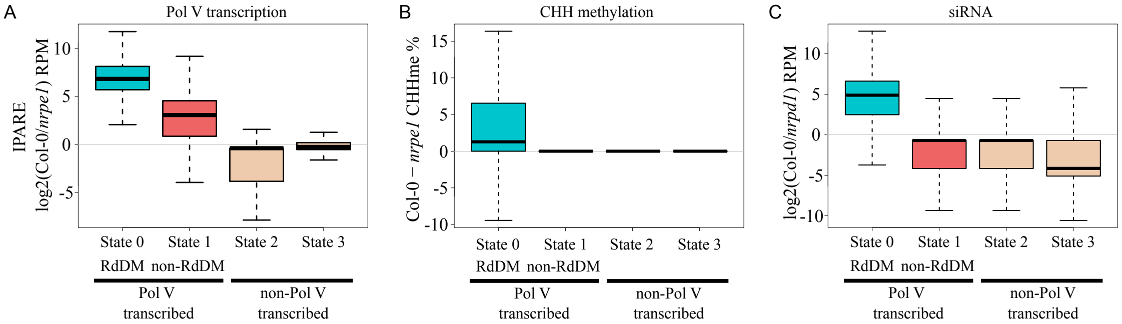

Figure 3.4: **Pol V transcription does not determine RdDM specificity:** **(A)** HMM analysis of Pol V IPARE identified two distinct states of Pol V transcription. Boxplot shows Pol V IPARE signal from all three biological replicates combined at each of the 4 emission states of the HMM output. States 0 and 1 are transcribed by Pol V. States 2 and 3 show no evidence of Pol V transcription. **(B)** DNA methylation in the CHH context is present only on a subset of Pol V-transcribed regions (State 0 but not on State 1). Boxplot shows the Pol V-dependent DNA methylation in the CHH context at each of the 4 emission states of HMM output. **(C)** RdDM is determined by siRNA. siRNA is enriched only in 1 of the 4 emission states of HMM. Boxplot shows Pol IV-dependent siRNA signal [102] at each of the 4 emission states of HMM output.

Figure 3.5A). Regions identified as State 0 were enriched in Pol V-dependent CHH methylation (Figure 3.4B), which indicates that they are *bona fide* RdDM targets. State 1 on the other hand, although Pol V-transcribed (Figure 3.4A) and abundant throughout the genome (Figure 3.5B), had no enrichment in Pol V-dependent CHH methylation (Figure 3.4B). This indicates that there is evidence of extensive Pol V transcription outside of RdDM targets.

Although non-RdDM Pol V transcription (State 1) is clearly detectable, it accumulates at substantially lower levels than RdDM Pol V transcription (State 0; Figure 3.4A). To confirm that non-RdDM Pol V transcription is not an artifact of IPARE, we performed real time RT-PCR using total RNA on arbitrarily selected non-RdDM Pol V transcribed regions. We found 10 primer pairs that showed a substantial signal reduction in *nrpe1* (Figure 3.5C), which is consistent with the presence of Pol V-dependent transcription. We further used IPARE to analyze the *drd3-3* and *dms5-1* alleles of NRPE1, which had significant reductions of signal on both RdDM (State 0) and non-RdDM (State 1) Pol V-transcribed regions (Figure 3.5D). This confirms that Pol V transcribes both RdDM and non-RdDM genomic regions and therefore, presence or absence of Pol V may not be essential for the determi-

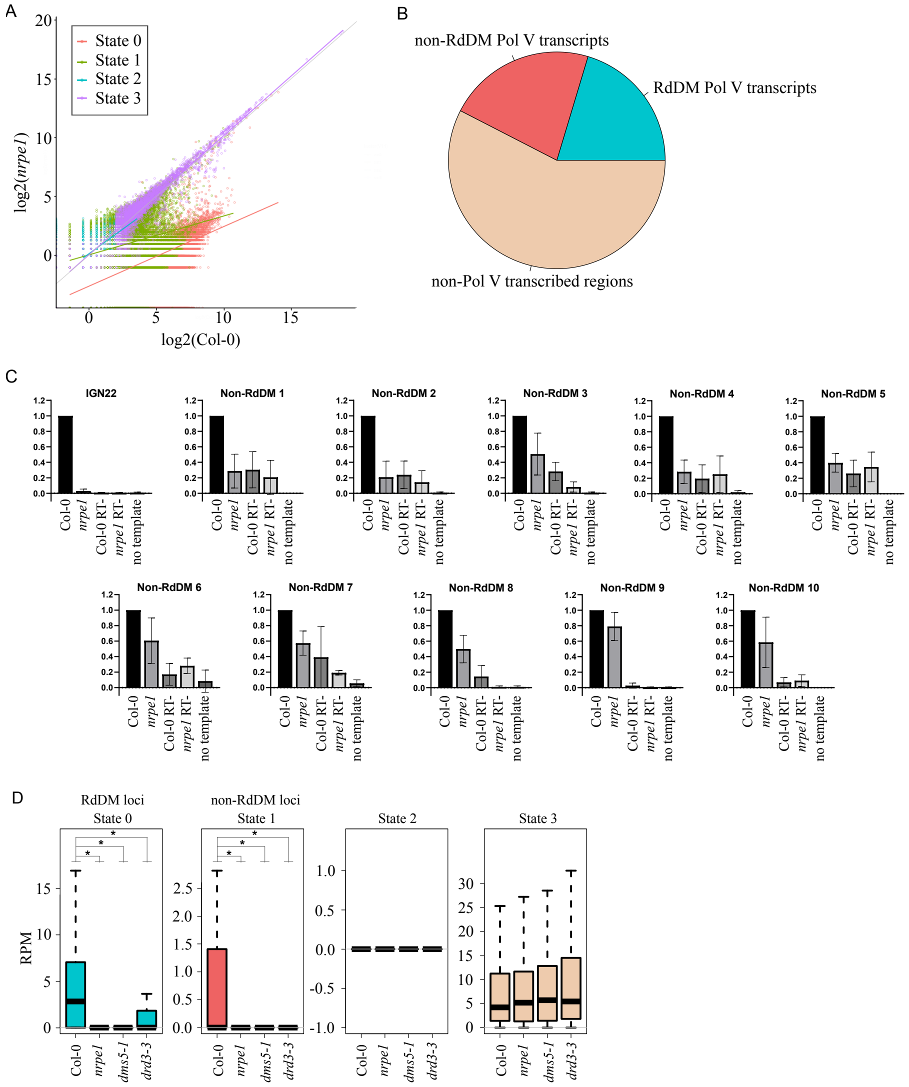

Figure 3.5: **Comparison of RdDM and non-RdDM Pol V transcription.:** **(A)** Pol V IPARE signal in Col-0 wild type compared to *nrpe1* at individual emission states identified by HMM. Scatterplot shows RPM-normalized signal levels in Col-0 and *nrpe1*. States 0 and 1 have more reads in Col-0 than *nrpe1* and represent Pol V transcribed loci. State 2 and 3 have comparable numbers of reads in Col-0 and *nrpe1* and are not Pol V transcribed.**(B)** Percentages of the *Arabidopsis* genome covered by Pol V transcription. The chart shows values for HMM state 0 (non-RdDM Pol V transcripts), state 1 (RdDM Pol V transcripts) and states 2 and 3 (others). Genomic bins were filtered to remove overlaps.**(C)** Locus-specific validation of non-RdDM Pol V transcription by real time RT-PCR using total RNA. IGN22 is a previously published RdDM Pol V transcribed locus. Signal is calculated relative to Col-0 and corrected using ACTIN2. Plots show averages and standard deviations from three or four biological replicates.**(D)** Non-RdDM Pol V transcription depends on the catalytic activity of Pol V. Boxplots show RPM-normalized IPARE signal levels at all four HMM states in Col-0, *nrpe1* (null allele), *dms5-1* (early termination allele of *NRPE1*) and *drd3-3* (catalytic active site point mutant of *NRPE1*). Stars indicate Wilcoxon test p < 2.2e-16.

nation of the specificity of RdDM.

We further analyzed regions of non-RdDM Pol V transcription to determine where in the genome Pol V transcribes independently of RdDM. We first performed HMM-based identification of Pol V transcription independently in each individual biological replicate of IPARE and determined the percentage of genomic bins identified in at least two independent replicates. While 77% of RdDM Pol V transcripts were identified more than once, 50% of non-RdDM Pol V transcripts were identified more than once. This is consistent with detection of non-RdDM Pol V transcription being limited by sequencing depth and this category of transcripts being possibly more widespread than detected at the sequencing coverage we used. Further analysis of non-RdDM Pol V transcription indicated that it is not associated with proximity to RdDM loci (Figure 3.6A), it is mostly euchromatic (Figure 3.6BC) and enriched on intergenic regions (Figure 3.6D). While RdDM regions had the expected high levels of CG methylation, a minor subset of non-RdDM Pol V transcribed regions also had elevated levels of CG methylation (Figure 3.6E), which may be explained by RdDM-independent silencing or gene body methylation. Together, these results support the possibility that non-RdDM Pol V transcription is produced stochastically and mostly non-specifically over a substantial fraction of the genome.

If Pol V has a limited role in determining the specificity of RdDM, Pol IV-dependent production of siRNA remains the expected alternative determinant of specificity[55]. To test this possibility, we quantified previously published Pol IV-dependent 24nt siRNA[102] on regions corresponding to four states identified by HMM. State 0, which corresponds to RdDM Pol V transcription, was enriched in Pol IV-dependent 24nt siRNA (Figure 3.4C). However, state 1, which corresponds to non-RdDM Pol V transcription was not enriched in siRNA (Figure 3.4C). We further tested the enrichment of small RNA clusters detected in Col-0 wild type on four states and found that small RNAs were enriched on RdDM Pol V transcripts but not non-RdDM Pol V transcripts or non-Pol V transcribed regions (Figure 3.6F). These results indicate that the presence of siRNA is associated with RdDM,

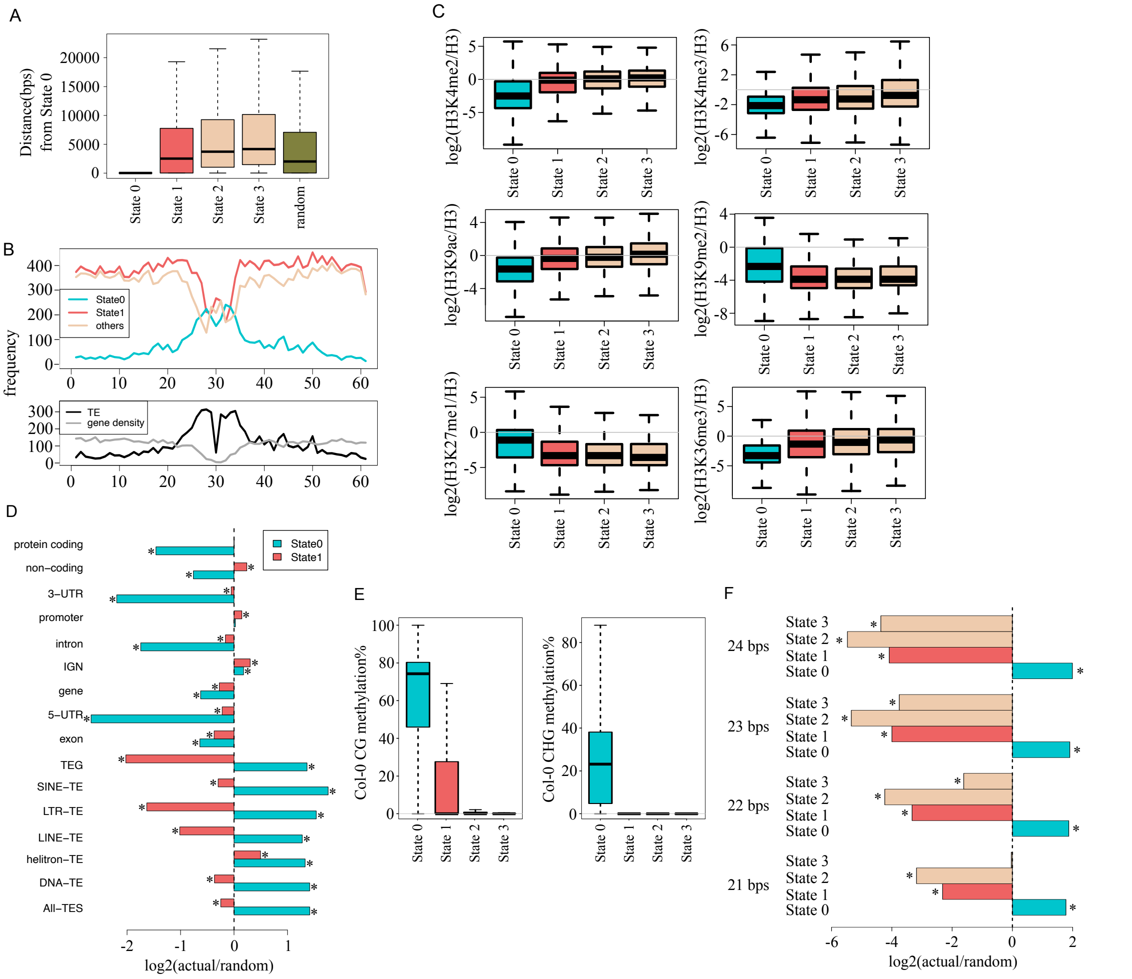

Figure 3.6: **Properties of non RdDM Pol V transcription:** **(A)** Non-RdDM Pol V transcription is not limited to the proximity of RdDM. Boxplot shows distances between regions assigned to each state and the closest State 0 region.**(B)** Non-RdDM Pol V transcription is enriched on chromosome arms. Plots show frequency of State 0 (RdDM Pol V transcription), State 1 (Non-RdDM Pol V transcription) and other regions in 500 kb genomic bins. Frequency of annotated TEs and genes (TAIR10) are shown as a reference. **(C)** Non-RdDM Pol V transcription is not associated with a strong enrichment in repressive chromatin marks. Boxplots show the levels of H3K4me2, H3K4me3, H3K9ac, H3K9me2, H3K27me1 and H3K36me3 corrected for nucleosome density. **(D)** Non-RdDM Pol V transcription is enriched on intergenic and non-coding regions. The plot shows enrichment of RdDM Pol V transcription (State 0) and non-RdDM Pol V transcription (State 1) on various genomic features. Ratio between observed overlaps and expected overlaps calculated as a mean of 1000 permutations of random genomic regions. Stars indicate $p < 0.001$. **(E)** DNA methylation in the CG and CHG contexts is enriched on RdDM Pol V-transcribed regions (State 0). Boxplot shows DNA methylation levels in Col-0 wild type at each of the 4 emission states of HMM output. **(F)** Small RNA clusters are enriched in RdDM Pol V transcription but not in non-RdDM Pol V transcription. Plots show ratio between observed overlaps and expected overlaps calculated as a mean of 1000 permutations of random genomic regions. Stars indicate $p < 0.001$.

which is consistent with siRNA being the primary determinant of RdDM specificity. We conclude that Pol V is unlikely to be the primary determinant of sequence specificity of RdDM.

3.4.3 Pol V is needed for TE resilencing

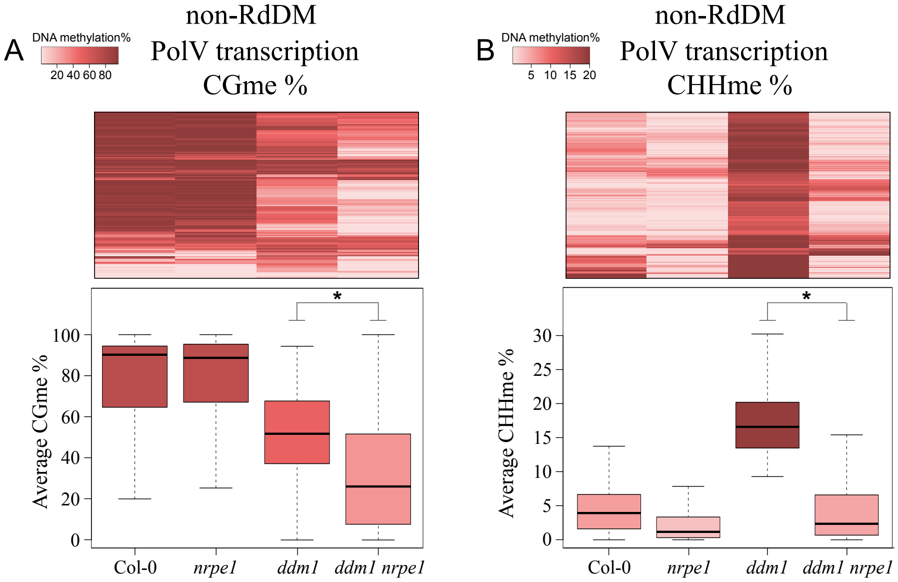

Figure 3.7: **Pol V transcription is required for TE resilencing:** **(A)** Loss of DNA methylation in the CG context in *ddm1*. Heatmap shows average CG methylation levels on non-RdDM Pol V-transcribed loci which gain CHH methylation in *ddm1*. Boxplot shows the distribution of datapoints shown in the heatmap. Stars indicate Wilcoxon test p < 2.2e-16. **(B)** *De novo* CHH methylation of TEs reactivated in *ddm1* requires Pol V. CHHme established in *ddm1* is dependent on Pol V. Heatmap shows average CHH methylation levels on non-RdDM Pol V-transcribed loci which gain CHH methylation in *ddm1*. Boxplot shows the distribution of datapoints shown in the heatmap. Stars indicate Wilcoxon test p < 2.2e-16.

Our observations that Pol V transcribes broadly and is not a determinant of RdDM explains a key inconsistency in the mechanistic understanding of this process. Although *de novo* silencing of newly integrated or activated TEs seems to always require Pol V [110], no mechanisms recruiting Pol V to previously unsilenced TEs are known. Our data indicate that non-RdDM Pol V transcription occurs broadly enough to facilitate *de novo* silencing of newly integrated or activated TEs. To test this possibility, we took advantage of the *ddm1* mutant, which disrupts the maintenance of CG methylation [159]. Because in *Ara-*

bidopsis a subset of TEs is silenced by CG methylation in an RNA-independent manner, disruption of CG methylation leads to reactivation of those TEs and establishment of *de novo* RdDM, manifested as CHH methylation [110]. We analyzed previously published methylome datasets from Col-0 wild type, *nrpe1*, *ddm1* and *ddm1 nrpe1* double mutant [110]. We identified differentially methylated regions (DMRs) between Col-0 wild type and *ddm1*, where CHH methylation is increased in *ddm1*. We then selected DMRs that overlap non-RdDM Pol V transcription identified by IPARE and HMM (State 1). As expected, these DMRs have reduced levels of CG methylation in *ddm1* (Figure 3.7A). We then tested if the increased CHH methylation requires Pol V transcription by analyzing the *ddm1 nrpe1* double mutant. Levels of CHH methylation were significantly lower in *ddm1 nrpe1* compared to *ddm1* (Figure 3.7B). This suggests that non-RdDM Pol V transcription is required for *de novo* establishment of RdDM in the *ddm1* mutant. This is consistent with Pol V transcribing the genome to make it competent for silencing if a transposon becomes reactivated. A similar mechanism may occur when new transposons are integrated into the genome.

3.4.4 Non-RdDM Pol V transcription requires the DDR complex

Non-RdDM Pol V transcription results in lower IPARE signals than RdDM transcription and therefore may be controlled in a unique manner. RdDM Pol V transcription requires the DDR complex, which is involved in transcription initiation and/or elongation [58, 112, 88, 86]. To test if non-RdDM Pol V transcription also depends on the DDR complex, we analyzed the RdDM and non-RdDM Pol V transcription IPARE signal in DDR mutants, *drd1* and *dms3*. Both mutants showed strong reductions of Pol V transcription on both RdDM and non-RdDM sites (Figure 3.8A). This indicates that Pol V requires the DDR complex even on non-RdDM sites.

One known mechanism of Pol V recruitment involves SUVH2 and SUVH9 proteins, which recognize preexisting DNA methylation[100, 101]. To test if this mechanism is

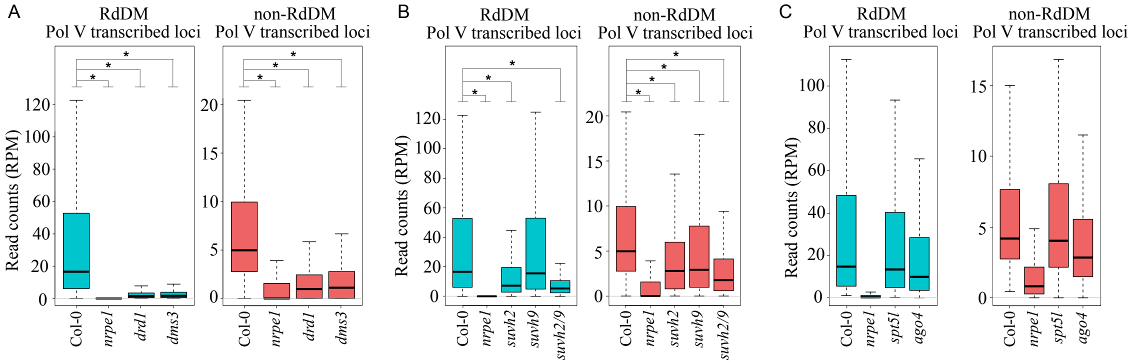

Figure 3.8: **Non-RdDM Pol V transcription requires the DDR complex:** **(A)** RdDM and non-RdDM Pol V transcription requires the DDR complex. Boxplots show Pol V IPARE signal levels in Col-0, *nrpe1*, *drd1* and *dms3* at RdDM and non-RdDM Pol V transcribed loci. Stars indicate Wilcoxon test p < 2.2e-16. **(B)** Proteins that work downstream of Pol V do not affect Pol V transcription. Boxplots show Pol V IPARE signal levels in Col-0, *nrpe1*, *spt5l*, *ago4* and *drm2* at RdDM and non-RdDM Pol V transcribed loci. **(C)** DRM2-dependent DNA methylation in CHH context does not affect the level of Pol V transcription. Scatterplots show effects of *drm2* on Pol V IPARE signal levels compared to the effects of *drm2* on CHH methylation at non-RdDM Pol V transcribed, RdDM Pol V transcribed and non-Pol V transcribed loci.

involved in both RdDM and non-RdDM Pol V transcription, we performed IPARE in *suvh2*, *suvh9* and *suvh2/suvh9* mutants. RdDM Pol V transcription was reduced in *suvh2* and *suvh2/suvh9* mutants (Figure 3.8B), which is consistent with previously published data[100, 101]. Non-RdDM Pol V transcription was slightly reduced in both *suvh2* and *suvh9* mutants and more substantially reduced in the *suvh2/suvh9* double mutant (Figure 3.8B). This suggests that SUVH2 and SUVH9 might play a role in non-RdDM Pol V transcription. Signal observed in the *suvh2/suvh9* double mutant was still substantially stronger than in *nrpe1* (Figure 3.8B), which indicates that other factors may also contribute to the initiation of both RdDM and non-RdDM Pol V transcription. Proteins that work downstream of Pol V have been shown to affect processing of Pol V transcripts through slicing by AGO4, which requires SPT5L[154]. To determine if non-RdDM Pol V transcription is affected by these downstream factors, we performed IPARE in *spt5l* and *ago4* mutants. Both mutants had no major effects on the median levels of accumulation of both RdDM and non-RdDM Pol V transcripts detected by IPARE (Figure 3.8C). This indicates that downstream factors do not affect the accumulation of nascent Pol V transcripts detected in our assay.

3.5 Discussion

We propose a speculative model where Pol V stochastically transcribes a significant fraction of the genome to make it competent for silencing (Figure 3.9). This includes surveillance of euchromatic sequences, which may harbor inactive transposons or may become landing sites for random integration of new transposons (Figure 3.9A). If there is no complementary siRNA, chromatin modifiers are not recruited, and Pol V transcripts are expected to quickly degrade with no consequences. However, any newly integrated or reactivated transposon triggers one of several pathways to produce siRNA [160]. Newly synthesized siRNA base-pairs with already available Pol V transcripts to establish initial DNA methyl marks (Figure 3.9B). This leads to the recruitment of Pol IV and further siRNA production. At the same time, Pol V transitions from a low-level surveillance status to a higher rate of transcription associated with maintenance of RdDM (Figure 3.9C).

Surveillance Pol V transcription occurs much more broadly than siRNA production and RdDM, including euchromatic loci and possibly also heterochromatic loci repressed by pathways other than RdDM. This indicates that previous studies of Pol V localization by ChIP-seq [56] were not sensitive enough to detect the actual breadth of Pol V transcription. Although we detected Pol V transcription on 42.4% of the genome, the absence of Pol V on any of the remaining 57.6% cannot be conclusively proven, especially on loci transcribed by Pol I, II or III. It is therefore possible that Pol V transcribes even more broadly and in the extreme case the entire genome could possibly be subject to at least occasional Pol V transcription. This would be consistent with the fact that Pol V appears to be universally required for de novo RdDM[108, 142, 109, 110, 156, 161]. Initiation of surveillance Pol V transcription is likely to be stochastic, which is consistent with lack of sequence-specificity detected for Pol V[154, 152, 86, 56]. The DDR complex and SUVH2/9 are likely responsible for sequence-independent initiation and/or

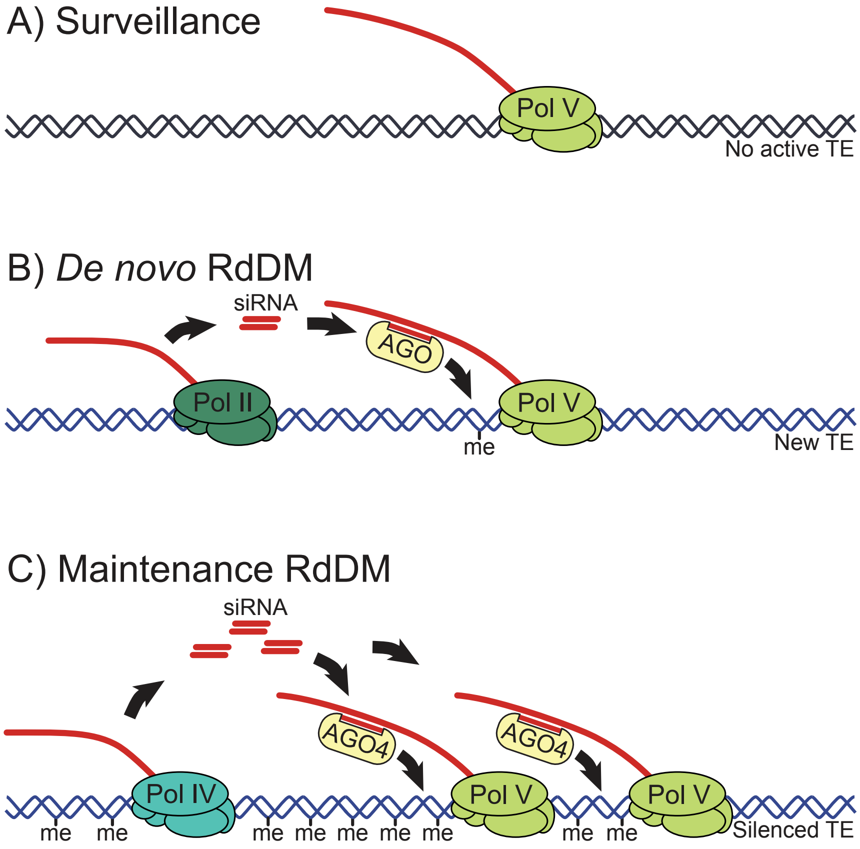

Figure 3.9: **Speculative model explaining the role of Pol V in *de novo* and maintenance RdDM:** **(A)** A large fraction of the genome is subject to infrequent surveillance transcription by Pol V. This includes euchromatic loci with no active TEs. The role of this transcription is to make the genome competent to initiate silencing if siRNAs become available. **(B)** Insertion and/or activation of a TE leads to siRNA production. This siRNA may initiate silencing by base-pairing with already available surveillance Pol V transcripts. This leads to the establishment of first repressive chromatin marks. **(C)** Presence of repressive chromatin marks leads to recruitment of Pol IV and enhanced production of siRNAs. At the same time, Pol V transitions to a higher rate of transcription, which facilitates efficient maintenance of RdDM.

elongation of Pol V transcription[112, 88, 86]. In contrast to loci where RdDM is already established[100], surveillance Pol V transcription is expected to be independent of pre-existing chromatin modifications, which indicates that SUVH2 and SUVH9 proteins may have a broader role than binding methylated DNA[100].

In our model, surveillance Pol V transcription is expected to have no independent impact on RdDM. However, Pol V has been proposed to have roles independent of 24nt siRNA[145] or gene silencing[162], both of which are consistent with our model. The role of genome surveillance by Pol V is tied to the inability of siRNA-AGO4 complexes to recognize complementary target loci in the absence of Pol V[88, 155]. Widespread Pol V transcription lets siRNA-AGO4 recognize target loci even if they were not previously silenced. This indicates that the specificity of base-pairing between siRNA and a highly complex pool of Pol V transcripts is essential for precise establishment of RdDM, which is consistent with high accuracy of ribonucleotide incorporation by Pol V [163]. Frequency of surveillance Pol V transcription remains unknown but low levels of those transcripts suggests that it is not a frequent event, which is consistent with a relatively low rate of ribonucleotide incorporation by Pol V[163]. Binding of siRNA-AGO4 to euchromatic surveillance Pol V transcripts leads to the recruitment of chromatin modifying machinery[90, 91] and the establishment of initial repressive chromatin marks. This leads to a series of events that result in robust and stable RdDM. First of those events is the repression of Pol II transcription and activation of Pol IV. This stops the production of initiating siRNAs[160], which are replaced by a strong accumulation of 24nt siRNA produced by Pol IV, RDR2 and DCL3. This is consistent with Pol IV being recruited by H3K9me2 recognized by SHH1[82, 103]. The second event is a strong increase in the level of Pol V transcription, which allows robust reestablishment of repressive chromatin marks and efficient maintenance of RdDM. Because assays used in previous studies were not sensitive enough to detect surveillance Pol V transcription and only reported more abundant Pol V transcription on already silenced loci, this is consistent with the reported importance of pre-existing DNA methylation for Pol V recruitment[100, 101]. The mechanism responsible for transition of Pol V from low level surveillance transcription to a higher rate of transcription may involve DNA methylation; however, the presence of CG methylation[152] and partial involvement of SUVH2 and SUVH9 proteins (Figure 3.8) do not fully explain this transition. This indicates that other properties of chromatin are likely to be important for transition of Pol V into high transcription rate. The surveillance model of Pol V transcription predicts that siRNA incorporated into a proper AGO protein should be sufficient to initiate RdDM.

This suggests the presence of a threshold mechanism, which prevents silencing by stochastically produced siRNAs. Existence of a threshold could explain why artificial tethering of Pol V to a reactivated FWA locus results in reestablishment of silencing[100, 164]. Active FWA in fwa-4 epiallele accumulates a low level of siRNAs of atypical lengths, which are insufficient to initiate RdDM [164]. We propose that the enhancement of Pol V transcription by tethering a DDR subunit to FWA lowers the threshold and allows re-initiation of RdDM. Transition of polymerase activities during the initial establishment of RdDM explains the functional relationship between three RNA polymerases involved in this process. Aberrant Pol II transcription is the initial signal that targets newly inserted or activated TEs for silencing and is replaced by Pol IV, as previously demonstrated[160]. A role of Pol II at this step of silencing is further supported by[129]. Pol V is functionally distinct in that it provides less or possibly even no sequence specificity for de novo RdDM. However, after silencing has been established a higher level of Pol V transcription facilitates efficient maintenance of repressive chromatin states. Broad transcription by Pol V changes our understanding of pervasive transcription, which in the absence of sequence conservation has been interpreted as non-functional[165]. It provides evidence that non-coding transcription of non-conserved sequences may serve an important role in maintaining genome integrity. Given the conservation of transcriptional silencing mechanisms[166], a similar process may exist outside of the plant kingdom.

3.6 Materials and Methods

3.6.1 Plant materials and growth condition

Arabidopsis thaliana Columbia-0 (Col-0) was used as a wild-type in all analyses. All mutant plants used in this study were also in the Col-0 background. We used the following alleles previously described: *nrpe1* (*nrpe1-11*: SALK_029919 [145]; *dms5-1* [167], *drd3-3* [78]), *drd1* (*drd1-6* from Marjori Matzke), *dms3* (SALK_125019C), *ago4* [88], *spt5l*

(SALK_001254), *drm2* (SAIL_70E12) [88, 112, 157]. The plants were grown at 22C under white fluorescent light in 16h/8h day/dark cycle.

3.6.2 Immunoprecipitation followed by analysis of RNA ends (IPARE)

The IPARE protocol was based on RNA-IP (RIP)-seq [168] and a modified NanoPARE protocol [169]. Chromatin extraction and RNA-IP were performed as described in [149]. At the final step of RNA-IP, extracted RNA was diluted in 5 µl of RNase-free H2O. RNA was then denatured by incubation at 65C for 5 min and put on ice immediately. Denatured RNA was polyadenylated with Poly(A) polymerase (NEB M0276S) using the following condition: 5 µl RNA, 1 µl 10x Poly(A) polymerase reaction buffer, 1 µl 10 mM ATP, 0.25 µl RiboLock RNase inhibitor (Thermo Fisher EO0381), 0.5 µl Poly(A) polymerase and 2.25 µl H2O in total 10 µl reaction volume. The reaction was performed at 37C for 30 min and denatured at 65C for 20 min in a thermal cycler. The polyadenylated RNA was ethanol-precipitated with 90 µl H2O, 0.5 µl Pellet Paint NF co-precipitant (Merck Millipore 70748), 10 µl 3M NaOAc and 250 µl 100% EtOH. After washing with 500 µl 80% EtOH, the pellet was air dried for 5 min and dissolved in 10 µl H2O. The RNA was then DNase treated with TURBO DNase (Thermo Fisher AM2238) using the conditions: 10 µl RNA, 1.5 µl 10x TURBO DNase buffer, 0.6 µl RiboLock RNase inhibitor, 1.5 µl TURBO DNase and 1.4 µL H2O in total 15 µl reaction volume. The reaction was performed at 25C for 60 min and the reaction was stopped by addition of 3 µl 25 mM EDTA and incubation at 65ºC for 10 min in a thermal cycler. The RNA was ethanol precipitated as described above and the pellet was dissolved in 1 µl H2O.

Reverse transcription reaction and addition of the 5' adapter by template switching were performed basically as described in the NanoPARE protocol [158]. Polyadenylated RNA and the anchored dT oligo were denatured at the following condition: 1 µl RNA, 1 µl 10 µM Anchored_dT_2_701_TS_UMI primer and 1 µl 10 mM dNTPs mixture were incubated at 72C for 3 min and put on ice immediately. The RT reaction mix (3 µl denatured RNA mixture, 2 µl 5x first-strand buffer, 0.25 µl 0.1 M DTT, 2 µl 5 M Betaine (Sigma), 1.8 µl 50 mM MgCl2, 1 µl 10 µM TSO-biotin primer, 0.25 µl RiboLock RNase inhibitor and 0.5 µl SuperScript II RT (Thermo Fisher 18064022)) was incubated in a thermal cycler using the following conditions: 42C for 90 min, 10 cycles of 50C for 2 min and 42C for 2 min and 70C for 15 min.

This was followed by the pre-amplification step. PCR was performed with the following reaction mix: 10 µl cDNA sample from the previous step, 0.5 µl 10µM ISPCR primer, 0.5 µl 10 µM ISPCR_3adapter_2 primer, 25 µl 2x KAPA HiFi Hot Start ReadyMix and 14 µl H2O in total 50 µl reaction volume. The PCR was performed using the following cycling conditions: 98C 3min, 20 cycles of 98C for 15s/67C for 20s/72C for 4min and 72C for 5 min. The PCR products were purified using the QIAquick PCR purification kit (Qiagen 28104) and eluted in 20 µL EB buffer. The purified DNA was then size selected by electrophoresis in 2% agarose and excision of gel slices corresponding to size range of 100 bp to 500 bp. Adapter dimers forming a band about 100 bp were carefully excluded. DNA was purified from the gel using the QIAquick gel extraction kit (Qiagen 28704) and eluted in 21 µl EB buffer. DNA concentration was assayed using Qubit dsDNA HS assay kit (Thermo Fisher Q32851) with 1 µl of the sample.

5 ng of the pre-amplified library was then barcoded in the second PCR. PCR reaction mix (5 ng pre-amplified library, 5 µl 10 µM P5_TSO_N5XX primer, 5 µl 10 mM P7_Tn5.2_N701 primer, 25 µl 2x KAPA Hot Start ReadyMix and H2O up to 50 µl in total) was amplified using the following cycling conditions: 98C for 3min, 5 cycles of 98C for 15s/63C for 20 s/72C for 4 min and 72C for 5min. The number of additional PCR cycles needed to properly amplify the barcoded library was optimized using real time PCR. 5 µl of PCR reaction after 5 cycles of amplification was mixed with 1 µl 10 µM P5_TSO_N50X, 1 µl 10 mM P7_Tn5.2_N701, 0.15 µl SYBR Green (1:400), 5 µl 2x KAPA HiFi Hot Start ReadyMix and 2.85 µl H2O in 10µl total reaction volume. Real time PCR was performed using the following cycling conditions: 98C for 30s, 40 cycles of 98C for 15s/63C for 20

Name	Sequence
TSO_seq_read	CTAGCAAGCAGTGGTATCAACGCAGAGTACGGG
TSO_seq_index_1 (i5)	CCCGTACTCTGCGTTGATACCACTGCTTGCTAG
TSO_seq_index_2	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG
TSO_seq_index_2 (i7)	CTGTCTCTTATACACATCTCCGAGCCCACGAGAC

Table 3.1: Oligonucleotides used in this study.

s/72C for 4 min. The number of additional PCR cycles was calculated to reach the amplification plateau. The reaction was performed with barcoding primers as described above. The barcoded library was purified using 1.5x volume of Ampure XP beads (Beckman Coulter A63880. DNA was eluted in 31 µL EB buffer and concentration was quantified using Qubit dsDNA HS assay kit with 1 µL sample.

The libraries were sequenced on the Illumina NextSeq 550 instrument at the University of Michigan Advanced Genomics Core using the following custom primers: TSO_seq_read (Read 1), TSO_seq_index_1 (Index i5), TSO_seq_read2 (Read 2) and TSO_seq_index_2 (Index i7).

3.6.3 Whole genome bisulfite-seq experimental procedure

Genomic DNA was isolated from approximately 3.5-week old Arabidopsis thaliana mature leaf tissue using DNeasy Plant Mini Kit (QIAGEN). DNA was processed for bisulfite treatment and library generation by the University of Michigan Epigenomics Core and Illumina sequencing was carried out at the University of Michigan Advanced Genomics Core.

3.6.4 Trimming, mapping and removing PCR duplicates

The fastq files obtained after sequencing were paired-end reads, with the first read representing the 60bp cDNA fragment and the second read containing the 8 bp long unique molecular identifier (UMI). UMIs for each read were first removed and appended to the read name of the second read pair using UMI-tools extract tool v0.5.4 [169]. Following this, each paired read was combined into a single read where the read name contains the UMI information from paired end read 2, and the sequence corresponds to the RNA sequence from paired end read 1. The reads were then trimmed to remove the 3' adapter and poly(A) sequences using cutadapt v1.9.1 [170] with a maximum allowed error-rate of 0.05 and a minimum length cut-off of 20bps. These trimmed, single-end reads were then mapped to the Arabidopsis TAIR10 genome using bowtie2 v2.2.9 [171] allowing one mismatch.

The mapped bam files were then sorted using samtools v0.1.19 [172]. This was followed by removal of the PCR duplicates using the UMI-tools dedup command [169]. These reads were then converted into the bed format using the bamToBed samtools command and further analysis was done on the bed files using BEDtools v2.25.0 [173]. Reads mapping to the nuclear chromosomes were used for further analysis.

3.6.5 Thinning reads

For every analysis that directly compares the number of reads in two datasets, the Col-0 wild-type reads were thinned by randomly selecting the same number of reads as in the *nrpe1* mutant to keep the comparison fair. The thinning of reads in Col-0 was done using the BEDTools command shuf.

3.6.6 Scoring reads

A scoring algorithm was determined which assigned scores to each IPARE read based on known properties of Pol II and Pol V transcripts. The Pol V properties considered were:

1. Pol V transcribed loci should have considerably higher number of reads in Col-0 compared to Pol V mutant.

2. Pol V transcription is not strand specific.

3. Pol V transcribed regions should not have reads in the Pol V mutant.

The Pol II properties considered were:

1. Pol II reads are enriched at genes.

2. Pol II shows strand-specific transcription.

These reads were assigned arbitrary scores: positive score of +2 for every Pol V property and negative score of -5 or -3 for the Pol II properties. Each read was thereby assigned a score and the total score for every read depicts the likely source of the read; read with the maximum score of +6, are most likely produced by Pol V whereas reads with the minimum score of -8 are most likely produced by Pol II. Possible scores range from -8 to +6 and reads with scores equal or greater than -2 were arbitrarily considered as originating from Pol V.

The genome coverages of the scored reads were calculated without further thinning of the reads as the scoring algorithm removes the noisy reads from the data. This resulted in higher coverage of the scored reads (Fig. 3.2E) compared to the plots of the thinned reads without any filtering (Fig. 3.2C).

3.6.7 Identification of Pol V transcription by HMM

To identify Pol V transcription we used the Hidden Markov Models (HMM) to split the genome into bins based on the IPARE read enrichment between Col-0 and *nrpe1*. We performed HMM using the hmmlearn (https://github.com/hmmlearn/) Python package, which implements the Gaussian-HMM method to cluster the genome in an unsupervised manner. The entire genome was divided into 200 bp windows with overlaps of 50 bp. The inputs to the HMM included the following information for every bin:

1. IPARE ratio of (Col-0 RPM/ *nrpe1* RPM) normalized to sequencing coverage on annotated genes.

2. CHH-methylation difference between Col-0 and *nrpe1*.

3. Col-0 strand bias ratio.

Four emission states were identified, and clusters were then classified as:

State 0 – RdDM Pol V transcripts

State 1 – non-RdDM Pol V transcripts

State 2 – Non-transcribed regions

State 3 – Pol II transcripts

3.6.8 Grouping of the emission states

The emission states were first combined based on whether they are transcribed by Pol V or not. This led to the classification of the emission states into two groups: Pol V transcribed loci (State 0 + State 1) and other, non-Pol V transcribed loci (State 2 + State 3). For other analyses, the HMM emission states were classified into 3 groups: RdDM Pol V transcripts (State 0), non-RdDM Pol V transcripts (State 1) and other loci with no Pol V transcription (State 2 + State 3). These clusters were then filtered further to remove the false positive bins. The following criteria were used:

1. In RdDM and non-RdDM Pol V transcribed subsets (States 0 and 1) bins with no reads in Col-0 were filtered and moved to other loci.

2. Non-RdDM Pol V transcribed bins (State 1) that overlapped genes were tested for the existence of at least one read on the antisense strand in Col-0 and for at least a 2-fold enrichment in Col-0 over *nrpe1* in the antisense strand. Gene-overlapping bins that did not meet these criteria were moved to other loci.

For all genome coverage plots, the overlaps between bins were assigned using the following rules:

1. Regions overlapping between RdDM and non-RdDM Pol V transcription (States 0 and 1) were classified as RdDM bins.

2. Regions overlapping between Pol V transcription and other loci were classified as other loci.

3.6.9 DNA Methylation data analysis

DNA methylation data used in this work were processed using bismark v0.16.1 [174]. The reads were first trimmed for adapters using trim_galore v0.4.1 [175] and reads shorter than 100bps were discarded. Mapping was performed using bismark's non-directional mapping setting. The methylation extraction from the mapped reads were done using bismark_methylation_extractor command and the read coverage information on each read was obtained.

3.6.10 DMR calling and analysis

Differentially methylated regions (DMRs) were identified using methylkit v1.8.0 [176] R package. The DMR calls in Figure 3.7 were done using methylation data from [110] for Col-0, *nrpe1*, *ddm1* and *ddm1 nrpe1* genotypes. CHH-DMRs were identified between *ddm1* and Col-0, where *ddm1* mutant showed a higher level of CHH methylation than Col-0 wild-type. The calls included 5 steps after extraction of methylation information from raw sequencing reads:

1. Read the methylation call data obtained from bisulfite sequencing with methRead function. Define the control and test datasets as Col-0 and *ddm1* respectively.

2. Tile the genome into 200bp windows with a step-size of 150bps using the tileMethylCounts() function. We also added an additional filter of a minimum cytosine coverage of 10 per window.

3. Keep only methylation information for regions with sufficient coverage in all samples using the unite() function.

4. Calculate the differential methylation values at each tile between Col-0 and *ddm1*, looking for hypermethylated regions using calculateDiffMeth() function.

5. Select DMRs with a methylation difference of at least 10% between the datasets and a q-value<0.01 using the getMethylDiff() function.

The *ddm1*-CHH-DMRs overlapping non-RdDM Pol V transcribed regions were identified using BEDTools and the CG and CHH methylation% were plotted as heatmaps at each locus or as a boxplot representing the distribution of the methylation levels at all the DMRs.

3.6.11 Data visualization on boxplots, scatterplots and heatmaps

Reads were counted using BEDTools v2.25.0 at each of the emission states of HMM and RPM normalized. Boxplots were made by plotting the number of RPM normalized reads at identified Pol V transcribed regions. To allow for the visualization of methylation at the individual non-RdDM Pol V transcribed regions overlapping *ddm1*-CHH-DMRs, CHHme and CGme were plotted as heatmaps at the DMRs using methylkit package. Scatterplots were made for (Col-0 - *drm2*) methylation% vs log2(Col-0/ *drm2*) RIP-RPM values calculated at each of the Pol V transcribed and non-transcribed loci using BEDTools.

3.6.12 Datasets produced

Table 3.2 shows all the datasets, sequencing depth and coverage information at every step of data processing for all obtained IPARE datasets. Numbers are in millions of reads.

3.6.13 Previously published sequencing datasets used

Arabidopsis genome annotations and Transposable Element annotations (TAIR10) were obtained from TAIR (www.arabidopsis.org). Pol V RIP-Seq data in Col-0 and *nrpe1* (GSE70290) were published previously [152]. Col-0 and *nrpe1* RNA-Seq data (GSE38464)

Datasets	Exp. group	GEO acc.	Total reads	reads post-trimming	mapped reads	deduplicated reads	nuclear
Col-0 IPARE	1	GSM4409524	24682849	8905307	4083058	3697649	3480799
nrpe1 IPARE	1	GSM4409525	20605406	6028768	1870969	1546483	1245185
spt5l IPARE	1	GSM4409526	22696094	8618717	4013124	3485780	3129984
cmt3 IPARE	1	GSM4409527	15689618	5782750	3094936	2743412	2540780
cmt2 IPARE	1	GSM4409528	19289299	7945143	3983147	3469076	3187090
Col-0 IPARE	2	GSM4409529	17734015	8117778	3673802	2876275	2457277
ago4 IPARE	2	GSM4409530	17681833	9480762	6459238	4499948	4230515
drm2 IPARE	2	GSM4409531	16778574	8329297	4508108	3475234	3114333
Col-0 IPARE	3	GSM4409533	17822479	10298133	4494395	3234001	2851516
nrpe1 IPARE	3	GSM4409534	15078232	8424752	3373309	2136438	1693972
spt5l IPARE	3	GSM4409535	16524586	9968584	5111131	3458921	2978168
ago4 IPARE	3	GSM4409536	14383772	8250043	3883098	2893695	2540634
drm2 IPARE	3	GSM4409537	15851954	9500166	4382153	2945698	2513272
cmt3 IPARE	3	GSM4409538	16817100	9007012	3884148	2922925	2599541
cmt2 IPARE	3	GSM4409539	15733428	8949374	4218221	2950154	2492313
Col-0 IPARE	4	GSM4409540	13305322	3831383	2259788	1805310	1597535
nrpe1 IPARE	4	GSM4409541	13827953	3098158	1956546	1278982	986153
drd1 IPARE	4	GSM4409542	18808526	4914439	3132627	2054232	1662883
dms3 IPARE	4	GSM4409543	16894343	4350932	2806516	1796201	1427323
Col-0 IPARE	5	GSM4409544	5979447	1800536	1002892	808171	741057
nrpe1 IPARE	5	GSM4409545	4336093	1456945	866638	625541	530716
dms5-1 IPARE	5	GSM4409546	5382364	1018487	464376	356695	309202
drd3-3 IPARE	5	GSM4409547	4592926	790364	403213	343871	265525
Col-0 IPARE	6	GSM4409548	4712815	1165390	608064	533196	484864
nrpe1 IPARE	6	GSM4409549	5590278	1077673	508193	406581	353784
dms5-1 IPARE	6	GSM4409550	5086476	994214	515224	406341	345995
drd3-3 IPARE	6	GSM4409551	4616853	822949	449397	369621	288240

Table 3.2: High throughput sequencing datasets obtained in this study. Experimental groups correspond to datasets generated in parallel from plants grown at the same time.

was obtained from [42]. siRNA (GSE99694) were reported previously in [102]. GRO-Seq data from Col-0 and *nrpe1* (GSE100010) were published in [154]. DNA methylation data for Col-0, *nrpe1*, *ddm1* and *ddm1 nrpe1* (GSE79746) were reported by [110].

3.6.14 Data and materials availability

The sequencing data from this study have been submitted to the NCBI Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE146913.

3.7 Authors Contributions

3.7.1 Authors

Masayuki Tsuzuki (M.T.)*: Department of Molecular, Cellular, and Developmental Biology, University of Michigan, Ann Arbor, United States

Shriya Sethuraman (S.S.)*: Bioinformatics Graduate Program, University of Michigan, Ann Arbor, United States

Adriana N. Coke (A.N.C.): Department of Molecular, Cellular, and Developmental Biology, University of Michigan, Ann Arbor, United States

Hafiz M Rothi (M.H.R.): Department of Molecular, Cellular, and Developmental Biology, University of Michigan, Ann Arbor, United States

Alan P Boyle (A.P.B.): Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, United States

Andrzej T Wierzbicki (A.T.W.): Department of Molecular, Cellular, and Developmental Biology, University of Michigan, Ann Arbor, United States

*- Authors contributed equally towards this work.

3.7.2 Individual author contributions

M.T. and S.S. contributed equally to this work.

M.T., S.S., A.N.C., M.H.R., and A.T.W. designed research;

M.T., S.S., A.N.C., and M.H.R. performed research;

M.T., S.S., A.N.C., A.P.B., and A.T.W. analyzed data;

M.T., S.S., and A.T.W. wrote the paper.

My contribution (S.S.) to this work involved all the bioinformatics analysis of the data, including filtering and mapping the data, scoring the reads and building the HMM model for splitting the genome into the different transcription states.

3.8 Publication

The work described in this chapter has been published in the PNAS journal and can be accessed at https://www.pnas.org/content/117/48/30799 [177].

CHAPTER IV

Non-coding RNA-mediated DNA Methylation Directs Nucleosome Positioning

M. Hafiz Rothi[1], ***Shriya Sethuraman***[1], *Jakub Dolata, Alan P. Boyle, Andrzej T. Wierzbicki*

4.1 Abstract

Repressive chromatin modifications are instrumental in regulation of gene expression and transposon silencing. In *Arabidopsis thaliana*, transcriptional silencing is performed by the RNA-directed DNA methylation (RdDM) pathway. In this process, two specialized RNA polymerases, Pol IV and Pol V, produce non-coding RNAs, which recruit several RNA-binding proteins and lead to the establishment of repressive chromatin marks. An important feature of chromatin is nucleosome positioning, which has also been implicated in RdDM. We show that RdDM affects nucleosomes via the SWI/SNF chromatin remodeling complex. This leads to the establishment of nucleosomes on methylated regions, which counteracts the general depletion of DNA methylation on nucleosomal regions. Nucleosome placement by RdDM has no detectable effects on the pattern of DNA methylation. Instead, DNA methylation by RdDM and other pathways affects nucleosome positioning. We propose a model where DNA methylation serves as one of the determinants of nucleosome positioning.

[1] co-first authors

4.2 Introduction

Transcriptional gene silencing (TGS) pathways play an important role in maintaining genomic integrity in eukaryotes. This is achieved through repressive chromatin modifications, which are specifically targeted to silence transposable elements (TE) present in the genome. TGS pathways are conserved in fungi, animal and plant kingdoms, denoting their importance in the proper control of genome stability[26]. In plants, TGS is established and partially maintained through RNA-directed DNA methylation (RdDM), which consists of two major steps, biogenesis of short interfering RNA (siRNA) and de novo DNA methylation[55].

In the first step, RNA polymerase IV (Pol IV) binds to loci targeted for silencing and produces noncoding RNA, which is then converted into a double-stranded form (dsRNA) by RNA-dependent RNA polymerase 2 (RDR2) and cleaved into 24-nucleotide siRNA by DICER-LIKE 3 (DCL3)[81, 178, 179, 82, 83]. siRNAs are then incorporated into ARGONAUTE 4 (AGO4) and other related AGOs, forming AGO-siRNA complexes[84]. In the second step, RNA polymerase V (Pol V) produces long noncoding RNA (lncRNA) that acts as a scaffold or otherwise helps recruit downstream effectors[155, 112, 88]. The AGO4-siRNA complex is recruited to Pol V-transcribed loci leading to step-wise binding of INVOLVED IN DE NOVO 2 (IDN2) and DOMAINS REARRANGED METHYLTRANSFERASE 2 (DRM2) which deposits DNA methylation[89, 90, 88, 87, 180]. However, the mechanisms by which DNA methylation and other repressive features of chromatin contribute to transcriptional gene silencing are not fully understood.

RNA-directed DNA methylation is functionally intertwined with nucleosome modifications and positioning. This includes the involvement of pre-existing histone modification and putative chromatin remodelers in recruitment of both Pol IV and Pol V[157, 82], as well as the the establishment of repressive histone modifications and nucleosome positioning in the second step of RdDM[24, 112, 42, 92, 93]. The involvement of active chromatin remodeling in transcriptional silencing by RdDM was suggested by an interaction of IDN2 with SWITCH 3B (SWI3B), a subunit of the Switch/Sucrose Non Fermenting (SWI/SNF) chromatin remodeling complex[42]. Subunits of this complex also interact with other silencing factors, including HISTONE DEACETYLASE 6 (HDA6) and MICRORCHIDIA 6 (MORC6), which indicates that SWI/SNF may be involved in various aspects of gene silencing[92, 93]. This is consistent with this complex being multi-functional and affecting not only gene silencing but also various other aspects of plant gene regulation [181, 182, 183, 184, 185, 186, 187].

There are several indications that nucleosome positioning and DNA methylation are somehow connected throughout plant genomes[188, 189, 190]. However, the exact nature of this connection varies depending on species and genomic regions tested[189, 190]. In Arabidopsis, nucleosomes determined by MNase digestion protections have been reported to generally correlate with DNA methylation[188]. However,the opposite correlation exists on a subset of Arabidopsis nucleosomes and throughout genomes of certain other species [189, 190]. This difference may be explained by the DNA binding of linker histones, which prevent methylation of linker DNA, and by the activity of DDM1, which facilitates methylation of nucleosomal DNA[190, 134]. In *Arabidopsis* these two proteins counteract the general preference to methylate linker DNA [190].

The involvement of linker histones, DDM1 and SWI/SNF in determining the pattern of DNA methylation indicates that the observed connection between nucleosomes and DNA methylation is primarily determined by nucleosomes being inaccessible to DNA methyltransferases. This is supported by *in vitro* data indicating preferential methylation of linker DNA[191]. However, the opposite relationship has been observed on a few individual loci, where nucleosomes were affected by the *drm2* mutation[42]. This indicates that DNA methylation may affect nucleosome positioning. This alternative causality is also supported by some *in vitro* data[192]. Therefore, the relationship between nucleosomes and DNA methylation remains only partially resolved. Here, we explore the mechanism by which RdDM affects nucleosome positioning in *Arabidopsis thaliana*. We demonstrate that

Pol V and more broadly RdDM affects nucleosomes through the SWI/SNF complex. The SWI/SNF complex is not required for DNA methylation on positioned nucleosomes. Instead, DNA methylation is needed for nucleosome positioning on differentially methylated regions. We propose a model where the RdDM pathway directs nucleosome positioning through DNA methylation to establish transcriptional gene silencing.

4.3 Results

4.3.1 Pol V affects nucleosomes by a combination of direct and indirect mechanisms

Pol V has been previously shown to affect protection to MNase digestion of certain genomic regions [42]. To conclusively attribute these protections to nucleosome positioning, we expanded this experiment by including immunoprecipitation with an anti-H3 antibody (MNase H3 ChIP-seq) in two biological replicates of Col-0 wild-type and *nrpe1*, a mutant of the largest subunit of Pol V (Figure 4.1A). We identified 690 nucleosomes stabilized by Pol V, where signal was at least 2-fold higher in Col-0 compared to *nrpe1* with a false discovery rate (FDR) of less than 0.05 (Figure 4.1B). We also identified 3082 Pol V destabilized nucleosomes, where signal was at least 2-fold higher in *nrpe1* compared to Col-0 with an FDR of less than 0.05 (Figure 4.2A). We validated a subset of Pol V stabilized nucleosomes by locus-specific MNase H3 ChIP-qPCR where we detected a significant decrease in nucleosome signal in *nrpe1* compared to Col-0 wild-type at several tested loci (Figure 4.1C). HSP70 was used as a negative control[193].

To test if Pol V stabilized and destabilized nucleosomes are located within Pol V-transcribed regions, we overlapped identified nucleosomes with previously published Pol V-transcribed regions[152]. Pol V stabilized nucleosomes showed a small overlap with annotated Pol V-transcribed regions (Figure 4.1D), which was still significantly more than expected by chance (Figure 4.1E). Consistently, the average level of Pol V transcription on Pol V stabilized nucleosomes was strongly enriched compared to adjacent regions or

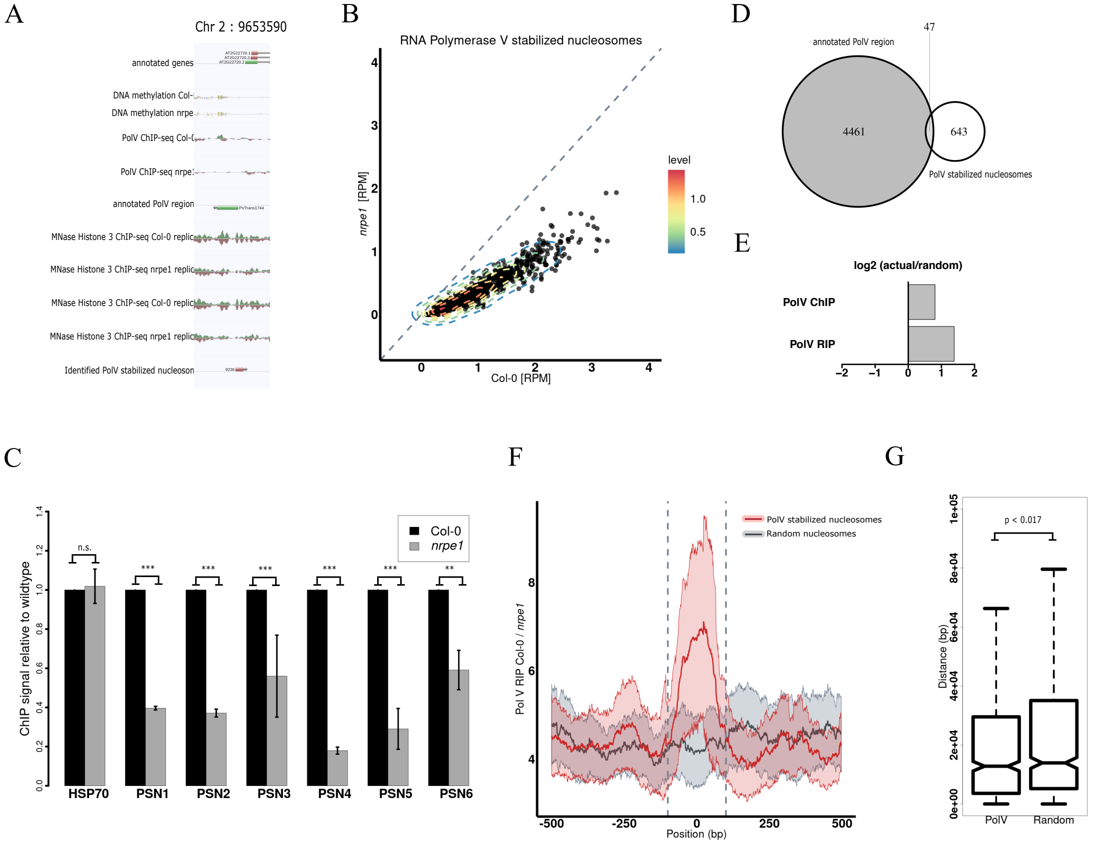

Figure 4.1: **Pol V affects nucleosomes by a combination of direct and indirect mechanisms: (A)** Genome browser screenshot showing a Pol V stabilized nucleosome. **(B)** Comparison of MNase H3 ChIP-seq signal in Col-0 and *nrpe1* on Pol V stabilized nucleosomes. **(C)** Locus-specific validation of Pol V stabilized nucleosomes. Significance tested using t-test (n.s. = not significant, ** = p-value < 0.01,*** = p-value < 0.001). ChIP signal values were normalized to ACTIN2 and Col-0 wildtype. Error bars show standard deviations from three biological replicates. **(D)** Overlap between Pol V stabilized nucleosomes and annotated Pol V transcribed regions. **(E)** Enrichment of Pol V stabilized nucleosomes on annotated Pol V transcribed or bound regions (random permutation test; 1000 iterations; p-value < 0.001). **(F)** Pol V RNA immunoprecipitation signal on Pol V stabilized nucleosomes and random nucleosomes. The nucleosomal regions are indicated with vertical dashed lines. **(G)** Distance of Pol V stabilized nucleosomes or random nucleosomes to annotated Pol V transcribed regions.

random sequences (Figure 4.1F). Furthermore, like Pol V transcription, Pol V stabilized nucleosomes are enriched in intergenic and promoter regions (Figure 4.2DE). On the other hand, overlaps between Pol V destabilized nucleosomes and annotated Pol V-transcribed regions were less likely than expected by chance (Figure 4.2BC). This indicates that Pol V stabilized nucleosomes are at least partially directly affected by Pol V and its downstream factors while Pol V destabilized nucleosome are most likely affected indirectly.

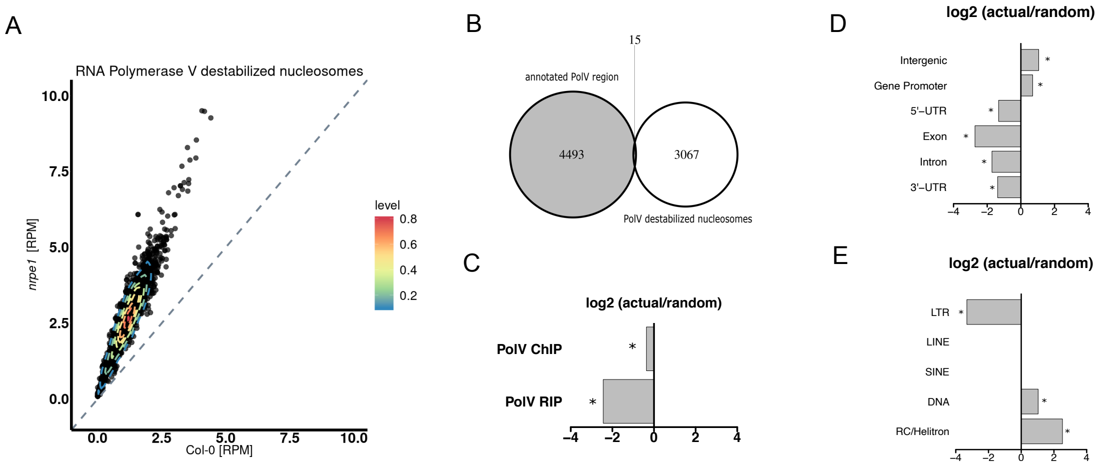

Figure 4.2: **Pol V affects nucleosomes by a combination of direct and indirect mechanisms (Supplementary): (A)** MNase H3-ChIP seq signal on Pol V destabilized nucleosomes. **(B)** Overlap between Pol V destabilized nucleosomes and annotated Pol V transcribed regions. **(C)** Enrichment of Pol V destabilized nucleosomes on annotated Pol V transcribed or bound regions (random permutation test; 1000 iterations; p-value < 0.001). **(D)** Enrichment of Pol V stabilized nucleosomes on various genomic regions (random permutation test; 1000 iterations; p-value < 0.001). **(E)** Enrichment of Pol V stabilized nucleosomes on annotated transposable element regions (random permutation test; 1000 iterations; p-value < 0.001).

To determine if Pol V stabilized nucleosomes that do not overlap Pol V-transcribed regions may still be directly affected by Pol V, we measured their distance from Pol V transcribed regions. The average distance between Pol V stabilized nucleosomes and annotated Pol V-transcribed regions was significantly smaller than the average distance between random nucleosomes and annotated Pol V-transcribed regions (Mann-Whitney test, p-value < 0.017) (Figure 4.1G). This indicates that Pol V may directly affect nucleosomes that do not overlap annotated Pol V transcripts. Altogether, we conclude that Pol V stabilizes a pool of nucleosomes and at least a subset of those nucleosomes is likely to be directly affected by RdDM.

4.3.2 Downstream RdDM components affect nucleosome positioning

Involvement of Pol V in nucleosome positioning suggests that other components of the RdDM pathway may also be involved in this process. To test this prediction, we performed

MNase-H3 ChIP followed by qPCR in Col-0 wild-type, *nrpe1*, *ago4-1* and *idn2-1* mutants. We detected a substantial decrease of the nucleosome signals in all three tested mutants compared to wild-type at Pol V stabilized nucleosomes (Figure 4.3A-G). While *nrpe1*, as expected, affected all tested nucleosomes, *ago4* and *idn2* had more locus-specific effects (Figure 4.3A-G). This indicates that AGO4 and IDN2 both contribute to Pol V-mediated nucleosome positioning. This could be interpreted as evidence that events occurring downstream of Pol V transcription are involved in the observed changes in nucleosome positioning.

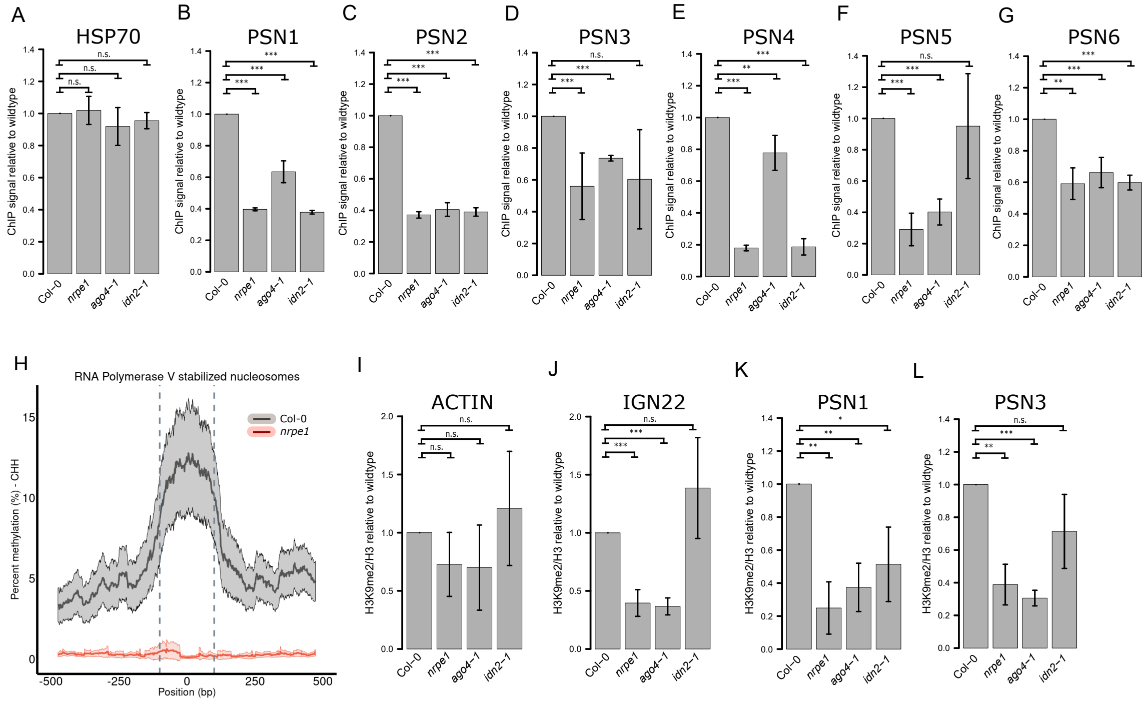

Figure 4.3: **IDN2 connects siRNA and lncRNA to nucleosome positioning: (A)-(G)** Locus-specific analysis of MNase H3-ChIP qPCR levels on Pol V stabilized nucleosomes in Col-0, *nrpe1*, *ago4-1* and *idn2-1*. Significance tested using t-test (n.s. = not significant, ** = p-value < 0.01,*** = p-value < 0.001). ChIP signal values were normalized to ACTIN2 and Col-0 wild-type. Error bars show standard deviations from three biological replicates. **(H)** Average profile of DNA methylation levels (CHH context) on Pol V stabilized nucleosome dyads. **(A)-(G)** Locus-specific analysis of H3K9me2 levels in ACTIN2, IGN22 and Pol V stabilized nucleosomes in Col-0, *nrpe1*, *ago4-1* and *idn2-1*. Significance tested using t-test (n.s. = not significant, * = p-value < 0.05, ** = p-value < 0.01,*** = p-value < 0.001). H3K9me2 ChIP signal values were normalized to H3 and Col-0 wild-type. Error bars show standard deviations from three biological replicates.

4.3.3 Pol V-stabilized nucleosomes are enriched in repressive chromatin marks

RdDM may affect nucleosome positioning in parallel with establishing repressive chromatin marks like DNA methylation and H3K9 dimethylation (H3K9me2). Alternatively, RdDM may establish nucleosomes and repressive chromatin marks on independent subsets of loci. To distinguish between those possibilities, we performed whole-genome bisulfite sequencing in Col-0 wild-type and *nrpe1* in two biological replicates. We plotted DNA methylation levels in the CHH context at Pol V stabilized nucleosomes and 500 bp adjacent regions (Figure 4.3H). CHH DNA methylation was significantly enriched on Pol V stabilized nucleosomes compared to both the adjacent regions and the *nrpe1* mutant (Figure 4.3H). To test if this enrichment is also dependent on AGO4 and IDN2, we used previously published whole-genome bisulfite sequencing datasets[120]. Likewise, we detected a reduction in the average DNA methylation levels in both *ago4-1* and *idn2-1* (Figure 4.4A). These findings indicate that Pol V affects nucleosome positioning in parallel with establishing DNA methylation. This is consistent with genome-wide enrichment of DNA methylation on nucleosomes reported by Chodavarapu et al.[188].

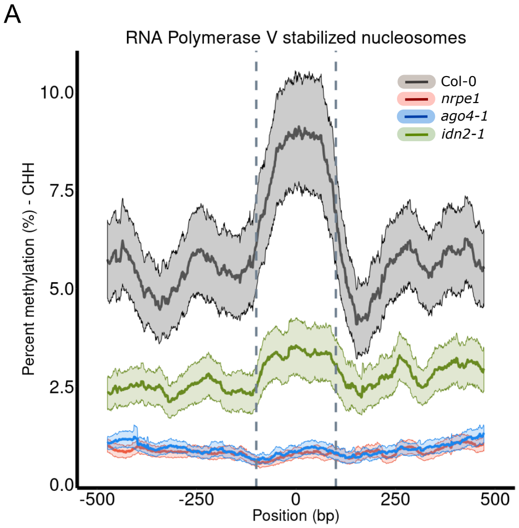

Figure 4.4: **IDN2 connects siRNA and lncRNA to nucleosome positioning (Supplementary): (A)** Average levels of CHH methylation on and around Pol V stabilized nucleosomes dyads using datasets from [120].

To determine if H3K9me2 is also established in parallel, we performed MNase ChIP-qPCR using anti-H3K9me2 antibody in wild-type, *nrpe1*, *ago4-1* and *idn2-1* in three biological replicates. We used the anti-H3 antibody as a reference. The levels of H3K9me2 relative to the levels of H3 were significantly reduced on tested Pol V stabilized nucleosomes in *nrpe1* and *ago4* (Figure 4.3K-L), unchanged on a negative control locus (Figure 4.3I) and reduced on a positive control locus (Figure 4.3J). The *idn2* mutant showed a locus-specific effect, which is consistent with demonstrated partial redundancy of IDN2 and its paralogs[194, 195]. This indicates that at least at the tested loci Pol V affects nucleosome positioning in parallel with establishing H3K9me2. Together, these results indicate that RdDM affects nucleosome positioning in parallel with establishing repressive chromatin marks. This is consistent with a model where Pol V-stabilized nucleosomes are placed and modified by the RdDM pathway.

4.3.4 Pol V positions nucleosomes via the SWI/SNF chromatin remodeling complex

The presence of a physical interaction between an RdDM factor IDN2 and SWI3B, a subunit of the SWI/SNF chromatin remodeling complex[42], indicates that Pol V may affect nucleosomes via the SWI/SNF complex. To test this possibility, we performed MNase H3 ChIP-seq in Col-0 wild-type and *swi3b* mutant. Although *swi3b* null mutants are embryo lethal[186], we took advantage of the well documented observation that SWI3B is haploinsufficient[196, 186, 42] and used the *swi3b/+* heterozygous plants. We plotted the average nucleosome signal at Pol V stabilized nucleosomes and adjacent regions (Figure 4.5A), compared to random nucleosomes identified in Col-0 wild type (Figure 4.5B). Nucleosome signal at Pol V stabilized nucleosomes was reduced in *swi3b/+* (Figure 4.5AC) but unchanged in adjacent regions (Figure 4.5AC) and at random nucleosomes (Figure 4.5BD). Although the effect observed in *swi3b/+* was statistically significant, it was much smaller than in *nrpe1* (Figure 4.5A). A significant but overall minor reduction in the nucleosomal signal in *swi3b/+* was confirmed by locus-specific MNase

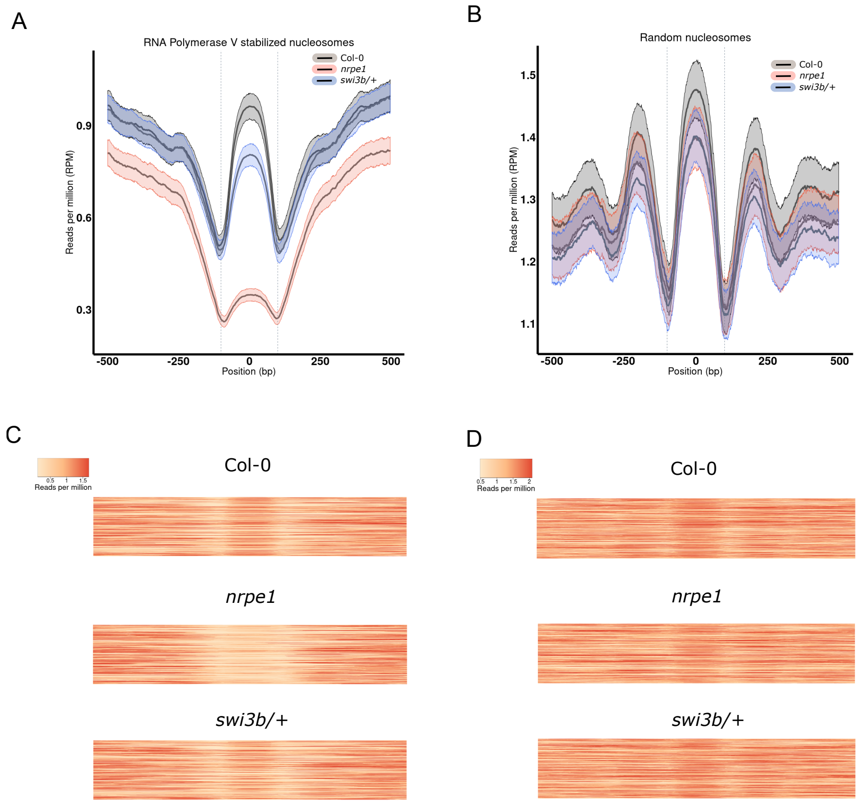

Figure 4.5: **Pol V positions nucleosomes through the SWI/SNF complex:** **(A)** Average levels of MNase H3 ChIP-seq signal on Pol V stabilized nucleosomes in Col-0, *nrpe1* and *swi3b/+*. Ribbons indicate confidence intervals with p < 0.05. **(B)** Average levels of MNase H3 ChIP-seq signal on random nucleosomes in Col-0, *nrpe1* and *swi3b/+*. Ribbons indicate confidence intervals with p < 0.05. **(C)** Heatmap of levels of MNase H3 ChIP-seq signal on Pol V stabilized nucleosomes in Col-0, *nrpe1* and *swi3b/+*. **(D)** Heatmap of levels of MNase H3 ChIP-seq signal on random nucleosomes in Col-0, *nrpe1* and *swi3b/+*.

H3 ChIP-qPCR, where we detected small but significant decreases in nucleosome signal in *swi3b/+* compared to wildtype at tested loci (Figure 4.6A). Partial reductions of nucleosome signals in *swi3b/+* may be explained by the presence of one allele of *SWI3B*, other SWI3 paralogs and other chromatin remodeling complexes. Overall, these results indicate that the SWI/SNF complex only partially contributes to nucleosome positioning by RdDM.

4.3.5 Preferential methylation of linker DNA

Our observation that RdDM establishes both nucleosome positioning and repressive chromatin marks provides a possible explanation to prior observations that nucleosome

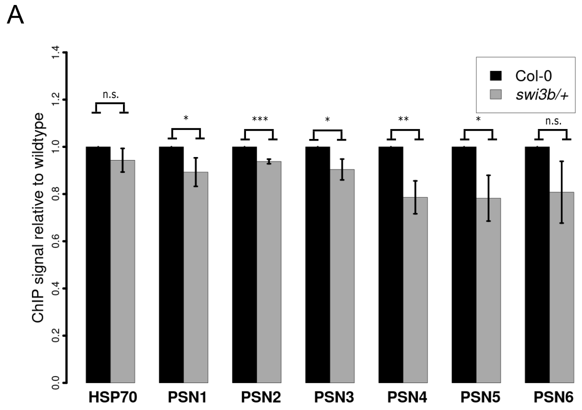

Figure 4.6: **Pol V positions nucleosomes through the SWI/SNF complex (Supplementary): (A)** Locus-specific validation of Pol V stabilized nucleosomes by MNase H3 ChIP followed by qPCR. Significance tested using t-test (n.s. = not significant, * = p-value < 0.05, ** = p-value < 0.01,*** = p-value < 0.001). ChIP signal values were normalized to ACTIN2 and Col-0 wild-type. Error bars indicate standard deviations from three biological replicates.

positioning and DNA methylation are correlated[188, 190]. However, prior studies in Arabidopsis used MNase digestion as the sole basis for the identification of nucleosomes[188, 190] and protections by DNA-binding proteins other than histones remain possible. To eliminate this possibility, we determined nucleosome positioning by MNase H3 ChIP-seq, which relies on MNase protection and binding of histone H3 to DNA to identify nucleosomes. We performed MNase H3 ChIP-seq and whole-genome bisulfite sequencing in two biological replicates of Col-0 wild-type. We first identified all nucleosome positions genome-wide (n=650,610) and measured the average DNA methylation levels in all contexts (CG, CHG and CHH) at nucleosomes and 500 bp adjacent regions. We observed that DNA methylation was enriched on linker regions and depleted on nucleosomes in CHG and CHH sequence contexts (Figure 4.7B-D). The CG methylation pattern was more complex, but linkers of neighboring nucleosomes showed strong enrichments in CG context (Figure 4.7A), which indicates that linkers are preferentially methylated in all sequence contexts. No enrichment was observed on matching random nucleosome-sized regions (Figure 4.8A-C). This indicates that when nucleosomes are identified based on MNase protection and the presence of histone H3, linker regions are enriched in DNA methylation

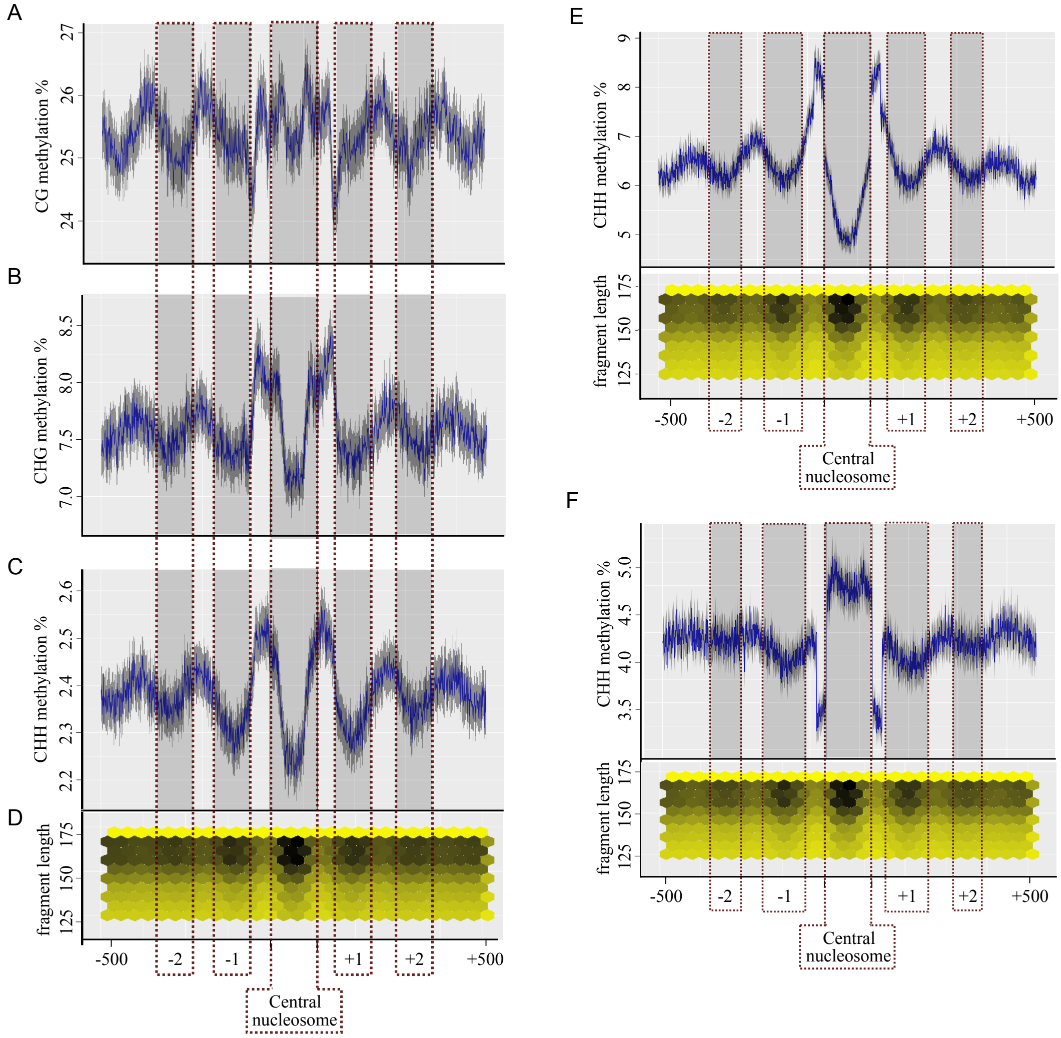

Figure 4.7: **Preferential methylation of linker DNA: (A)-(C)** Average CG **(A)**, CHG **(B)** and CHH **(C)** methylation levels on and around all annotated nucleosomes. Dark grey shading indicates the annotated nucleosome and four neighboring nucleosomes. Ribbon indicates confidence intervals with $p < 0.05$. **(D)** Average MNase H3 ChIP signal levels at and around annotated nucleosomes (X axis) by sequenced fragment length (y axis). **(E)** Average levels of CHH methylation around hypomethylated nucleosomes. Dark grey shading indicates the annotated nucleosome and four neighboring nucleosomes. Ribbon indicates confidence intervals with $p < 0.05$. Scatterplot below shows average MNase H3 ChIP signal levels at and around hypomethylated nucleosomes (X axis) by sequenced fragment length (y axis). **(F)** Average levels of CHH methylation around hypermethylated nucleosomes. Dark grey shading indicates the annotated nucleosome and four neighboring nucleosomes. Ribbon indicates confidence intervals with $p < 0.05$. Scatterplot below shows average MNase H3 ChIP signal levels at and around hypermethylated nucleosomes (X axis) by sequenced fragment length (y axis).

compared to regions occupied by nucleosomes. This correlation is apparent when analyzing all identified nucleosomes (Figure 4.7A-D) and in most subsets of nucleosomes present on specific genomic regions (Figure 4.8D-E).

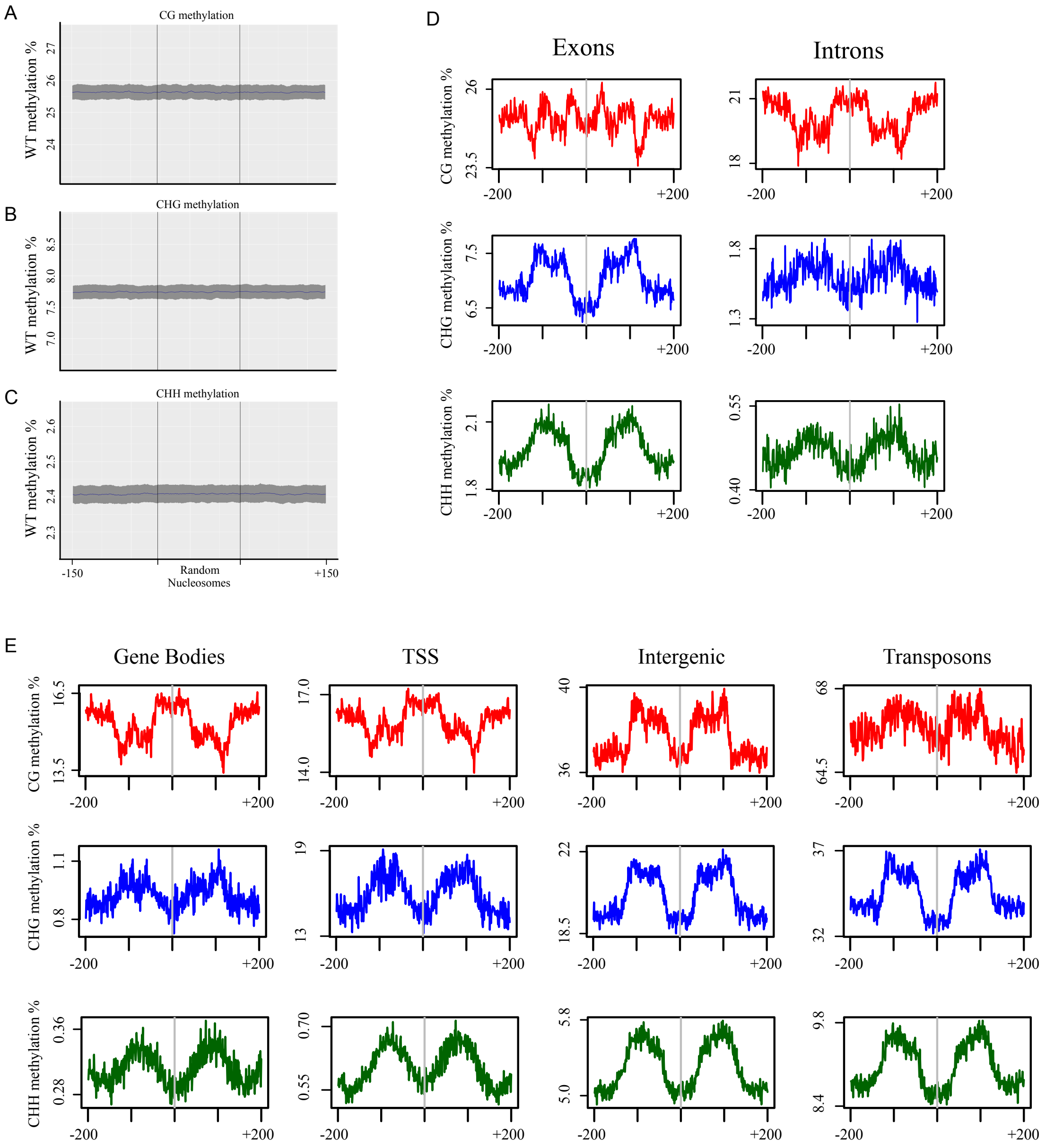

Figure 4.8: **Preferential methylation of linker DNA (Supplementary): (A)** Average levels of CG methylation at random nucleosome-sized regions. Ribbon indicates confidence intervals with p < 0.05. **(B)** Average levels of CHG methylation at random nucleosome-sized regions. Ribbon indicates confidence intervals with p < 0.05. **(C)** Average levels of CHH methylation at random nucleosome-sized regions. Ribbon indicates confidence intervals with p < 0.05. **(D)** Average levels of CG, CHG and CHH methylation at and around annotated nucleosomes overlapping exons or introns. **(E)** Average levels of CG, CHG and CHH methylation at and around annotated nucleosomes overlapping gene bodies, TSS, intergenic regions and transposable elements.

Although the average levels of DNA methylation were higher on linker regions than

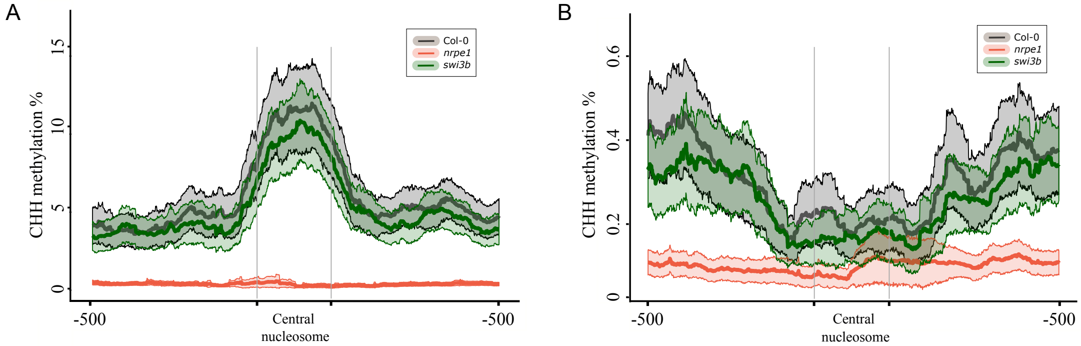

Figure 4.9: **SWI/SNF complex is not required for DNA methylation on positioned nucleosomes: (A)** Average levels of CHH methylation on and around Pol V stabilized nucleosomes. X axis indicates position (bp). Ribbons indicate confidence intervals with $p < 0.05$. **(B)** Average levels of CHH methylation on and around SWI3B stabilized nucleosomes. X axis indicates position (bp). Ribbons indicate confidence intervals with $p < 0.05$.

on nucleosomes, a substantial subset of nucleosomes did not follow this general trend, including Pol V stabilized nucleosomes (Figure 4.3H). To determine if enrichment of DNA methylation on linkers is generally applicable, we focused on subsets of nucleosomes that show enrichment (Figure 4.7E) or depletion (Figure 4.7F) of CHH methylation on their linkers. We then used these subsets to test the enrichment of CHH methylation on linkers of neighboring nucleosomes, which are not expected to be biased by the selection of the central nucleosome. Nucleosomes that follow the general trend, showed the expected enrichment of CHH methylation on linkers of neighboring nucleosomes (Figure 4.7E). Interestingly, nucleosomes filtered for depletion of CHH methylation on their linkers, still showed enrichment of CHH methylation on linkers of neighboring nucleosomes (Figure 4.7F). This further confirms our observation that on average, DNA methylation is enriched on linker regions. We conclude that while Pol V stabilized nucleosomes are enriched in DNA methylation, the general genome-wide trend is preferential methylation of linker DNA.

4.3.6 SWI/SNF complex is not required for DNA methylation on positioned nucleosomes

The general trend of methylation depletion on nucleosomal DNA (Figure 4.7C) is not followed by Pol V stabilized nucleosomes, which are enriched in CHH methylation (Figure 4.3H). This indicates that RdDM overrides general preferences of DNA methylation in respect to nucleosome positioning. This may be explained by a hypothesis that nucleosomes positioned by SWI/SNF are preferential substrates for DNA methyltransferases. To test this hypothesis, we assayed DNA methylation by whole-genome bisulfite sequencing in Col-0 wild-type and *swi3b/+* in two biological replicates. We first analyzed CHH methylation levels on and around Pol V stabilized nucleosomes and observed no change in DNA methylation levels in *swi3b/+* compared to Col-0 wild type (Figure 4.9A). This indicates that the activity of SWI/SNF on Pol V-stabilized nucleosomes has no strong effect on DNA methylation.

To further test if nucleosomes positioned by SWI/SNF affect DNA methylation, we identified SWI3B stabilized nucleosomes, which are defined as nucleosomes that have a higher MNase H3 ChIP-seq signal level in wild-type compared to *swi3b/+* with FDR of less than 0.05. In total, we identified 4089 SWI3B stabilized nucleosomes, where the average nucleosome signal was significantly and reproducibly decreased in swi3b/+ (Figure 4.10A-B). CHH methylation levels on SWI3B stabilized nucleosomes were not significantly changed in *swi3b/+* compared to Col-0 wild type [(Figure 4.9B), (Figure 4.10C)]. This indicates that changed patterns of nucleosome positioning in *swi3b/+* do not affect the levels of CHH methylation. This suggests that nucleosomes positioned by the activity of SWI/SNF do not determine the pattern of DNA methylation.

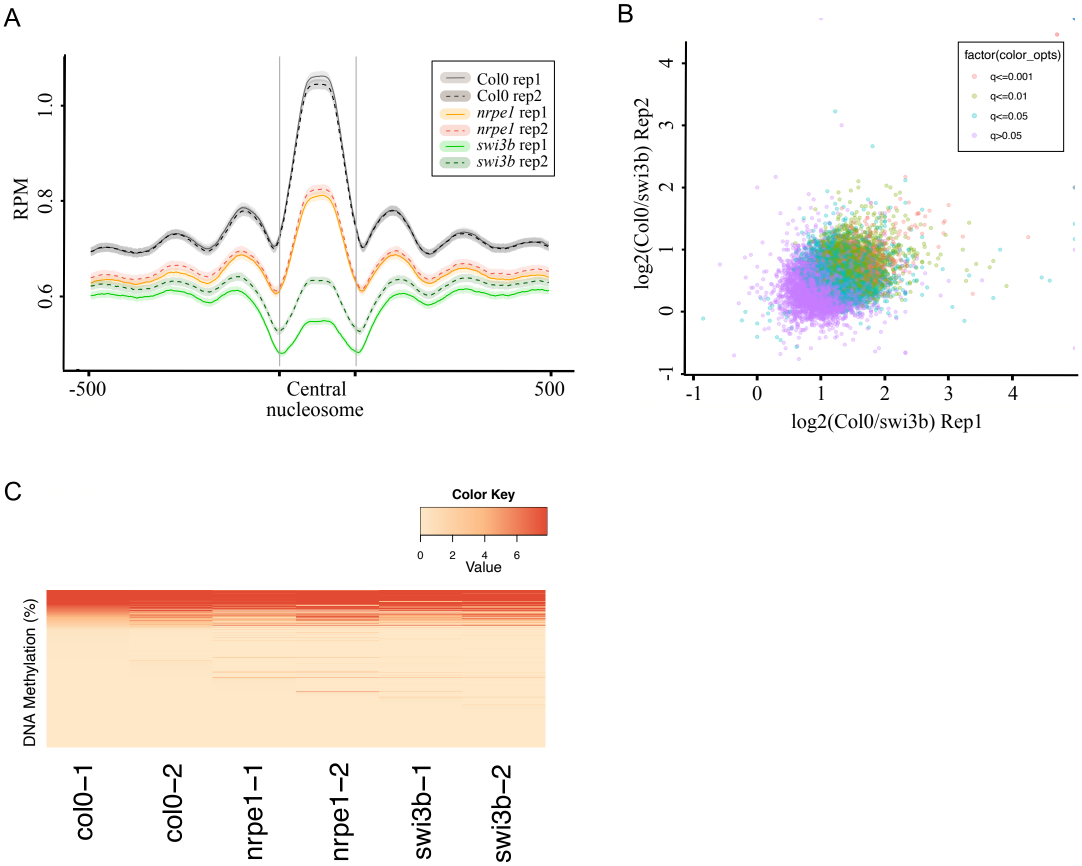

Figure 4.10: **SWI/SNF complex is not required for DNA methylation on positioned nucleosomes (Supplementary): (A)** Average levels of MNase H3 ChIP in two biological replicates of Col-0, *nrpe1* and *swi3b/+* at and around SWI3B stabilized nucleosomes. **(B)** Comparison of biological replicates of MNase H3 ChIP in Col-0 and *swi3b/+*. Colors scale represents FDR cutoff values. **(C)** DNA methylation levels at individual SWI3B stabilized nucleosomes.

4.3.7 DNA methylation is needed for positioning nucleosomes at differentially methylated regions

Our observation that correlations between nucleosomes and CHH methylation are not determined by nucleosome positioning suggests an alternative possibility that DNA methylation may participate in determining positions of nucleosomes. To test this prediction, we used previously published datasets[120] to identify differentially methylated regions (DMRs), where CHH methylation is affected by DRM2. We then assayed nucleosome positioning by MNase H3 ChIP-seq in two biological replicates of Col-0 wild-type and *drm2* mutant. At DRM2 DMRs, where lack of DRM2 resulted in strong reductions of CHH

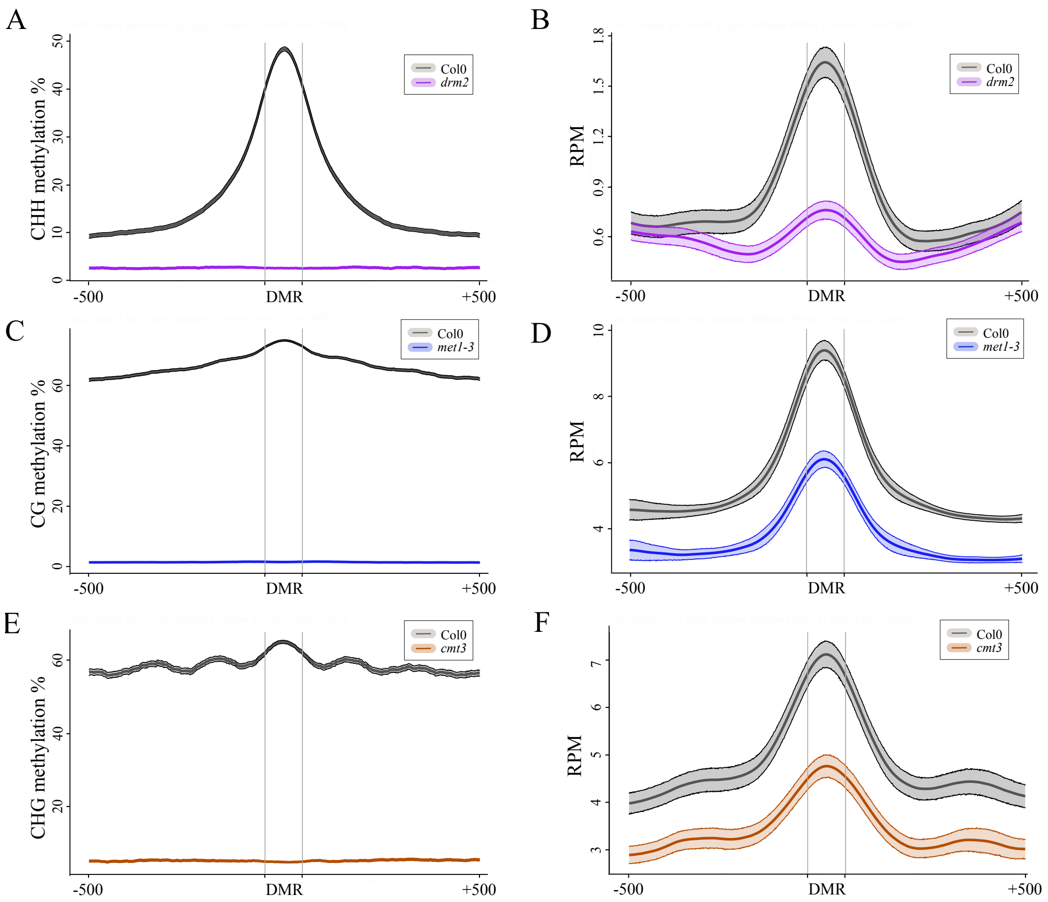

Figure 4.11: **DNA methylation is needed for positioning nucleosomes at differentially methylated regions: (A)** Average levels of CHH methylation at and around regions that lose CHH methylation in the *drm2* mutant (DRM2 DMRs). Ribbons indicate confidence intervals with p < 0.05. **(B)** Average levels of MNase H3 ChIP signal at and around DRM2 DMRs. Ribbons indicate confidence intervals with p < 0.05. **(C)** Average levels of CG methylation at and around regions that lose CG methylation in the *met1* mutant (MET1 DMRs). Ribbons indicate confidence intervals with p < 0.05. **(D)** Average levels of MNase H3 ChIP signal at and around MET1 DMRs. Ribbons indicate confidence intervals with p < 0.05. **(E)** Average levels of CHG methylation at and around regions that lose CHG methylation in the *cmt3* mutant (CMT3 DMRs). Ribbons indicate confidence intervals with p < 0.05. **(F)** Average levels of MNase H3 ChIP signal at and around CMT3 DMRs. Ribbons indicate confidence intervals with p < 0.05.

methylation (Figure 4.11A), the nucleosome signal was generally enriched in Col-0 wild type (Figure 4.11B). This is consistent with Pol V stabilized nucleosomes being enriched in CHH methylation (Figure 4.3H). It is also in agreement with the observation that nucleosomes overlapping DRM2 DMRs behave like Pol V stabilized nucleosomes and are enriched in both CHH methylation and nucleosome signal (Figure 4.12A-B). Importantly, in the *drm2* mutant, DRM2 DMRs had a strong reduction in the nucleosome signal (Figure 4.11B). This indicates that DNA methylation in the CHH context established by the RdDM pathway affects nucleosome positioning.

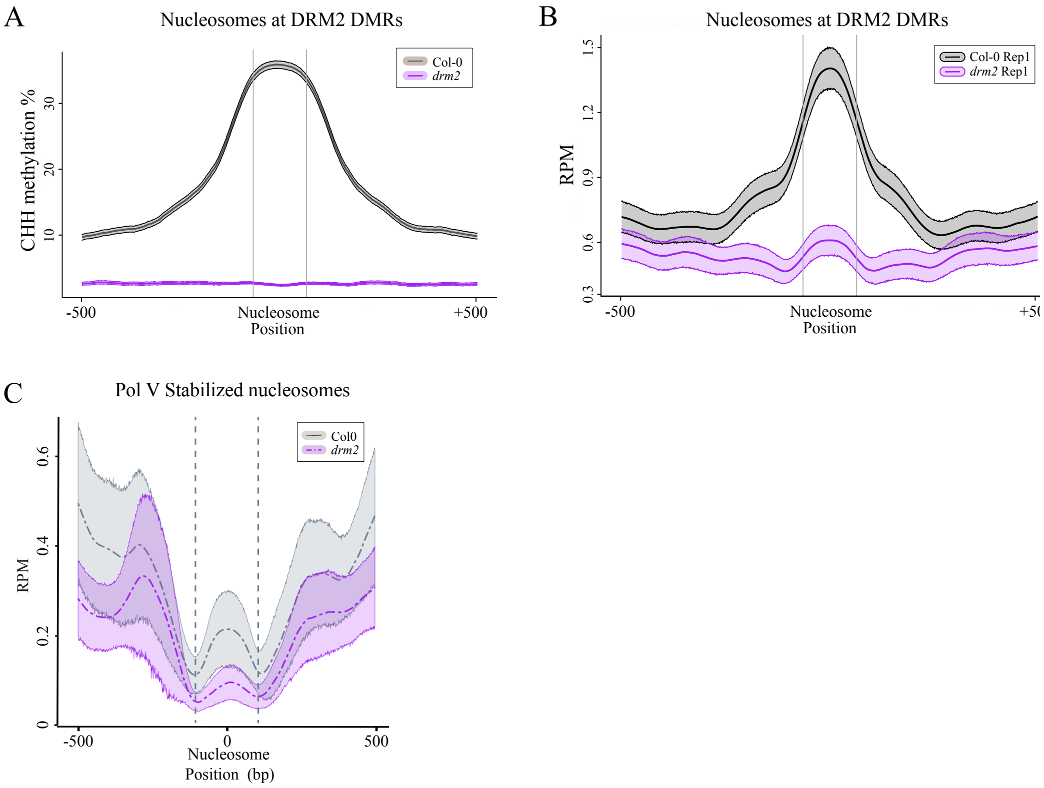

Figure 4.12: **DNA methylation is needed for positioning nucleosomes at differentially methylated regions (Supplementary): (A)** Average levels of CHH methylation at and around annotated nucleosomes that overlap DRM2 DMRs. Ribbons indicate confidence intervals with $p < 0.05$. **(B)** Average levels of MNase H3 ChIP signal at nucleosomes overlapping DRM2 DMRs. Ribbons indicate confidence intervals with $p < 0.05$. **(C)** Average levels of MNase H3 ChIP signal at Pol V stabilized nucleosomes. Ribbons indicate confidence intervals with $p < 0.05$.

To test if this also applies to Pol V stabilized nucleosomes, we plotted the MNase H3 ChIP signal on Pol V stabilized nucleosomes in the *drm2* mutant compared to Col-0 wild type. Consistently with observations on DRM2 DMRs (Figure 4.11B) and nucleosomes overlapping DRM2 DMRs (Figure 4.12B), Pol V stabilized nucleosomes also had a reduction in nucleosome signal in the *drm2* mutant (Figure 4.12C). This further supports our observation that DNA methylation in the CHH context established by the RdDM pathway affects nucleosome positioning.

To test if DNA methylation in CG and CHG contexts also affects nucleosome positioning we performed similar experiments and analysis in *met1* and *cmt3* mutants. CG and CHG DMRs identified in *met1* and *cmt3*, respectively, were enriched in the nucleosomal signal (Figure 4.11DF). At MET1 DMRs, where lack of MET1 resulted in strong reductions of CG methylation (Figure 4.11C), the nucleosome signal was significantly reduced in *met1* (Figure 4.11D). Similarly, at CMT3 DMRs, where lack of CMT3 resulted in strong reductions of CHG methylation (Figure 4.11E), the nucleosome signal was also significantly reduced in *cmt3* (Figure 4.11F). This indicates that DNA methylation affects nucleosome positioning irrespective of the sequence context.

4.4 Discussion

We propose a model, where DNA methylation is a determinant of nucleosome positioning in RdDM. In this model, non-coding transcription by both Pol IV and Pol V leads to the recruitment of AGO4 and IDN2. IDN2 interacts with a subunit of SWI/SNF, which is however not sufficient to affect nucleosome positioning. Instead, the subsequent recruitment of DRM2 and establishment of DNA methylation activates chromatin remodelers and leads to changes in nucleosome positioning. Coordinated establishment of various chromatin marks leads to repression of Pol II promoters within the silenced region of the genome.

The effect of DNA methylation on nucleosome positioning may be explained by distinct intrinsic properties of DNA containing 5-methylcytosines, as suggested by [192]. Alternatively, DNA methylation may facilitate the recruitment or activation of SWI/SNF, either directly or by the involvement of other proteins that are sensitive to the presence of 5-methylcytosines. Another possibility is that DNA methylation may affect nucleosome positioning by changing the pattern of posttranslational histone modifications. This includes H3K9me2, which may recruit proteins that modulate the activity of chromatin remodelers. This also includes histone deacetylation, which may affect physical properties of the nucleosomes [93, 197].

The importance of DNA methylation for nucleosome positioning has a significant impact on our understanding of the RdDM pathway. It argues against the pathway being branched after IDN2 recruitment[42]. Instead, it supports the notion that events occurring co-transcriptionally at the sites of Pol V transcription are organized in a step-wise genetic pathway[90]. Although when studied genetically, this pathway appears linear, various steps of the pathway are likely to rely on the cooperative recruitment or activation of subsequent factors. One example of such a connection is the requirement of both IDN2-SWI3B interaction and DNA methylation for nucleosome positioning. Other examples include the recruitment of AGO4, which has been proposed to rely on the interaction of AGO4 with NRPE1 C-terminal domain and with Pol V transcripts[133, 88]. Similarly, there is evidence of DRM2 being recruited by interactions with AGO4 and other RdDM factors[107, 91].

Our model is consistent with the notion that events in the late stages of RdDM lead to a concerted establishment of DNA methylation, posttranslational histone modifications and nucleosome positioning, which together form a repressive chromatin structure. This explains the robustness of transcriptional silencing, where coordinated establishment of various repressive chromatin marks leads to efficient repression of Pol II transcription. It is also consistent with the general difficulty to experimentally tease apart various repressive chromatin modifications established by this pathway.

The involvement of SWI/SNF and nucleosome positioning in RdDM may also be considered in context of this pathway performing mostly maintenance of silencing. Transcription of heterochromatic regions by Pol IV and Pol V may involve the removal or repositioning of previously positioned nucleosomes. This is supported by the involvement of putative chromatin remodelers in initiation and/or elongation of transcription by both of those polymerases[105, 112, 102, 198]. Nucleosome positioning established as an outcome of RdDM may serve to re-create the pattern of nucleosomes disrupted by Pol IV and Pol V. *De novo* RdDM in newly inserted TEs is a distinct scenario, since no pre-existing repressive chromatin modifications are expected to exist. The role of nucleosome positioning in this *de novo* process remains unexplored.

The involvement of DNA methylation in determining the pattern of nucleosomes extends beyond RdDM targets. The impact of MET1-dependent CG methylation and CMT3-dependent CHG methylation on stabilizing nucleosomes indicates that DNA methylation may affect nucleosome patterns throughout the genome. This is consistent with findings in other eukaryotes[199, 192]. Such a general effect of DNA methylation on nucleosome positioning would counteract the general preference to methylate linkers and contribute to local correlations between nucleosomes and DNA methylation. This property of nucleosomes is consistent with previous reports[188] and may involve the activity of DDM1[190]. It illustrates the general interdependence between nucleosomes and DNA methylation.

Existing evidence does not support the view that DNA methylation is the primary determinant of the nucleosome pattern. This role remains reserved for a combination of intrinsic factors and active chromatin remodeling. The role of DNA methylation is more limited and probabilistic, clearly visible in meta-analysis of large pools of sequences. Therefore, opposite behaviors of individual loci are expected. Moreover, global losses of DNA methylation in RdDM and DNA methyltransferase mutants may affect the patterns of nucleosomes by a combination of cis- and trans-acting factors, which could only be distinguished using tools targeting DNA methylation to specific loci.

4.5 Materials and Methods

4.5.1 Plant material

Col-0 ecotype (wild-type), *nrpe1* (nrpd1b-11 [145]), *ago4-1*(introgressed into the Col-0 background [88]), *idn2-1* [89], *drm2-2* (SAIL_70_E12) was described previously [112], *swi3b-2* (GABI_302G08 [186] and *cmt3-11* (SALK_148381). *met1-3* which was described previously [146] were grown at 22C under white LED light in 16h/8h day/night cycle.

4.5.2 Antibodies

Rabbit polyclonal anti- histone H3 antibody (ab1791) and mouse monoclonal anti-H3K9me2 antibody (ab1220) were obtained from Abcam.

4.5.3 MNase H3 ChIP-seq

2g of approximately 3.5-week old *Arabidopsis thaliana* mature leaf tissue, which was cross-linked with 0.5% formaldehyde, was ground in liquid nitrogen. MNase H3 ChIP of Col-0, *met1*, *cmt3* and *drm2* was carried out as described previously[42]. MNase H3 ChIP of Col-0, *nrpe1* and *swi3b* was carried out using the following protocol. Cold nuclei isolation buffer I (10 mM Tris HCl pH8, 10mM MgCl2, 0.4 M sucrose, 0.035% β-mercaptoethanol, 1mM phenylmethylsulfonyl fluoride (PMSF)) was added. Tissue was resuspended by vigorous vortexing and shaking. Sample was filtered using Miracloth into new 50 ml tube on ice. Miracloth was washed with 10 ml of nuclei isolation buffer I. Sample was centrifuged 15 min, 4000 g, 4C.

Supernatant was discarded and nuclei pellet was resuspended using 1 ml of cold nuclei isolation buffer II (10 mM Tris HCl pH8, 10 mM MgCl2, 0.4 M sucrose, 1% Triton X-100, 0.035% β-mercaptoethanol, 1mM phenylmethylsulfonyl fluoride (PMSF), 0.02 tab/ml complete EDTA-free, 0.004 mg/ml Pepstatin A). Sample was transferred to 1.5 ml tube and centrifuged for 5 min, 2000 g, 4C. This step was repeated two more times. Pellet was resuspended using 300 μl of cold Nuclei isolation buffer II and layered on top of cold 900 ml Nuclei isolation buffer III (10 mM Tris HCl pH8, 2 mM MgCl2, 1.7 M sucrose, 0.15% Triton X-100, 0.035% β-mercaptoethanol, 1 mM phenylmethylsulfonyl fluoride (PMSF), 0.02 tab/ml complete EDTA-free, 0.004 mg/ml Pepstatin A) in 1.5 ml tube. Sample was centrifuged for 30 min, 16000 g, 4C and supernatant was discarded.

Isolated nuclei were washed twice with Micrococcal Nuclease (MNase) reaction buffer (10 mM Tris HCl pH8, 15 mM NaCl, 60 μM KCl, 1mM CaCl2) and resuspended in the same buffer. MNase enzyme (NEB; 200 Kunitz unit/μl) was added and samples were mixed by vortexing. Samples were digested for 10 minutes at 30°C. 1 volume of MNase stop buffer (30 mM Tris HCl pH8, 225 mM NaCl, 10 mM ethylenediaminetetraacetic acid (EDTA), 10 mM egtazic acid (EGTA), 0.2% sodium dodecyl sulphate (SDS), 2% Tween 20) was then added to stop the reaction. To release the chromatin from the nuclei, the sample was vortexed vigorously 5 times and centrifuged for 10 min, 14000 g. The supernatant was then transferred to a new tube. Samples for H3 ChIP were then diluted in 1 volume ChIP dilution buffer (16.7 mM Tris HCl pH8, 1.2 mM ethylenediaminetetraacetic acid (EDTA), 167 mM NaCl, 1.1% Triton X-100, 1 mM phenylmethylsulfonyl fluoride (PMSF), 0.02 tab/ml cOmplete EDTA-free, 0.004 mg/ml Pepstatin A). H3 antibody was added and sample was incubated 12-16 hours, 4C with rotation.

Protein A magnetic beads (PierceTM) were washed three times with IP buffer (50 mM HEPES pH7.5, 150 mM NaCl, 10 µM ZnSO4, 1% Triton X-100, 0.05% sodium dodecyl sulphate (SDS), 1 mM phenylmethylsulfonyl fluoride (PMSF), 0.02 tab/ml cOmplete EDTA-free, 0.004 mg/ml Pepstatin A) and resuspended in 50 µl IP buffer. Beads were added to IP sample and incubated for 1 hour, 4C with rotation. Immunoprecipitated chromatin bounded to magnetic beads was collected using magnetic separator. Beads were washed 5 min with cold buffers: two times with low salt buffer (20 mM Tris HCl pH8, 2 mM ethylenediaminetetraacetic acid (EDTA), 150 mM NaCl, 1% Triton X-100, 0.1% sodium dodecyl sulphate (SDS)), once with high salt buffer (20 mM Tris HCl pH8, 2 mM ethylenediaminetetraacetic acid (EDTA), 0.5 M NaCl, 1% Triton X-100, 0.1% sodium dodecyl sulphate (SDS)), once with LiCl buffer (20 mM Tris HCl pH8, 2 mM ethylenediaminetetraacetic acid (EDTA), 250 mM LiCl, 1% NP-100, 1% sodium deoxycholate)) and twice with TE buffer (10 mM Tris HCl pH8, 1 mM ethylenediaminetetraacetic acid (EDTA)). After the last wash, samples were transferred into new a tube and beads were collected using a magnetic separator.

For library preparation, magnetic beads were incubated with 100 µl Elution buffer (10 mM Tris HCl pH8, 1 mM ethylenediaminetetraacetic acid (EDTA), 1% sodium dodecyl sulphate (SDS)) in a thermomixer (65C, 1400 rpm, 30 min). Beads were collected using magnetic separator and supernatant was transferred into new tube. Step was repeated and both supernatants combined. IP samples were de-crosslinked by Proteinase K treatment (5 µl, 65C, 12 h). Samples were purified using QIAquick PCR Purification Kit (35 µl of EB buffer were used). Library for Illumina sequencing was prepared using either MicroPlex Library PreparationTM Kit (Diagenode) according manufacturer instruction, using in-house library preparation based on Bowman et al[200], or prepared by the University of Michigan Advanced Genomics Core. MNase ChIP-seq experiments were performed in two biological replicates and sequenced by either 50 bp or 150 bp paired-end sequencing at the University of Michigan Advanced Genomics Core.

4.5.4 MNase H3 and H3K9me2 ChIP-qPCR

Nuclei were extracted from 2g of approximately 3.5-week old Arabidopsis thaliana mature leaf tissue which was cross-linked with formaldehyde [0.5%] as described previously [42] and were digested with Micrococcal Nuclease (MNase ; NEB) for 10 minutes at 30C. MNase-digested chromatin was immunoprecipitated with anti-histone H3 antibody or anti-H3K9me2 antibody. DNA was purified and used for qPCR analysis. MNase ChIP-qPCR experiments were performed in three biological replicates with region-specific primers listed in Table 4.1.

4.5.5 Whole genome bisulfite-seq (WGBS)

Genomic DNA was isolated from approximately 3.5-week old *Arabidopsis thaliana* mature leaf tissue using DNeasy Plant Mini Kit (QIAGEN). DNA was processed for bisulfite treatment and library generation at the University of Michigan Advanced Genomics Core.

Locus	Name	Sequence (5’-3’)	Application
Nucleosome validation			
PSN1	MH487	caggttgtgagttcgaatcgt	ChIP-qPCR
	MH488	catctccgttagccaccttt	ChIP-qPCR
PSN2	MH489	tgagattttaccgggtccac	ChIP-qPCR
	MH490	cccttatacgtaatttccatcaca	ChIP-qPCR
PSN3	MH491	ggagtgggatgtagactcgaa	ChIP-qPCR
	MH492	ctagtggtaccgcagggttt	ChIP-qPCR
PSN4	MH493	cgatcggttcgatctcctta	ChIP-qPCR
	MH494	taacggttcaacccgagaaa	ChIP-qPCR
PSN5	MH495	tctcccccacaatttctgtc	ChIP-qPCR
	MH496	aaatggacccctcattgtca	ChIP-qPCR
PSN6	MH501	acagatagcgctgtacagatttta	ChIP-qPCR
	MH502	tcatttgatatgcgttttgttt	ChIP-qPCR
ACTIN2	Actin2-A118	gagagattcagatgcccagaagtc	ChIP-qPCR[112]
	Actin2-A119	tggattccagcagcttcca	ChIP-qPCR[112]
HSP70	A512	ctcttcctcacacaatcataaaca	ChIP-qPCR[193]
	A513	cagaattgttcgccggaaag	ChIP-qPCR[193]
IGN22	MH537	cgggtccttggactcctgat	ChIP-qPCR[168]
	MH538	tcgtgaccggaataattaaatgg	ChIP-qPCR[168]
H3K9me2 validation			
ACTIN2	Actin2-A118	gagagattcagatgcccagaagtc	ChIP-qPCR[112]
	Actin2-A119	tggattccagcagcttcca	ChIP-qPCR[112]
PSN1	MH487	caggttgtgagttcgaatcgt	ChIP-qPCR
	MH488	catctccgttagccaccttt	ChIP-qPCR
PSN3	MH491	ggagtgggatgtagactcgaa	ChIP-qPCR
	MH492	ctagtggtaccgcagggttt	ChIP-qPCR

Table 4.1: Oligonucleotides used in this study

4.5.6 Bioinformatic analysis

MNase H3 ChIP-seq paired-end reads from two independent biological replicates were aligned and processed to the Arabidopsis TAIR10 genome with Bowtie2 [171]. Mapped reads were deduplicated using PICARD tools (http://broadinstitute.github.io/picard) and filtered by fragment length between 120-170 bp and MAPQ value of >=2. Differential nucleosomes were identified using DANPOS2[201] by filtering nucleosomes with more than 2 fold enrichment in either in Col-0 for Pol V stabilized nucleosomes or in *nrpe1* for Pol V destabilized nucleosomes and FDR< 0.05. Nucleosomes are then filtered using the negative binomial test with reads from biological replicates using the NBPseq R package[151]. For subsequent analysis we selected nucleosomes which showed more than 2 fold-change and FDR < 0.05. We further refined the nucleosome positions for well-positioned nucleosomes by filtering for main peak nucleosomes using iNPS[202]. Nucleosome data was (RPM) normalized and visualized on heatmaps and profiles by calculating the number of reads using BEDTools 2.15.0 at nucleosome dyads[173]. Overlap analyses with nucleosomes were performed with 1000 permuted genomic regions to obtain expected numbers and p-values. SWI3B-stabilized nucleosomes were filtered for higher read counts in Col-0 than the *swi3b* mutant and an FDR<0.05. These nucleosomes were then further filtered using the negative binomial test with reads from biological duplicates using NBPseq and the nucleosomes with FDR<0.01 were selected for further analysis.

Hypermethylated nucleosomes were identified by filtering for nucleosomes with higher DNA methylation level in CHH-context inside the nucleosomes (140bps), compared to their adjacent DNA linker regions (30bps upstream and downstream of nucleosomes). The nucleosomes were filtered for the presence of more than 8 CHH-context cytosines within the nucleosomes and more than 2 CHH-context Cs in each of the adjacent linkers to correct for the sizes of the regions and frequencies of Cs. Hypomethylated nucleosomes were similarly identified, except these regions had higher levels of CHH-context DNA methylation in the adjacent DNA linker regions than the nucleosomes.

The sequencing reads from whole genome bisulfite-seq datasets were mapped to the TAIR10 genome using the Bismark software allowing no mismatches[203]. DNA methylation levels were calculated by the ratio of #C/(#C+#T) after selecting for Cs with at least 5 sequenced reads. Differentially Methylated Regions (DMRs) were identified using methylKit package in R[204]. The bin sizes used were 100bp bins with a step-size of 50bps. 10 minimum bases were required in each tile. A 10% minimum methylation difference was selected for in each of the tiles and an FDR value of 0.01 was used. The number of MNase-H3 ChIP-seq reads overlapping these DMRs were then plotted as a profile.

4.5.7 Other datasets used in this study

Arabidopsis genome annotations (TAIR10) were obtained from TAIR. Pol V ChIP-seq data (SRA054962) and peak list and Pol V RIP-seq data (GSE70290) and annotated regions were published previously [152, 56]. DNA methylation data from *idn2*, *ago4*, *drm2* and *cmt3* mutants as well as corresponding Col-0 and *nrpe1* controls were obtained from GSE39901 [120].

4.5.8 Data access

The sequencing data from this study have been submitted to the NCBI Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE148173.

4.6 Author Contributions

4.6.1 Authors

Hafiz M Rothi [M.H.R.]*: Department of Molecular, Cellular, and Developmental Biology, University of Michigan, Ann Arbor, United States

Shriya Sethuraman [S.S.]*: Bioinformatics Graduate Program, University of Michigan, Ann Arbor, United States

Jakub Dolata [J.D.]: Department of Molecular, Cellular, and Developmental Biology, University of Michigan, Ann Arbor, United States

Alan P Boyle [A.P.B.]: Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, United States

Andrzej T Wierzbicki [A.T.W.]: Department of Molecular, Cellular, and Developmental Biology, University of Michigan, Ann Arbor, United States

*- Authors contributed equally towards this work.

4.6.2 Individual author contributions

M.H.R. and S.S. contributed equally to this work.

M.H.R., S.S., J.D., and A.T.W. designed research.

M.H.R., S.S., and J.D. performed research.

M.H.R., S.S., A.P.B., and A.T.W. analyzed data.

M.H.R., S.S., and A.T.W. wrote the paper.

M.H.R. was instrumental in performing the experimental design, planning, data acquisition and analysis of the first half of the manuscript, including the sections: 4.3.1 – 4.3.4. He was also instrumental in writing up the manuscript.

My contribution (S.S.) to this project involved the experimental design, planning, data acquisition and analysis of the second half of the manuscript, including the sections: 4.3.5 – 4.3.7. I also assisted in writing the final manuscript.

4.7 Publication

The work described in this chapter has been submitted and is also accessible in bioRxiv [205].

CHAPTER V

Concluding Remarks and Future Directions

The main focus of this dissertation has been to better understand the RdDM pathway with extensive new insights into the mechanism of Pol V transcription and its impact on gene regulation. Through this dissertation, I have identified Pol V transcribed regions genome-wide to better understand the RdDM mechanism and the factors involved in the pathway (Chapter II). I have also shown that Pol V transcription is not limited to RdDM loci but it is more widespread and pervasive, depicting a possible surveillance mechanism of transcription (Chapter III). I have also focused on the downstream effects of the RdDM pathway showing that Pol V transcription regulates gene expression by controlling DNA methylation and nucleosome positioning. I have demonstrated how DNA methylation directs nucleosomes to the RdDM loci which, in turn, leads to regulation of gene expression (Chapter IV). The analysis of various epigenomic and transcriptomic datasets has not only led to these important new findings but also paved the way to explore many new areas of research in the field of transcriptional gene silencing.

5.1 Genome-wide identification of lncRNA transcribed by Pol V

Chapters II and III have led to the substantial expansion of the known RdDM loci. This was made possible by identification of lncRNAs produced by Pol V through RIP-Seq and IPARE-Seq, respectively. Before the start of this study, our knowledge of *in vivo* functions of Pol V and the lncRNAs produced by it were limited to certain locus-specific assays [88, 57, 42, 90, 112].

Chapter II led to the identification of ~4500 highly significant Pol V transcribed lncRNAs covering approximately 2.6% of the genome. These identified Pol V transcribed regions further solidified our preexisting knowledge about RdDM. These lncRNAs depicted interaction with other proteins linked to the RdDM pathway like AGO4 and IDN2 along with enrichment of essential features of RdDM like CHH methylation and CG methylation. They were also enriched at TEs. This depicted that the loci identified in the study using genome-wide assays were, indeed, Pol V transcribed RdDM loci.

Chapter III showed a further improvement and expansion of the identification of the Pol V transcribed lncRNAs by the development of a new, modified RIP technique called IPARE. The improved quality of sequencing data and a new and better analysis technique utilizing an unsupervised HMM model, led to the identification of RdDM Pol V transcripts that covered 23% of the entire genome. The study also identified non-RdDM Pol V transcribed regions showing that Pol V transcription is more pervasive, covering 42%, if not more, of the entire genome. This also led to the new understanding that Pol V might function at regions outside of RdDM loci and might play an important role in the maintenance of genomic integrity.

Our initial understanding of the RdDM pathway was that Pol V produces lncRNA to silence TEs and maintain the TEs in a repressed state[55]. Our knowledge of Pol V transcription suggested that recruitment of Pol V to the TEs mostly involved preexisting chromatin modifications [100, 82, 101, 103] and not sequence-encoded promoters [160]. This explained the maintenance mechanism of the RdDM pathway well, where Pol V recruitment to its loci depended on the previously established repressive modifications by Pol V itself. This self-reinforcing loop mechanism, however, has never been able to explain the mechanism of recruitment of Pol V to *de novo* RdDM loci, which include novel TE insertions with no preexisting chromatin modifications. The currently improved techniques for identification of these Pol V transcripts have greatly increased our understanding about the mechanism of Pol V transcription.

Our model in Figure 3.9 suggests that Pol V exists in three transcription states spanning 42% of the genome. The maintenance RdDM state occurs at silenced TEs and depicts a high level of Pol V transcription supported by the self-reinforcing Pol V recruitment mechanism [100, 82, 101, 103]. The surveillance transcription state occurs at non-RdDM loci that are scanning for possible random insertions of TEs and are extremely lowly transcribed. The fate of these surveillance transcripts could be: the possibility that they locate a *de novo* RdDM locus at a newly activated or inserted TE, which could in turn recruit the RdDM machinery for silencing or, more often than not, the possibility that they do not encounter any aberrant transcript and are most likely directed for degradation by exonucleases. Despite the great improvements in sequencing techniques, we are still unable to sequence lowly transcribed or quickly degraded regions of the genome, which explains our previous inability to identify these surveillance Pol V transcripts. With our new findings, there exists a great number of unanswered questions that would not only help us better understand Pol V transcription but also explain the RdDM mechanism in light of surveillance transcription.

One interesting direction to better understand the expanse of Pol V transcription would be to disrupt RNA degradation and test the accumulation of Pol V transcripts throughout the genome. If Pol V transcribes a large portion of the genome, the surveillance transcripts that are lowly transcribed and not required for RdDM would be marked for degradation by exonucleases present in the cells. Testing the accumulation of Pol V transcripts in exonuclease mutants, which lack the ability to degrade RNA, would give us a better idea of the spread of Pol V transcripts across the genome. This can also shed some light on the loci targeted or preferentially transcribed by Pol V and thereby help us better understand how Pol V functions.

5.2 Pol V surveillance transcription is important for maintaining genomic integrity

Chapter III shows that Pol V transcribes a very broad portion of the genome (42%) and about half of these Pol V transcribed regions do not lead to RdDM. These regions are shown to be transcribed to a much lower extent and possibly less frequently. These non-RdDM Pol V transcripts, or surveillance transcripts, are predicted to scan the genome to identify newly activated/inserted transposons and other aberrant transcripts that are in turn targeted for silencing by RdDM. In the chance that an aberrant transcript is identified, the surveillance transcripts can act as a template to recruit siRNAs produced by the inserted/activated transposons [160] and initiate the RdDM pathway to establish *de novo* DNA methylation and other repressive chromatin modifications to silence these loci. In the alternate scenario that there are no new insertions or disruptions to the genome, the surveillance transcripts are targeted for degradation. This model (Figure 3.9) explains the importance of pervasive Pol V transcription for targeting chromatin modifications to non-conserved regions in the genome and thus, for maintenance of genomic integrity of the organism.

Pervasive transcription of the eukaryotic genome has been considered non-functional due to the absence of sequence conservation in these transcripts [165]. Additionally, the rapid degradation of most of these pervasive non-coding transcripts has been believed to be the lack of observable function of these transcripts[206, 207]. Thus, pervasively transcribed RNA have been believed to be noisy transcripts that the cell tries to dampen[206, 207]. However, it has been argued that many of the non-coding transcription primarily provides a cache of RNA molecules that can eventually evolve useful functions[208]. Our model depicts one such function for pervasive transcription where these transcripts are produced throughout the genome as a way to scan the genome for the presence of spurious transcription from newly integrated TEs. These transcripts act as a scaffold for recruiting chromatin modifying machinery to the regions if a newly incorporated TE is encountered. Previous studies have shown that siRNA-AGO4 complexes that target RdDM to genomic loci are not capable of recognizing or binding to their target loci in the absence of lncRNAs produced by Pol V[88, 155]. This greatly hints at the importance of these surveillance Pol V transcripts in identifying novel TE insertions that might occur randomly throughout the genome and make it competent for immediate silencing by recruitment of the RdDM machinery. Our model also suggests that if the surveillance transcripts do not identify a novel TE, it would not recruit the complementary siRNA-AGO complex and thereby end up being quickly degraded by the cell, thus depicting the short-lived and lowly transcribed nature of all pervasive transcripts. Thus, our model suggests a possible function for pervasive transcription in maintaining the genomic integrity by scanning the genome for possible aberrant transcripts and targeting them for silencing almost immediately.

Despite providing a possible explanation of the need for pervasive transcription in organisms, our model does not explain the mechanism of transition of Pol V from a low transcription, surveillance state to a high transcription, RdDM state, once a novel TE insertion has been identified. In the event that a novel TE insertion has been found, a series of proteins would need to be recruited leading to the establishment of repressive chromatin modifications as in RdDM. However, this would need to be followed by switching of Pol V transcription from a low-level surveillance state to the higher-level, maintenance state of transcription. Identifying the proteins and components that bring about this transition would be a very interesting next step for this study. Another possible method to validate the Pol V transition would be by introducing a new transposon by CRISPR targeted to a surveillance locus and tracking the increase in Pol V transcript accumulation with the new insertion. This would provide a subsequent locus-specific validation of the model described in Chapter III.

5.3 lncRNA determines heterochromatin boundaries

RdDM has been shown to be involved in many biological processes [55], however its primary function has always been considered to be transposon silencing. Despite knowing that RdDM can silence transposons by directing *de novo* DNA methylation to transposons, the mechanism bringing about the silencing has been unclear. Transposon silencing occurs through a combination of pathways, three of which have been studied in great depth. RdDM is the only *de novo* silencing pathway that has been elucidated. The other two pathways are maintenance pathways that work through MET1 based CG methylation[114] and CMT2 and CMT3 based CHG and CHH methylation[24, 209]. Chapter II shows the enrichment of Pol V transcription and Pol V-dependent CHH methylation at the edges of transposon. Another important conclusion from this study is that Pol V preferentially transcribes into transposons from both ends of the transposons. This led to the hypothesis that Pol V plays an important role in determining heterochromatin boundaries which could be important to prevent the spread of heterochromatin. The role of RdDM in the determination of heterochromatin boundaries has been previously studied in maize, where mutations in RdDM components depicted the spreading of euchromatin from genes into nearby transposons[144]. From this, it appears that our data also hints at the possibility of Pol V and RdDM playing a role in determining and restricting heterochromatin–euchromatin boundaries.

These findings raise questions about how Pol V identifies and targets RdDM to edges of heterochromatin. One possibility is that Pol V enrichment at the heterochromatin edges could be the result of preexisting chromatin marks that already outline heterochromatin and recruit Pol V to these loci. Another possibility is that Pol V is required to establish these heterochromatin boundaries by directing repressive chromatin modifications to the edges. This would greatly explain the role of Pol V in maintaining euchromatin–heterochromatin boundaries. Another interesting expansion to this work would be to test the impact of Pol V on Pol II transcription and gene expression regulation. This can be done by testing the effect of Pol V and its downstream repressive chromatin marks on the recruitment of Pol II to the heterochromatin boundaries.

5.4 RNA-directed DNA methylation positions nucleosomes to regulate gene expression

RNA-directed DNA methylation is functionally intertwined with chromatin modifications in the form of DNA methylation and nucleosome positioning. It is known that chromatin modifications play a role both upstream [82, 157] and downstream [24, 112, 42, 55] of recruitment of Pol IV and Pol V to RdDM loci. Chapter IV shows that at RdDM loci, DNA methylation determines nucleosome positioning. The results show that non-coding transcription by Pol V recruits AGO4 and IDN2 proteins. IDN2, in turn, recruits the SWI/SNF nucleosome remodeling complex, which is capable of but not sufficient to alter nucleosome positions. IDN2 also recruits the *de novo* DNA methyltransferase, DRM2, which in turn establishes DNA methylation that are essential to activate chromatin remodelers and alter nucleosome positions. Thus, at RdDM loci, we have identified that nucleosome positions are dependent on DNA methylation and their coordinated effect leads to repression of genes. Our model of DNA methylation directing nucleosomes to regulate genes is depicted in Figure 5.1.

The importance of DNA methylation for nucleosome positioning has a significant impact on our understanding of the RdDM pathway. Our conclusion suggests that the RdDM pathway might not be branched after IDN2 recruitment[42] but on the contrary, these events occurring at Pol V transcribed loci are organized in a stepwise genetic pathway[90]. Although when studied genetically, this pathway appears linear, various steps of the pathway are likely to rely on the cooperative recruitment or activation of subsequent factors. One example of such a connection is the requirement of both SWI3B and DRM2 based DNA methylation to direct nucleosomes. Our model is also consistent with the idea that the

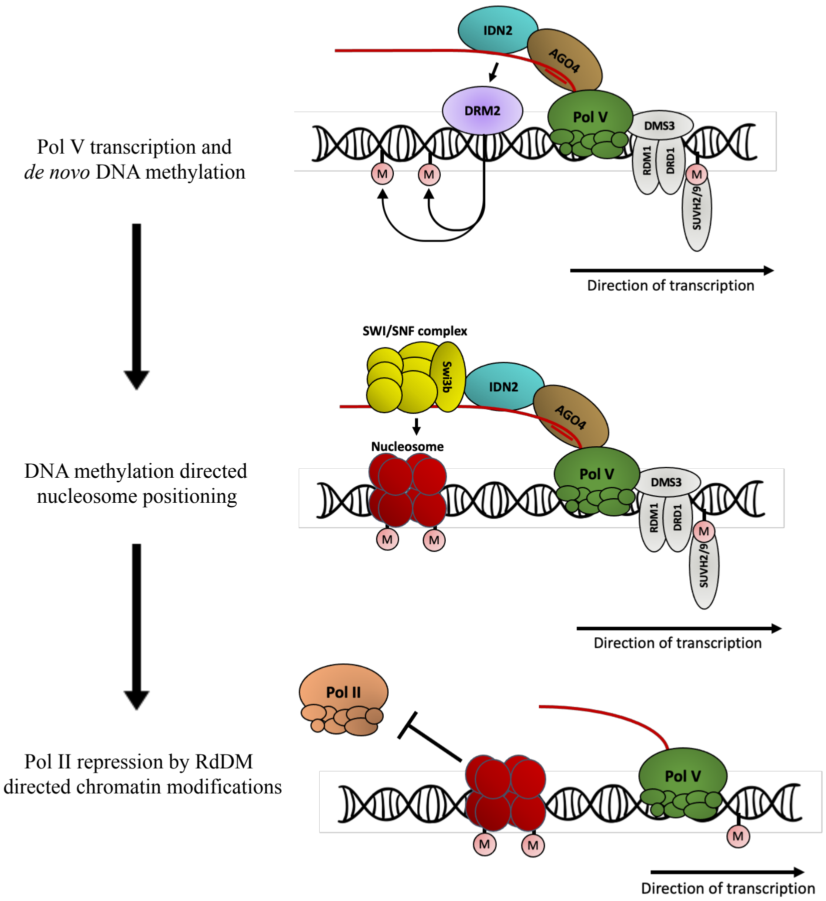

Figure 5.1: **DNA methylation directs nucleosome positioning at RdDM loci**

events in the late stages of RdDM lead to a concerted establishment of repressive chromatin modifications like DNA methylation and nucleosome positioning leading to an efficient repression of Pol II transcription. It is also consistent with the general difficulty to experimentally tease apart various repressive chromatin modifications established by this pathway.

The involvement of SWI/SNF and nucleosome positioning in RdDM may also be considered in context of this pathway mostly involved in maintenance of silencing of TEs. Transcription of heterochromatic regions by Pol IV and Pol V may be preceded by the removal or repositioning of previously positioned nucleosomes. This is supported by the involvement of putative chromatin remodelers in initiation and/or elongation of transcription by both of those polymerases[105, 198, 112, 102]. Nucleosome positioning established as an outcome of RdDM might also be a mechanism to re-create the pattern of nucleosomes disrupted by Pol IV and Pol V transcription. However, at *de novo* RdDM loci around newly inserted TEs, we do not expect to see preexisting repressive chromatin modifications. The role of nucleosome positioning in this *de novo* process remains unexplored.

The involvement of DNA methylation in determining the pattern of nucleosomes is not restricted to RdDM targets. The impact of MET1-dependent CG methylation and CMT3-dependent CHG methylation on stabilizing nucleosomes suggests that DNA methylation may affect nucleosome patterns throughout the genome. This has also been observed in other eukaryotes[192, 199]. This general effect of DNA methylation on nucleosome positioning would counteract the general preference to methylate linkers and contribute to local correlations between nucleosomes and DNA methylation. This property of nucleosomes is consistent with previous reports[188] and may involve the activity of DDM1[190]. It illustrates the general interdependence between nucleosomes and DNA methylation.

Existing evidence, however, does not support the view that DNA methylation is the primary determinant of the nucleosome pattern. Nucleosome positions can be controlled by a combination of intrinsic factors and active chromatin remodeling. The role of DNA methylation is more limited and probabilistic. We believe that global losses of DNA methylation in RdDM and DNA methyltransferase mutants may affect the patterns of nucleosomes by a combination of cis- and trans-acting factors. This question could be tested by targeting or depleting DNA methylation to/from specific loci in the genome and studying the effect it has on the nucleosome positions around the region, rendering a more targeted locus-specific approach to address this question.

5.5 Implications of new findings

RNA-directed DNA methylation is the transcriptional gene silencing pathway in plants that utilizes ncRNA to direct repressive chromatin modifications like *de novo* DNA methylation and nucleosomes, to specific target loci for silencing. Target loci of RdDM have been known to include transposons and repeats distributed throughout the genome with its presence particularly notable at smaller and younger transposons and repeats in euchromatic chromosome arms[86, 209, 119] Over time, we have understood that RdDM mechanism greatly depends on Pol IV and Pol V transcription of siRNA and lncRNA, where lncRNA acts as a scaffold to recruit a series of proteins and chromatin remodelers to eventually direct repressive chromatin modifications to the loci[55]. However, there remain a lot of specific unanswered questions that are essential for a better understanding of the RdDM pathway, which would not just help understand this process of gene regulation in plants but can be expanded to explain the mechanism of transcriptional gene silencing in other organisms.

This thesis addresses some of these unanswered mechanistic questions of the TGS pathway by trying to understand the role that lncRNA and Pol V play in directing and targeting repressive chromatin marks and thereby regulating genes. This study tries to understand Pol V transcription in relation to the RdDM pathway at three different stages:

A) Upstream of RdDM pathway or the pre-transcription stage.

B) The RdDM pathway or transcription stage.

C) Downstream of RdDM pathway or the post-transcription stage.

Upstream of RdDM, there still exist unanswered questions about how Pol V identifies and transcribes its target loci in the genome. Prior to this study, the only knowledge we had about Pol V recruitment was the existence of a self-reinforcing loop, wherein DNA methylation established by RdDM is read by SUVH2 and SUVH9 proteins, which in turn have been shown to be important for Pol V recruitment to the same loci [58, 86, 100, 101]. This mechanism, however, was only able to explain the recruitment of Pol V to loci that have already been targeted for silencing by RdDM but still didn't explain how *de novo* loci lacking chromatin modifications were identified by Pol V. This thesis has shown that Pol V is not necessarily recruited to its targets but instead is pervasively transcribing the genome, scanning large portions of it. Thus, at locations where *de novo* TE expression occurs, Pol V could already be present. This surveillance mechanism of transcription explains how Pol V is able to find and quickly silence newly activated transposons through RdDM.

In the RdDM pathway, studies have already shown a list of proteins recruited to specific target loci silenced by RdDM. At these specific RdDM targets, it is known that Pol V transcribes lncRNA that acts as a scaffold to recruit 24-nt siRNA bound AGO4 protein followed by recruitment of IDN2 and eventually the *de novo* DNA methyltransferase, DRM2, which directs DNA methylation to the target loci to silence them [87, 88, 89, 90, 91]. Most of this information was obtained from assays at specific loci previously identified as RdDM but not on a genome-wide scale[88, 57, 42]. This thesis has been successful in identifying genome-wide RdDM loci. Identification of these new sites has made it possible to expand our knowledge of RdDM and mechanism of Pol V transcription at these loci. The mechanism of Pol V transcription has also been better explained by this study depicting that Pol V transcription is not strand-specific and it transcribes into TEs from both sides, thereby marking the boundaries of heterochromatin. Thus, with the expansion of our knowledge of Pol V transcribed loci, this thesis has paved way to a better understanding of RdDM.

As for the downstream effects of RdDM, it has been known that RdDM targets DNA methylation to the Pol V transcribed regions that need to be silenced[55, 120]. It has also been observed that Pol V transcription leads to the recruitment of the SWI/SNF remodeling complex that could also alter nucleosome positions as a result of Pol V transcription[42]. However, what was not completely known was how these different repressive chromatin marks targeted by RdDM interact with each other. This thesis has addressed and found some answers to the interaction between DNA methylation and nucleosome positioning caused by RdDM. It has shown that RdDM directs DNA methylation which in turn positions nucleosomes at the methylated loci.

This thesis has addressed many important questions about Pol V transcription and lncRNAs and their involvement in RdDM. The studies have expanded our knowledge about lncRNAs and their mechanism of directing repressive modifications to their target loci. This information about the role of ncRNAs in TGS can be expanded to other organisms to better understand the TGS pathways in these organisms too.

In addition to the TGS pathway, this thesis has also significantly improved our understanding on the role and need of pervasive transcription in plants, and possibly all organisms. Pervasive transcripts have been believed to be noisy transcripts that are non-functional, non-conserved and targeted for rapid degradation by the cell[206, 207]. This thesis has provided a possible function for these pervasive transcripts. This study has shed light onto the possibility that pervasive transcripts are essential for maintaining the integrity of the genome. Pervasive transcripts are produced to survey the entire genome for possible random integrations, on account of which these transcripts act as a scaffold to recruit chromatin modifying machinery to silence the random insertion. On the other hand, in the absence of a random insertion, these pervasive transcripts are targeted for degradation.

Thus, this study shows that pervasive transcription might be essential to identify novel TE insertions that might occur randomly throughout the genome and it makes the genome competent for immediate silencing, thereby playing a very important role in maintaining the integrity of the genome[177] (Chapter III). Once this random insertion is identified, the RdDM machinery is quickly recruited to target DNA methylation to the target loci[152] (Chapter II), which in turn recruits and positions nucleosomes through the SWI/SNF complex[205] (Chapter IV).

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Eric S. Lander, Lauren M. Linton, Bruce Birren, Chad Nusbaum, Michael C. Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, Roel Funke, Diane Gage, Katrina Harris, Andrew Heaford, John Howland, Lisa Kann, Jessica Lehoczky, Rosie LeVine, Paul McEwan, Kevin McKernan, James Meldrim, Jill P. Mesirov, Cher Miranda, William Morris, Jerome Naylor, Christina Raymond, Mark Rosetti, Ralph Santos, Andrew Sheridan, Carrie Sougnez, Nicole Stange-Thomann, Nikola Stojanovic, Aravind Subramanian, Dudley Wyman, Jane Rogers, John Sulston, Rachael Ainscough, Stephan Beck, David Bentley, John Burton, Christopher Clee, Nigel Carter, Alan Coulson, Rebecca Deadman, Panos Deloukas, Andrew Dunham, Ian Dunham, Richard Durbin, Lisa French, Darren Grafham, Simon Gregory, Tim Hubbard, Sean Humphray, Adrienne Hunt, Matthew Jones, Christine Lloyd, Amanda McMurray, Lucy Matthews, Simon Mercer, Sarah Milne, James C. Mullikin, Andrew Mungall, Robert Plumb, Mark Ross, Ratna Shownkeen, Sarah Sims, Robert H. Waterston, Richard K. Wilson, LaDeana W. Hillier, John D. McPherson, Marco A. Marra, Elaine R. Mardis, Lucinda A. Fulton, Asif T. Chinwalla, Kymberlie H. Pepin, Warren R. Gish, Stephanie L. Chissoe, Michael C. Wendl, Kim D. Delehaunty, Tracie L. Miner, Andrew Delehaunty, Jason B. Kramer, Lisa L. Cook, Robert S. Fulton, Douglas L. Johnson, Patrick J. Minx, Sandra W. Clifton, Trevor Hawkins, Elbert Branscomb, Paul Predki, Paul Richardson, Sarah Wenning, Tom Slezak, Norman Doggett, Jan-Fang Cheng, Anne Olsen, Susan Lucas, Christopher Elkin, Edward Uberbacher, Marvin Frazier, Richard A. Gibbs, Donna M. Muzny, Steven E. Scherer, John B. Bouck, Erica J. Sodergren, Kim C. Worley, Catherine M. Rives, James H. Gorrell, Michael L. Metzker, Susan L. Naylor, Raju S. Kucherlapati, David L. Nelson, George M. Weinstock, Yoshiyuki Sakaki, Asao Fujiyama, Masahira Hattori, Tetsushi Yada, Atsushi Toyoda, Takehiko Itoh, Chiharu Kawagoe, Hidemi Watanabe, Yasushi Totoki, Todd Taylor, Jean Weissenbach, Roland Heilig, William Saurin, Francois Artiguenave, Philippe Brottier, Thomas Bruls, Eric Pelletier, Catherine Robert, Patrick Wincker, André Rosenthal, Matthias Platzer, Gerald Nyakatura, Stefan Taudien, Andreas Rump, Douglas R. Smith, Lynn Doucette-Stamm, Marc Rubenfield, Keith Weinstock, Hong Mei Lee, JoAnn Dubois, Huanming Yang, Jun Yu, Jian Wang, Guyang Huang, Jun Gu, Leroy Hood, Lee Rowen, Anup Madan, Shizen Qin, Ronald W. Davis, Nancy A. Federspiel, A. Pia Abola, Michael J. Proctor, Bruce A. Roe, Feng Chen, Huaqin Pan, Juliane Ramser, Hans Lehrach, Richard Reinhardt, W. Richard McCombie, Melissa de la Bastide, Neilay Dedhia, Helmut Blöcker, Klaus Hornischer, Gabriele Nordsiek, Richa Agarwala, L. Aravind, Jeffrey A. Bailey, Alex Bateman, Serafim Batzoglou,

Ewan Birney, Peer Bork, Daniel G. Brown, Christopher B. Burge, Lorenzo Cerutti, Hsiu-Chuan Chen, Deanna Church, Michele Clamp, Richard R. Copley, Tobias Doerks, Sean R. Eddy, Evan E. Eichler, Terrence S. Furey, James Galagan, James G. R. Gilbert, Cyrus Harmon, Yoshihide Hayashizaki, David Haussler, Henning Hermjakob, Karsten Hokamp, Wonhee Jang, L. Steven Johnson, Thomas A. Jones, Simon Kasif, Arek Kaspryzk, Scot Kennedy, W. James Kent, Paul Kitts, Eugene V. Koonin, Ian Korf, David Kulp, Doron Lancet, Todd M. Lowe, Aoife McLysaght, Tarjei Mikkelsen, John V. Moran, Nicola Mulder, Victor J. Pollara, Chris P. Ponting, Greg Schuler, Jörg Schultz, Guy Slater, Arian F. A. Smit, Elia Stupka, Joseph Szustakowki, Danielle Thierry-Mieg, Jean Thierry-Mieg, Lukas Wagner, John Wallis, Raymond Wheeler, Alan Williams, Yuri I. Wolf, Kenneth H. Wolfe, Shiaw-Pyng Yang, Ru-Fang Yeh, Francis Collins, Mark S. Guyer, Jane Peterson, Adam Felsenfeld, Kris A. Wetterstrand, Richard M. Myers, Jeremy Schmutz, Mark Dickson, Jane Grimwood, David R. Cox, Maynard V. Olson, Rajinder Kaul, Christopher Raymond, Nobuyoshi Shimizu, Kazuhiko Kawasaki, Shinsei Minoshima, Glen A. Evans, Maria Athanasiou, Roger Schultz, Aristides Patrinos, Michael J. Morgan, International Human Genome Sequencing Consortium, Center for Genome Research: Whitehead Institute for Biomedical Research, The Sanger Centre:, Washington University Genome Sequencing Center, US DOE Joint Genome Institute:, Baylor College of Medicine Human Genome Sequencing Center:, RIKEN Genomic Sciences Center:, Genoscope and CNRS UMR-8030:, Institute of Molecular Biotechnology: Department of Genome Analysis, GTC Sequencing Center:, Beijing Genomics Institute/Human Genome Center:, The Institute for Systems Biology: Multimegabase Sequencing Center, Stanford Genome Technology Center:, University of Oklahoma's Advanced Center for Genome Technology:, Max Planck Institute for Molecular Genetics:, Lita Annenberg Hazen Genome Center: Cold Spring Harbor Laboratory, GBF—German Research Centre for Biotechnology:, also includes individuals listed under other headings): *Genome Analysis Group (listed in alphabetical order, US National Institutes of Health: Scientific management: National Human Genome Research Institute, Stanford Human Genome Center:, University of Washington Genome Center:, Keio University School of Medicine: Department of Molecular Biology, University of Texas Southwestern Medical Center at Dallas:, US Department of Energy: Office of Science, and The Wellcome Trust:. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, February 2001. Number: 6822 Publisher: Nature Publishing Group.

[2] Elizabeth Pennisi. ENCODE Project Writes Eulogy for Junk DNA. *Science*, 337(6099):1159–1161, September 2012. Publisher: American Association for the Advancement of Science Section: News &amp; Analysis.

[3] Sarah Geisler and Jeff Coller. RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. *Nature Reviews Molecular Cell Biology*, 14(11):699–712, November 2013. Number: 11 Publisher: Nature Publishing Group.

[4] Jean-Jack M. Riethoven. Regulatory regions in DNA: promoters, enhancers, si-

lencers, and insulators. *Methods in Molecular Biology (Clifton, N.J.)*, 674:33–42, 2010.

[5] Glenn A. Maston, Sara K. Evans, and Michael R. Green. Transcriptional Regulatory Elements in the Human Genome. *Annual Review of Genomics and Human Genetics*, 7(1):29–59, September 2006. Publisher: Annual Reviews.

[6] Len A. Pennacchio, Wendy Bickmore, Ann Dean, Marcelo A. Nobrega, and Gill Bejerano. Enhancers: five essential questions. *Nature Reviews Genetics*, 14(4):288–295, April 2013. Number: 4 Publisher: Nature Publishing Group.

[7] Petros Kolovos, Tobias A. Knoch, Frank G. Grosveld, Peter R. Cook, and Argyris Papantonis. Enhancers and silencers: an integrated and simple model for their function. *Epigenetics & Chromatin*, 5(1):1, January 2012.

[8] Jan Bednar, Rachel A. Horowitz, Sergei A. Grigoryev, Lenny M. Carruthers, Jeffrey C. Hansen, Abraham J. Koster, and Christopher L. Woodcock. Nucleosomes, linker DNA, and linker histone form a unique structural motif that directs the higher-order folding and compaction of chromatin. *Proceedings of the National Academy of Sciences*, 95(24):14173–14178, November 1998. Publisher: National Academy of Sciences Section: Biological Sciences.

[9] Timothy J. Richmond and Curt A. Davey. The structure of DNA in the nucleosome core. *Nature*, 423(6936):145–150, May 2003.

[10] Cizhong Jiang and B. Franklin Pugh. Nucleosome positioning and gene regulation: advances through genomics. *Nature Reviews Genetics*, 10(3):161–172, March 2009. Number: 3 Publisher: Nature Publishing Group.

[11] B. Wasylyk and P. Chambon. Transcription by eukaryotic RNA polymerases A and B of chromatin assembled in vitro. *European Journal of Biochemistry*, 98(2):317–327, August 1979.

[12] D. Y. Lee, J. J. Hayes, D. Pruss, and A. P. Wolffe. A positive role for histone acetylation in transcription factor access to nucleosomal DNA. *Cell*, 72(1):73–84, January 1993.

[13] Kaifu Chen, Qingshu Meng, Lina Ma, Qingyou Liu, Petrus Tang, Chungshung Chiu, Songnian Hu, and Jun Yu. A novel DNA sequence periodicity decodes nucleosome positioning. *Nucleic Acids Research*, 36(19):6228–6236, November 2008.

[14] Craig L Peterson and Marc-André Laniel. Histones and histone modifications. *Current Biology*, 14(14):R546–R551, July 2004.

[15] Carsten Carlberg and Ferdinand Molnár. Chromatin Modifiers. In Carsten Carlberg and Ferdinand Molnár, editors, *Mechanisms of Gene Regulation*, pages 125–141. Springer Netherlands, Dordrecht, 2014.

[16] Taiping Chen and Sharon Y. R. Dent. Chromatin modifiers and remodellers: regulators of cellular differentiation. *Nature Reviews Genetics*, 15(2):93–106, February 2014. Number: 2 Publisher: Nature Publishing Group.

[17] Tianyi Zhang, Sarah Cooper, and Neil Brockdorff. The interplay of histone modifications – writers that read. *EMBO Reports*, 16(11):1467–1481, November 2015.

[18] Jeffrey A. Rosenfeld, Zhibin Wang, Dustin E. Schones, Keji Zhao, Rob DeSalle, and Michael Q. Zhang. Determination of enriched histone modifications in non-genic portions of the human genome. *BMC Genomics*, 10(1):143, March 2009.

[19] Artem Barski, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Dustin E. Schones, Zhibin Wang, Gang Wei, Iouri Chepelev, and Keji Zhao. High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell*, 129(4):823–837, May 2007.

[20] Julie A. Law and Steven E. Jacobsen. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature Reviews Genetics*, 11(3):204–220, March 2010. Number: 3 Publisher: Nature Publishing Group.

[21] Zachary D. Smith and Alexander Meissner. DNA methylation: roles in mammalian development. *Nature Reviews Genetics*, 14(3):204–220, March 2013. Number: 3 Publisher: Nature Publishing Group.

[22] John R. Edwards, Olya Yarychkivska, Mathieu Boulard, and Timothy H. Bestor. DNA methylation and DNA methyltransferases. *Epigenetics & Chromatin*, 10, May 2017.

[23] Olga V. Dyachenko, Tara V. Schevchuk, Leo Kretzner, Yaroslav I. Buryanov, and Steven S. Smith. Human non-CG methylation. *Epigenetics*, 5(7):569–572, October 2010. Publisher: Taylor & Francis _eprint: https://doi.org/10.4161/epi.5.7.12702.

[24] Hume Stroud, Truman Do, Jiamu Du, Xuehua Zhong, Suhua Feng, Lianna Johnson, Dinshaw J. Patel, and Steven E. Jacobsen. The roles of non-CG methylation in Arabidopsis. *Nature structural & molecular biology*, 21(1):64–72, January 2014.

[25] Lena Ho and Gerald R. Crabtree. Chromatin remodelling during development. *Nature*, 463(7280):474–484, January 2010. Number: 7280 Publisher: Nature Publishing Group.

[26] Jiamu Du, Lianna M. Johnson, Steven E. Jacobsen, and Dinshaw J. Patel. DNA methylation pathways and their crosstalk with histone methylation. *Nature Reviews. Molecular Cell Biology*, 16(9):519–532, September 2015.

[27] Howard Cedar and Yehudit Bergman. Linking DNA methylation and histone modification: patterns and paradigms. *Nature Reviews Genetics*, 10(5):295–304, May 2009. Number: 5 Publisher: Nature Publishing Group.

[28] Steen K. T. Ooi, Chen Qiu, Emily Bernstein, Keqin Li, Da Jia, Zhe Yang, Hediye Erdjument-Bromage, Paul Tempst, Shau-Ping Lin, C. David Allis, Xiaodong Cheng, and Timothy H. Bestor. DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. *Nature*, 448(7154):714–717, August 2007.

[29] Bernhard Lehnertz, Yoshihide Ueda, Alwin A. H. A. Derijck, Ulrich Braunschweig, Laura Perez-Burgos, Stefan Kubicek, Taiping Chen, En Li, Thomas Jenuwein, and Antoine H. F. M. Peters. Suv39h-Mediated Histone H3 Lysine 9 Methylation Directs DNA Methylation to Major Satellite Repeats at Pericentric Heterochromatin. *Current Biology*, 13(14):1192–1200, July 2003.

[30] H. Tamaru and E. U. Selker. A histone H3 methyltransferase controls DNA methylation in Neurospora crassa. *Nature*, 414(6861):277–283, November 2001.

[31] James P. Jackson, Anders M. Lindroth, Xiaofeng Cao, and Steven E. Jacobsen. Control of CpNpG DNA methylation by the KRYPTONITE histone H3 methyltransferase. *Nature*, 416(6880):556–560, April 2002. Number: 6880 Publisher: Nature Publishing Group.

[32] Olivier Mathieu, Aline V Probst, and Jerzy Paszkowski. Distinct regulation of histone H3 methylation at lysines 27 and 9 by CpG methylation in Arabidopsis. *The EMBO Journal*, 24(15):2783–2791, August 2005.

[33] Cedric R. Clapier, Janet Iwasa, Bradley R. Cairns, and Craig L. Peterson. Mechanisms of action and regulation of ATP-dependent chromatin-remodelling complexes. *Nature Reviews Molecular Cell Biology*, 18(7):407–422, July 2017. Number: 7 Publisher: Nature Publishing Group.

[34] Geeta J. Narlikar, Ramasubramanian Sundaramoorthy, and Tom Owen-Hughes. Mechanisms and Functions of ATP-Dependent Chromatin-Remodeling Enzymes. *Cell*, 154(3):490–503, August 2013.

[35] Wen Du, Daimo Guo, and Wei Du. ATP-Dependent Chromatin Remodeling Complex in the Lineage Specification of Mesenchymal Stem Cells, September 2020. ISSN: 1687-966X Pages: e8839703 Publisher: Hindawi Volume: 2020.

[36] Ana Belén Sanz, Raúl García, José Manuel Rodríguez-Peña, César Nombela, and Javier Arroyo. Cooperation between SAGA and SWI/SNF complexes is required for efficient transcriptional responses regulated by the yeast MAPK Slt2. *Nucleic Acids Research*, 44(15):7159–7172, September 2016. Publisher: Oxford Academic.

[37] H. Steven Zhang and Douglas C. Dean. Rb-mediated chromatin structure regulation and transcriptional repression. *Oncogene*, 20(24):3134–3138, May 2001. Number: 24 Publisher: Nature Publishing Group.

[38] Kathleen Dennis, Tao Fan, Theresa Geiman, Qingsheng Yan, and Kathrin Muegge. Lsh, a member of the SNF2 family, is required for genome-wide methylation. *Genes*

& Development, 15(22):2940–2944, November 2001. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[39] Jeffrey A. Jeddeloh, Trevor L. Stokes, and Eric J. Richards. Maintenance of genomic methylation requires a SWI2/SNF2-like protein. *Nature Genetics*, 22(1):94–97, May 1999. Number: 1 Publisher: Nature Publishing Group.

[40] Richard J. Gibbons, Tarra L. McDowell, Sundhya Raman, Delia M. O'Rourke, David Garrick, Helena Ayyub, and Douglas R. Higgs. Mutations in ATRX , encoding a SWI/SNF-like protein, cause diverse changes in the pattern of DNA methylation. *Nature Genetics*, 24(4):368–371, April 2000. Number: 4 Publisher: Nature Publishing Group.

[41] Fatima Banine, Christopher Bartlett, Ranjaka Gunawardena, Christian Muchardt, Moshe Yaniv, Erik S. Knudsen, Bernard E. Weissman, and Larry S. Sherman. SWI/SNF chromatin-remodeling factors induce changes in DNA methylation to promote transcriptional activation. *Cancer Research*, 65(9):3542–3547, May 2005.

[42] Yongyou Zhu, M. Jordan Rowley, Gudrun Böhmdorfer, and Andrzej T. Wierzbicki. A SWI/SNF chromatin-remodeling complex acts in noncoding RNA-mediated transcriptional silencing. *Molecular Cell*, 49(2):298–309, January 2013.

[43] Mihaela Pertea. The Human Transcriptome: An Unfinished Story. *Genes*, 3(3):344–360, June 2012.

[44] Zhipeng Qu and David L. Adelson. Evolutionary conservation and functional roles of ncRNA. *Frontiers in Genetics*, 3, October 2012.

[45] John S. Mattick and Igor V. Makunin. Small regulatory RNAs in mammals. *Human Molecular Genetics*, 14(suppl_1):R121–R132, April 2005. Publisher: Oxford Academic.

[46] Cheng Lu, Shivakundan Singh Tej, Shujun Luo, Christian D. Haudenschild, Blake C. Meyers, and Pamela J. Green. Elucidation of the Small RNA Component of the Transcriptome. *Science*, 309(5740):1567–1569, September 2005. Publisher: American Association for the Advancement of Science Section: Report.

[47] David P. Bartel. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, 116(2):281–297, January 2004.

[48] Bryan R. Cullen. Derivation and function of small interfering RNAs and microRNAs. *Virus Research*, 102(1):3–9, June 2004.

[49] Marc Robert Fabian, Nahum Sonenberg, and Witold Filipowicz. Regulation of mRNA Translation and Stability by microRNAs. *Annual Review of Biochemistry*, 79(1):351–379, June 2010. Publisher: Annual Reviews.

[50] Di Zhang, Juan Zhou, Jie Gao, Ri-Ying Wu, Ying-Long Huang, Qin-Wen Jin, Jian-Si Chen, Wei-Zhong Tang, and Lin-Hai Yan. Targeting snoRNAs as an emerging method of therapeutic development for cancer. *American Journal of Cancer Research*, 9(8):1504–1516, August 2019.

[51] Andrew Hamilton, Olivier Voinnet, Louise Chappell, and David Baulcombe. Two classes of short interfering RNA in RNA silencing. *The EMBO journal*, 21(17):4671–4679, September 2002.

[52] Cesar Llave, Kristin D. Kasschau, Maggie A. Rector, and James C. Carrington. Endogenous and Silencing-Associated Small RNAs in Plants. *The Plant Cell*, 14(7):1605–1619, July 2002.

[53] Liling Tang, Eva Nogales, and Claudio Ciferri. Structure and function of SWI/SNF chromatin remodeling complexes and mechanistic implications for transcription. *Progress in Biophysics and Molecular Biology*, 102(2-3):122–128, July 2010.

[54] ZHENGYING HE and ERIK J. SONTHEIMER. “siRNAs and miRNAs”: A meeting report on RNA silencing. *RNA*, 10(8):1165–1173, August 2004.

[55] Marjori A. Matzke and Rebecca A. Mosher. RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. *Nature Reviews. Genetics*, 15(6):394–408, June 2014.

[56] Andrzej T. Wierzbicki, Ross Cocklin, Anoop Mayampurath, Ryan Lister, M. Jordan Rowley, Brian D. Gregory, Joseph R. Ecker, Haixu Tang, and Craig S. Pikaard. 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, August 2012.

[57] Qi Zheng, M. Jordan Rowley, Gudrun Böhmdorfer, Davinder Sandhu, Brian D. Gregory, and Andrzej T. Wierzbicki. RNA polymerase V targets transcriptional silencing components to promoters of protein-coding genes. *The Plant journal : for cell and molecular biology*, 73(2):179–189, January 2013.

[58] Julie A. Law, Israel Ausin, Lianna M. Johnson, Ajay A. Vashisht, Jian-Kang Zhu, James A. Wohlschlegel, and Steven E. Jacobsen. A protein complex required for polymerase V transcripts and RNA- directed DNA methylation in Arabidopsis. *Current biology: CB*, 20(10):951–956, May 2010.

[59] Danesh Moazed. Small RNAs in transcriptional gene silencing and genome defence. *Nature*, 457(7228):413–420, January 2009. Number: 7228 Publisher: Nature Publishing Group.

[60] Javier Martinez, Agnieszka Patkaniowska, Henning Urlaub, Reinhard Lührmann, and Thomas Tuschl. Single-Stranded Antisense siRNAs Guide Target RNA Cleavage in RNAi. *Cell*, 110(5):563–574, September 2002.

[61] Zhixin Xie, Lisa K. Johansen, Adam M. Gustafson, Kristin D. Kasschau, Andrew D. Lellis, Daniel Zilberman, Steven E. Jacobsen, and James C. Carrington. Genetic and Functional Diversification of Small RNA Pathways in Plants. *PLOS Biology*, 2(5):e104, February 2004. Publisher: Public Library of Science.

[62] Hsiao-Lin V. Wang and Julia A. Chekanova. Long Noncoding RNAs in Plants. *Advances in Experimental Medicine and Biology*, 1008:133–154, 2017.

[63] Zhijin Li, Weiling Zhao, Maode Wang, and Xiaobo Zhou. The Role of Long Noncoding RNAs in Gene Expression Regulation. *Gene Expression Profiling in Cancer*, January 2019. Publisher: IntechOpen.

[64] Tim R. Mercer and John S. Mattick. Structure and function of long noncoding RNAs in epigenetic regulation. *Nature Structural & Molecular Biology*, 20(3):300–307, March 2013.

[65] Yicheng Long, Xueyin Wang, Daniel T. Youmans, and Thomas R. Cech. How do lncRNAs regulate transcription? *Science Advances*, 3(9), September 2017.

[66] Jing-Wen Shih, Wei-Fan Chiang, Alexander T. H. Wu, Ming-Heng Wu, Ling-Yu Wang, Yen-Ling Yu, Yu-Wen Hung, Wen-Chang Wang, Cheng-Ying Chu, Chiu-Lien Hung, Chun A. Changou, Yun Yen, and Hsing-Jien Kung. Long noncoding RNA LncHIFCAR/MIR31HG is a HIF-1α co-activator driving oral cancer progression. *Nature Communications*, 8(1):15874, June 2017. Number: 1 Publisher: Nature Publishing Group.

[67] Yifei Miao, Nassim E. Ajami, Tse-Shun Huang, Feng-Mao Lin, Chih-Hong Lou, Yun-Ting Wang, Shuai Li, Jian Kang, Hannah Munkacsi, Mano R. Maurya, Shakti Gupta, Shu Chien, Shankar Subramaniam, and Zhen Chen. Enhancer-associated long non-coding RNA LEENE regulates endothelial nitric oxide synthase and endothelial function. *Nature Communications*, 9(1):292, 2018.

[68] Seung Woo Cho, Jin Xu, Ruping Sun, Maxwell R. Mumbach, Ava C. Carter, Y. Grace Chen, Kathryn E. Yost, Jeewon Kim, Jing He, Stephanie A. Nevins, Suet-Feung Chin, Carlos Caldas, S. John Liu, Max A. Horlbeck, Daniel A. Lim, Jonathan S. Weissman, Christina Curtis, and Howard Y. Chang. Promoter of lncRNA Gene PVT1 Is a Tumor-Suppressor DNA Boundary Element. *Cell*, 173(6):1398–1412.e22, 2018.

[69] John L. Rinn, Michael Kertesz, Jordon K. Wang, Sharon L. Squazzo, Xiao Xu, Samantha A. Brugmann, L. Henry Goodnough, Jill A. Helms, Peggy J. Farnham, Eran Segal, and Howard Y. Chang. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell*, 129(7):1311–1323, June 2007.

[70] John E. Froberg, Lin Yang, and Jeannie T. Lee. Guided by RNAs: X-inactivation as a model for lncRNA function. *Journal of Molecular Biology*, 425(19):3698–3706, October 2013.

[71] Emily Maclary, Michael Hinten, Clair Harris, Shriya Sethuraman, Srimonta Gayen, and Sundeep Kalantry. PRC2 represses transcribed genes on the imprinted inactive X chromosome in mice. *Genome Biology*, 18(1):82, May 2017.

[72] Hervé Vaucheret and Mathilde Fagard. Transcriptional gene silencing in plants: targets, inducers and regulators. *Trends in Genetics*, 17(1):29–35, January 2001. Publisher: Elsevier.

[73] Marc S. Weinberg and Kevin V. Morris. Transcriptional gene silencing in humans. *Nucleic Acids Research*, 44(14):6505–6517, August 2016.

[74] Andrzej T Wierzbicki. The role of long non-coding RNA in transcriptional gene silencing. *Current Opinion in Plant Biology*, 15(5):517–522, November 2012.

[75] Thomas S. Ream, Jeremy R. Haag, Andrzej T. Wierzbicki, Carrie D. Nicora, Angela D. Norbeck, Jian-Kang Zhu, Gretchen Hagen, Thomas J. Guilfoyle, Ljiljana Pasa-Tolić, and Craig S. Pikaard. 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, January 2009.

[76] Jeremy R. Haag and Craig S. Pikaard. Multisubunit RNA polymerases IV and V: purveyors of non-coding RNA for plant gene silencing. *Nature Reviews. Molecular Cell Biology*, 12(8):483–492, July 2011.

[77] Ming Zhou and Julie A. Law. RNA Pol IV and V in Gene Silencing: Rebel Polymerases Evolving Away From Pol II's Rules. *Current opinion in plant biology*, 27:154–164, October 2015.

[78] Tatsuo Kanno, Bruno Huettel, M. Florian Mette, Werner Aufsatz, Estelle Jaligot, Lucia Daxinger, David P. Kreil, Marjori Matzke, and Antonius J. M. Matzke. Atypical RNA polymerase subunits required for RNA-directed DNA methylation. *Nature Genetics*, 37(7):761–765, July 2005.

[79] Dominique Pontier, Galina Yahubyan, Danielle Vega, Agnès Bulski, Julio Saez-Vasquez, Mohamed-Ali Hakimi, Silva Lerbs-Mache, Vincent Colot, and Thierry Lagrange. 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, September 2005.

[80] Julie A. Law, Ajay A. Vashisht, James A. Wohlschlegel, and Steven E. Jacobsen. 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), July 2011.

[81] Todd Blevins, Ram Podicheti, Vibhor Mishra, Michelle Marasco, Jing Wang, Doug Rusch, Haixu Tang, and Craig S. Pikaard. Identification of Pol IV and RDR2-dependent precursors of 24 nt siRNAs guiding de novo DNA methylation in Arabidopsis. *eLife*, 4, 2015. Publisher: eLife Sciences Publications, Ltd.

[82] Julie A. Law, Jiamu Du, Christopher J. Hale, Suhua Feng, Krzysztof Krajewski, Ana Marie S. Palanca, Brian D. Strahl, Dinshaw J. Patel, and Steven E. Jacobsen. Polymerase IV occupancy at RNA-directed DNA methylation sites requires SHH1. *Nature*, 498(7454):385–389, June 2013.

[83] Jixian Zhai, Sylvain Bischof, Haifeng Wang, Suhua Feng, Tzuu-fen Lee, Chong Teng, Xinyuan Chen, Soo Young Park, Linshan Liu, Javier Gallego-Bartolome, Wanlu Liu, Ian R. Henderson, Blake C. Meyers, Israel Ausin, and Steven E. Jacobsen. One precursor One siRNA model for Pol IV- dependent siRNAs Biogenesis. *Cell*, 163(2):445–455, October 2015.

[84] Ruiqiang Ye, Wei Wang, Taichiro Iki, Chang Liu, Yang Wu, Masayuki Ishikawa, Xueping Zhou, and Yijun Qi. Cytoplasmic assembly and selective nuclear import of Arabidopsis Argonaute4/siRNA complexes. *Molecular Cell*, 46(6):859–870, June 2012.

[85] Markus Kuhlmann and Michael Florian Mette. Developmentally non-redundant SET domain proteins SUVH2 and SUVH9 are required for transcriptional gene silencing in Arabidopsis thaliana. *Plant Molecular Biology*, 79(6):623–633, August 2012.

[86] Xuehua Zhong, Christopher J. Hale, Julie A. Law, Lianna M. Johnson, Suhua Feng, Andy Tu, and Steven E. Jacobsen. DDR complex facilitates global association of RNA Polymerase V to promoters and evolutionarily young transposons. *Nature structural & molecular biology*, 19(9):870–875, September 2012.

[87] Daniel Zilberman, Xiaofeng Cao, and Steven E. Jacobsen. ARGONAUTE4 control of locus-specific siRNA accumulation and DNA and histone methylation. *Science (New York, N.Y.)*, 299(5607):716–719, January 2003.

[88] Andrzej T. Wierzbicki, Thomas Ream, Jeremy R. Haag, and Craig S. Pikaard. RNA Polymerase V transcription guides ARGONAUTE4 to chromatin. *Nature genetics*, 41(5):630–634, May 2009.

[89] Israel Ausin, Todd C. Mockler, Joanne Chory, and Steven E. Jacobsen. IDN1 and IDN2 are required for de novo DNA methylation in Arabidopsis thaliana. *Nature Structural & Molecular Biology*, 16(12):1325–1327, December 2009.

[90] Gudrun Böhmdorfer, M. Jordan Rowley, Jan Kuciński, Yongyou Zhu, Ivan Amies, and Andrzej T. Wierzbicki. RNA-directed DNA methylation requires stepwise binding of silencing factors to long non-coding RNA. *The Plant Journal*, 79(2):181, July 2014. Publisher: Wiley-Blackwell.

[91] Xuehua Zhong, Jiamu Du, Christopher J. Hale, Javier Gallego-Bartolome, Suhua Feng, Ajay A. Vashisht, Joanne Chory, James A. Wohlschlegel, Dinshaw J. Patel, and Steven E. Jacobsen. Molecular Mechanism of Action of Plant DRM De Novo DNA Methyltransferases. *Cell*, 157(5):1050–1060, May 2014.

[92] Zhang-Wei Liu, Jin-Xing Zhou, Huan-Wei Huang, Yong-Qiang Li, Chang-Rong Shao, Lin Li, Tao Cai, She Chen, and Xin-Jian He. Two Components of the RNA-Directed DNA Methylation Pathway Associate with MORC6 and Silence Loci Targeted by MORC6 in Arabidopsis. *PLOS Genetics*, 12(5):e1006026, May 2016. Publisher: Public Library of Science.

[93] Jie Yang, Lianyu Yuan, Ming-Ren Yen, Feng Zheng, Rujun Ji, Tao Peng, Dachuan Gu, Songguang Yang, Yuhai Cui, Pao-Yang Chen, Keqiang Wu, and Xuncheng Liu. SWI3B and HDA6 interact and are required for transposon silencing in Arabidopsis. *The Plant Journal*, 102(4):809–822, 2020. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/tpj.14666.

[94] Shaofang Li, Lee E. Vandivier, Bin Tu, Lei Gao, So Youn Won, Shengben Li, Binglian Zheng, Brian D. Gregory, and Xuemei Chen. Detection of Pol IV/RDR2-dependent transcripts at the genomic scale in Arabidopsis reveals features and regulation of siRNA biogenesis. *Genome Research*, 25(2):235–245, February 2015.

[95] Marvin R. Paule and Robert J. White. Transcription by RNA polymerases I and III. *Nucleic Acids Research*, 28(6):1283–1298, March 2000.

[96] Sarah J. Goodfellow and Joost C. B. M. Zomerdijk. Basic Mechanisms in RNA Polymerase I Transcription of the Ribosomal RNA Genes. *Sub-cellular biochemistry*, 61, 2012.

[97] Tomasz W. Turowski and David Tollervey. Transcription by RNA polymerase III: insights into mechanism and regulation. *Biochemical Society Transactions*, 44(5):1367–1375, October 2016.

[98] D. B. Nikolov and S. K. Burley. RNA polymerase II transcription initiation: A structural view. *Proceedings of the National Academy of Sciences*, 94(1):15–22, January 1997. Publisher: National Academy of Sciences Section: Review.

[99] Jered M. Wendte and Craig S. Pikaard. The RNAs of RNA-directed DNA methylation. *Biochimica Et Biophysica Acta. Gene Regulatory Mechanisms*, 1860(1):140–148, January 2017.

[100] Lianna M. Johnson, Jiamu Du, Christopher J. Hale, Sylvain Bischof, Suhua Feng, Ramakrishna K. Chodavarapu, Xuehua Zhong, Giuseppe Marson, Matteo Pellegrini, David J. Segal, Dinshaw J. Patel, and Steven E. Jacobsen. SRA/SET domain-containing proteins link RNA polymerase V occupancy to DNA methylation. *Nature*, 507(7490):124–128, March 2014.

[101] Zhang-Wei Liu, Chang-Rong Shao, Cui-Jun Zhang, Jin-Xing Zhou, Su-Wei Zhang, Lin Li, She Chen, Huan-Wei Huang, Tao Cai, and Xin-Jian He. The SET Domain Proteins SUVH2 and SUVH9 Are Required for Pol V Occupancy at RNA-Directed DNA Methylation Loci. *PLoS Genetics*, 10(1), January 2014.

[102] Ming Zhou, Ana Marie S. Palanca, and Julie A. Law. Locus-specific control of the de novo DNA methylation pathway in Arabidopsis by the CLASSY family. *Nature genetics*, 50(6):865–873, June 2018.

[103] Heng Zhang, Ze-Yang Ma, Liang Zeng, Kaori Tanaka, Cui-Jun Zhang, Jun Ma, Ge Bai, Pengcheng Wang, Su-Wei Zhang, Zhang-Wei Liu, Tao Cai, Kai Tang, Renyi Liu, Xiaobing Shi, Xin-Jian He, and Jian-Kang Zhu. DTF1 is a core component of RNA-directed DNA methylation and may assist in the recruitment of Pol IV. *Proceedings of the National Academy of Sciences of the United States of America*, 110(20):8290–8295, May 2013.

[104] Shiming Liu, Yu Yu, Ying Ruan, Denise Meyer, Michel Wolff, Lin Xu, Ning Wang, Andre Steinmetz, and Wen-Hui Shen. Plant SET- and RING-associated domain proteins in heterochromatinization. *The Plant Journal*, 52(5):914–926, 2007. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-313X.2007.03286.x.

[105] Tatsuo Kanno, M. Florian Mette, David P. Kreil, Werner Aufsatz, Marjori Matzke, and Antonius J. M. Matzke. Involvement of putative SNF2 chromatin remodeling protein DRD1 in RNA-directed DNA methylation. *Current biology: CB*, 14(9):801–805, May 2004.

[106] Tatsuo Kanno, Etienne Bucher, Lucia Daxinger, Bruno Huettel, Gudrun Böhmdorfer, Wolfgang Gregor, David P. Kreil, Marjori Matzke, and Antonius J. M. Matzke. A structural-maintenance-of-chromosomes hinge domain-containing protein is required for RNA-directed DNA methylation. *Nature Genetics*, 40(5):670–675, May 2008.

[107] Zhihuan Gao, Hai-Liang Liu, Lucia Daxinger, Olga Pontes, Xinjian He, Weiqiang Qian, Huixin Lin, Mingtang Xie, Zdravko J. Lorkovic, Shoudong Zhang, Daisuke Miki, Xianqiang Zhan, Dominique Pontier, Thierry Lagrange, Hailing Jin, Antonius J. Matzke, Marjori Matzke, Craig S. Pikaard, and Jian-Kang Zhu. An RNA polymerase II- and AGO4-associated protein acts in RNA-directed DNA methylation. *Nature*, 465(7294):106–109, May 2010.

[108] Dalen Fultz and R. Keith Slotkin. Exogenous Transposable Elements Circumvent Identity-Based Silencing, Permitting the Dissection of Expression-Dependent Silencing. *The Plant Cell*, 29(2):360–376, 2017.

[109] Saivageethi Nuthikattu, Andrea D. McCue, Kaushik Panda, Dalen Fultz, Christopher DeFraia, Erica N. Thomas, and R. Keith Slotkin. 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, May 2013.

[110] Kaushik Panda, Lexiang Ji, Drexel A. Neumann, Josquin Daron, Robert J. Schmitz, and R. Keith Slotkin. Full-length autonomous transposable elements are preferentially targeted by expression-dependent forms of RNA-directed DNA methylation. *Genome Biology*, 17, August 2016.

[111] Daniel Holoch and Danesh Moazed. RNA-mediated epigenetic regulation of gene expression. *Nature reviews. Genetics*, 16(2):71–84, February 2015.

[112] Andrzej T. Wierzbicki, Jeremy R. Haag, and Craig S. Pikaard. Noncoding Transcription by RNA Polymerase Pol IVb/Pol V Mediates Transcriptional Silencing of Overlapping and Adjacent Genes. *Cell*, 135(4):635–648, November 2008.

[113] Sylvie Lahmy, Dominique Pontier, Emilie Cavel, Danielle Vega, Mahmoud El-Shami, Tatsuo Kanno, and Thierry Lagrange. PolV(PolIVb) function in RNA-directed DNA methylation requires the conserved active site and an additional plant-specific subunit. *Proceedings of the National Academy of Sciences of the United States of America*, 106(3):941–946, January 2009.

[114] Ryan Lister, Ronan C. O'Malley, Julian Tonti-Filippini, Brian D. Gregory, Charles C. Berry, A. Harvey Millar, and Joseph R. Ecker. Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. *Cell*, 133(3):523–536, May 2008.

[115] A. J. Herr, M. B. Jensen, T. Dalmay, and D. C. Baulcombe. RNA polymerase IV directs silencing of endogenous DNA. *Science (New York, N.Y.)*, 308(5718):118–120, April 2005.

[116] Yasuyuki Onodera, Jeremy R. Haag, Thomas Ream, Pedro Costa Nunes, Olga Pontes, and Craig S. Pikaard. Plant nuclear RNA polymerase IV mediates siRNA and DNA methylation-dependent heterochromatin formation. *Cell*, 120(5):613–622, March 2005.

[117] Jeremy R. Haag, Thomas S. Ream, Michelle Marasco, Carrie D. Nicora, Angela D. Norbeck, Ljiljana Pasa-Tolic, and Craig S. Pikaard. 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, December 2012.

[118] Athanasios Dalakouras and Michael Wassenegger. Revisiting RNA-directed DNA methylation. *RNA Biology*, 10(3):453, March 2013. Publisher: Taylor & Francis.

[119] Marjori A. Matzke, Tatsuo Kanno, and Antonius J. M. Matzke. RNA-Directed DNA Methylation: The Evolution of a Complex Epigenetic Pathway in Flowering Plants. *Annual Review of Plant Biology*, 66:243–267, 2015.

[120] Hume Stroud, Maxim V.C. Greenberg, Suhua Feng, Yana V. Bernatavichute, and Steven E. Jacobsen. Comprehensive Analysis of Silencing Mutants Reveals Complex Regulation of the Arabidopsis Methylome. *Cell*, 152(1-2):352–364, January 2013.

[121] Carlos Molina and Erich Grotewold. Genome wide analysis of Arabidopsis core promoters. *BMC Genomics*, 6:25, February 2005.

[122] Guillaume Moissiard, Shawn J. Cokus, Joshua Cary, Suhua Feng, Allison C. Billi, Hume Stroud, Dylan Husmann, Ye Zhan, Bryan R. Lajoie, Rachel Patton McCord, Christopher J. Hale, Wei Feng, Scott D. Michaels, Alison R. Frand, Matteo Pellegrini, Job Dekker, John K. Kim, and Steve Jacobsen. MORC Family ATPases Required for Heterochromatin Condensation and Gene Silencing. *Science (New York, N.Y.)*, 336(6087):1448–1451, June 2012.

[123] Chongyuan Luo, David J. Sidote, Yi Zhang, Randall A. Kerstetter, Todd P. Michael, and Eric Lam. Integrative analysis of chromatin states in Arabidopsis identified potential regulatory mechanisms for natural antisense transcript production. *The Plant Journal: For Cell and Molecular Biology*, 73(1):77–90, January 2013.

[124] Maxim V. C. Greenberg, Angelique Deleris, Christopher J. Hale, Ao Liu, Suhua Feng, and Steven E. Jacobsen. Interplay between Active Chromatin Marks and RNA-Directed DNA Methylation in Arabidopsis thaliana. *PLoS Genetics*, 9(11), November 2013.

[125] Adam J. Bewick, Lexiang Ji, Chad E. Niederhuth, Eva-Maria Willing, Brigitte T. Hofmeister, Xiuling Shi, Li Wang, Zefu Lu, Nicholas A. Rohr, Benjamin Hartwig, Christiane Kiefer, Roger B. Deal, Jeremy Schmutz, Jane Grimwood, Hume Stroud, Steven E. Jacobsen, Korbinian Schneeberger, Xiaoyu Zhang, and Robert J. Schmitz. On the origin and evolutionary consequences of gene body DNA methylation. *Proceedings of the National Academy of Sciences of the United States of America*, 113(32):9111, August 2016. Publisher: National Academy of Sciences.

[126] Hidetoshi Saze, Junko Kitayama, Kazuya Takashima, Saori Miura, Yoshiko Harukawa, Tasuku Ito, and Tetsuji Kakutani. Mechanism for full-length RNA processing of Arabidopsis genes containing intragenic heterochromatin. *Nature Communications*, 4:2301, 2013.

[127] E. P. Geiduschek and G. P. Tocchini-Valentini. Transcription by RNA polymerase III. *Annual Review of Biochemistry*, 57:873–914, 1988.

[128] Tzuu-fen Lee, Sai Guna Ranjan Gurazada, Jixian Zhai, Shengben Li, Stacey A. Simon, Marjori A. Matzke, Xuemei Chen, and Blake C. Meyers. RNA polymerase V-dependent small RNAs in Arabidopsis originate from small, intergenic loci including most SINE repeats. *Epigenetics*, 7(7):781–795, July 2012.

[129] Binglian Zheng, Zhengming Wang, Shengben Li, Bin Yu, Jin-Yuan Liu, and Xuemei Chen. 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, December 2009.

[130] Yijun Qi, Xingyue He, Xiu-Jie Wang, Oleksiy Kohany, Jerzy Jurka, and Gregory J. Hannon. Distinct catalytic and non-catalytic roles of ARGONAUTE4 in RNA-directed DNA methylation. *Nature*, 443(7114):1008–1012, October 2006.

[131] Ruchi Jain, Nahid Iglesias, and Danesh Moazed. Distinct Functions of Argonaute Slicer in siRNA Maturation and Heterochromatin Formation. *Molecular cell*, 63(2):191–205, July 2016.

[132] Carey Fei Li, Olga Pontes, Mahmoud El-Shami, Ian R. Henderson, Yana V. Bernatavichute, Simon W.-L. Chan, Thierry Lagrange, Craig S. Pikaard, and Steven E. Jacobsen. An ARGONAUTE4-containing nuclear processing center colocalized with Cajal bodies in Arabidopsis thaliana. *Cell*, 126(1):93–106, July 2006.

[133] Mahmoud El-Shami, Dominique Pontier, Sylvie Lahmy, Laurence Braun, Claire Picart, Danielle Vega, Mohamed-Ali Hakimi, Steven E. Jacobsen, Richard Cooke, and Thierry Lagrange. Reiterated WG/GW motifs form functionally and evolutionarily conserved ARGONAUTE-binding platforms in RNAi-related components. *Genes & Development*, 21(20):2539–2544, October 2007.

[134] Assaf Zemach, M. Yvonne Kim, Ping-Hung Hsieh, Devin Coleman-Derr, Leor Eshed-Williams, Ka Thao, Stacey L. Harmer, and Daniel Zilberman. The nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin. *Cell*, 153(1):193–205, March 2013.

[135] Allison L. Cohen and Songtao Jia. Non-coding RNAs and the borders of heterochromatin. *Wiley interdisciplinary reviews. RNA*, 5(6):835, November 2014. Publisher: NIH Public Access.

[136] Lianna M. Johnson, Julie A. Law, Anuj Khattar, Ian R. Henderson, and Steven E. Jacobsen. SRA-Domain Proteins Required for DRM2-Mediated De Novo DNA Methylation. *PLoS Genetics*, 4(11), November 2008.

[137] S. L. Tucker, J. Reece, T. S. Ream, and C. S. Pikaard. Evolutionary history of plant multisubunit RNA polymerases IV and V: subunit origins via genome-wide and segmental gene duplications, retrotransposition, and lineage-specific subfunctionalization. *Cold Spring Harbor Symposia on Quantitative Biology*, 75:285–297, 2010.

[138] Lucia Daxinger, Tatsuo Kanno, Etienne Bucher, Johannes van der Winden, Ulf Naumann, Antonius J. M. Matzke, and Marjori Matzke. A stepwise pathway for biogenesis of 24-nt secondary siRNAs and spreading of DNA methylation. *The EMBO Journal*, 28(1):48, January 2009. Publisher: European Molecular Biology Organization.

[139] T Pélissier and M Wassenegger. A DNA target of 30 bp is sufficient for RNA-directed DNA methylation. *RNA*, 6(1):55–65, January 2000.

[140] Taku Sasaki, Tzuu-fen Lee, Wen-Wei Liao, Ulf Naumann, Jo-Ling Liao, Changho Eun, Ya-Yi Huang, Jason L. Fu, Pao-Yang Chen, Blake C. Meyers, Antonius J. M. Matzke, and Marjori Matzke. Distinct and concurrent pathways of Pol II- and Pol IV-dependent siRNA biogenesis at a repetitive trans-silencer locus in Arabidopsis

thaliana. *The Plant Journal: For Cell and Molecular Biology*, 79(1):127–138, July 2014.

[141] Dalen Fultz, Sarah G. Choudury, and R. Keith Slotkin. Silencing of active transposable elements in plants. *Current Opinion in Plant Biology*, 27:67–76, October 2015.

[142] Andrea D McCue, Kaushik Panda, Saivageethi Nuthikattu, Sarah G Choudury, Erica N Thomas, and R Keith Slotkin. ARGONAUTE 6 bridges transposable element mRNA-derived siRNAs to the establishment of DNA methylation. *The EMBO Journal*, 34(1):20–35, January 2015.

[143] Claudia Keller, Raghavendran Kulasegaran-Shylini, Yukiko Shimada, Hans-Rudolf Hotz, and Marc Bühler. Noncoding RNAs prevent spreading of a repressive histone mark. *Nature Structural & Molecular Biology*, 20(8):994–1000, August 2013.

[144] Qing Li, Jonathan I. Gent, Greg Zynda, Jawon Song, Irina Makarevitch, Cory D. Hirsch, Candice N. Hirsch, R. Kelly Dawe, Thelma F. Madzima, Karen M. McGinnis, Damon Lisch, Robert J. Schmitz, Matthew W. Vaughn, and Nathan M. Springer. RNA-directed DNA methylation enforces boundaries between heterochromatin and euchromatin in the maize genome. *Proceedings of the National Academy of Sciences of the United States of America*, 112(47):14728–14733, November 2015.

[145] Olga Pontes, Carey Fei Li, Pedro Costa Nunes, Jeremy Haag, Thomas Ream, Alexa Vitins, Steven E. Jacobsen, and Craig S. Pikaard. The Arabidopsis Chromatin-Modifying Nuclear siRNA Pathway Involves a Nucleolar RNA Processing Center. *Cell*, 126(1):79–92, July 2006.

[146] Hidetoshi Saze, Ortrun Mittelsten Scheid, and Jerzy Paszkowski. Maintenance of CpG methylation is essential for epigenetic inheritance during plant gametogenesis. *Nature Genetics*, 34(1):65–69, May 2003.

[147] Michelle L. Ebbs and Judith Bender. Locus-Specific Control of DNA Methylation by the Arabidopsis SUVH5 Histone Methyltransferase. *The Plant Cell*, 18(5):1166–1176, May 2006.

[148] Frédéric Pontvianne, Todd Blevins, Chinmayi Chandrasekhara, Wei Feng, Hume Stroud, Steven E. Jacobsen, Scott D. Michaels, and Craig S. Pikaard. Histone methyltransferases regulating rRNA gene dose and dosage control in Arabidopsis. *Genes & Development*, 26(9):945–957, May 2012.

[149] M. Jordan Rowley, Gudrun Böhmdorfer, and Andrzej T. Wierzbicki. Analysis of long non-coding RNAs produced by a specialized RNA polymerase in Arabidopsis thaliana. *Methods*, 63(2):160–169, September 2013.

[150] Songbo Huang, Jinbo Zhang, Ruiqiang Li, Wenqian Zhang, Zengquan He, Tak-Wah Lam, Zhiyu Peng, and Siu-Ming Yiu. SOAPsplice: Genome-Wide ab initio Detection of Splice Junctions from RNA-Seq Data. *Frontiers in Genetics*, 2, July 2011.

[151] Yanming Di, Daniel W. Schafer, Jason S. Cumbie, and Jeff H. Chang. The NBP negative binomial model for assessing differential gene expression from RNA-Seq. *Statistical applications in genetics and molecular biology*, 10(1), 2011. ISBN: 1544-6115 Publisher: De Gruyter.

[152] Gudrun Böhmdorfer, Shriya Sethuraman, M Jordan Rowley, Michal Krzyszton, M Hafiz Rothi, Lilia Bouzit, and Andrzej T Wierzbicki. Long non-coding RNA produced by RNA polymerase V determines boundaries of heterochromatin. *eLife*, 5:e19092, October 2016.

[153] Robert Martienssen and Danesh Moazed. RNAi and heterochromatin assembly. *Cold Spring Harbor Perspectives in Biology*, 7(8):a019323, August 2015.

[154] Wanlu Liu, Sascha H. Duttke, Jonathan Hetzel, Martin Groth, Suhua Feng, Javier Gallego-Bartolome, Zhenhui Zhong, Hsuan Yu Kuo, Zonghua Wang, Jixian Zhai, Joanne Chory, and Steven E. Jacobsen. RNA-directed DNA methylation involves co-transcriptional small RNA-guided slicing of Pol V transcripts in Arabidopsis. *Nature plants*, 4(3):181–188, March 2018.

[155] Sylvie Lahmy, Dominique Pontier, Natacha Bies-Etheve, Michèle Laudié, Suhua Feng, Edouard Jobet, Christopher J. Hale, Richard Cooke, Mohamed-Ali Hakimi, Dimitar Angelov, Steven E. Jacobsen, and Thierry Lagrange. Evidence for ARGONAUTE4-DNA interactions in RNA-directed DNA methylation in plants. *Genes & Development*, 30(23):2565–2570, 2016.

[156] Liang Wu, Long Mao, and Yijun Qi. Roles of dicer-like and argonaute proteins in TAS-derived small interfering RNA-triggered DNA methylation. *Plant Physiology*, 160(2):990–999, October 2012.

[157] Lianna M. Johnson, Jiamu Du, Christopher J. Hale, Sylvain Bischof, Suhua Feng, Ramakrishna K. Chodavarapu, Xuehua Zhong, Giuseppe Marson, Matteo Pellegrini, David J. Segal, Dinshaw J. Patel, and Steven E. Jacobsen. SRA- and SET-domain-containing proteins link RNA polymerase V occupancy to DNA methylation. *Nature*, 507(7490):124–128, 2014.

[158] Michael A. Schon, Max J. Kellner, Alexandra Plotnikova, Falko Hofmann, and Michael D. Nodine. NanoPARE: parallel analysis of RNA 5' ends from low-input RNA. *Genome Research*, 28(12):1931–1942, 2018.

[159] A. Vongs, T. Kakutani, R. A. Martienssen, and E. J. Richards. Arabidopsis thaliana DNA methylation mutants. *Science (New York, N.Y.)*, 260(5116):1926–1928, June 1993.

[160] Diego Cuerda-Gil and R. Keith Slotkin. Non-canonical RNA-directed DNA methylation. *Nature Plants*, 2(11):16163, 2016.

[161] Donna M. Bond and David C. Baulcombe. Epigenetic transitions leading to heritable, RNA-mediated de novo silencing in Arabidopsis thaliana. *Proceedings of the*

National Academy of Sciences, 112(3):917–922, January 2015. Publisher: National Academy of Sciences Section: Biological Sciences.

[162] Role of the arabidopsis DRM methyltransferases in de novo DNA methylation and gene silencing. - PubMed - NCBI.

[163] Michelle Marasco, Weiyi Li, Michael Lynch, and Craig S. Pikaard. Catalytic properties of RNA polymerases IV and V: accuracy, nucleotide incorporation and rNTP/dNTP discrimination. *Nucleic Acids Research*, 45(19):11315–11326, November 2017. Publisher: Oxford Academic.

[164] Javier Gallego-Bartolomé, Wanlu Liu, Peggy Hsuanyu Kuo, Suhua Feng, Basudev Ghoshal, Jason Gardiner, Jenny Miao-Chi Zhao, Soo Young Park, Joanne Chory, and Steven E. Jacobsen. Co-targeting RNA Polymerases IV and V Promotes Efficient De Novo DNA Methylation in Arabidopsis. *Cell*, 176(5):1068–1082.e19, February 2019.

[165] Torben Heick Jensen, Alain Jacquier, and Domenico Libri. Dealing with pervasive transcription. *Molecular Cell*, 52(4):473–484, November 2013.

[166] Ansgar Zoch, Tania Auchynnikava, Rebecca V. Berrens, Yuka Kabayama, Theresa Schöpp, Madeleine Heep, Lina Vasiliauskaitė, Yuvia A. Pérez-Rico, Atlanta G. Cook, Alena Shkumatava, Juri Rappsilber, Robin C. Allshire, and Dónal O'Carroll. SPOCD1 is an essential executor of piRNA-directed de novo DNA methylation. *Nature*, 584(7822):635–639, August 2020. Number: 7822 Publisher: Nature Publishing Group.

[167] A structural-maintenance-of-chromosomes hinge domain–containing protein is required for RNA-directed DNA methylation | Nature Genetics.

[168] M. Jordan Rowley, Maria I. Avrutsky, Christopher J. Sifuentes, Ligia Pereira, and Andrzej T. Wierzbicki. Independent Chromatin Binding of ARGONAUTE4 and SPT5L/KTF1 Mediates Transcriptional Gene Silencing. *PLOS Genetics*, 7(6):e1002120, June 2011.

[169] Tom Sean Smith, Andreas Heger, and Ian Sudbery. UMI-tools: Modelling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Research*, page gr.209601.116, January 2017.

[170] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10–12, May 2011.

[171] Ben Langmead and Steven L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359, March 2012.

[172] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16):2078–2079, August 2009.

[173] Aaron R. Quinlan and Ira M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, 26(6):841–842, March 2010.

[174] Felix Krueger and Simon R. Andrews. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, 27(11):1571–1572, April 2011.

[175] Babraham Bioinformatics - Trim Galore!

[176] Altuna Akalin, Matthias Kormaksson, Sheng Li, Francine E Garrett-Bakelman, Maria E Figueroa, Ari Melnick, and Christopher E Mason. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome biology*, 13(10):R87–R87, October 2012.

[177] Masayuki Tsuzuki, Shriya Sethuraman, Adriana N. Coke, M. Hafiz Rothi, Alan P. Boyle, and Andrzej T. Wierzbicki. Broad noncoding transcription suggests genome surveillance by RNA polymerase V. *Proceedings of the National Academy of Sciences*, November 2020. Publisher: National Academy of Sciences Section: Biological Sciences.

[178] M F Mette, J van der Winden, M A Matzke, and A J Matzke. Production of aberrant promoter transcripts contributes to methylation and silencing of unlinked homologous promoters in trans. *The EMBO Journal*, 18(1):241–248, January 1999.

[179] M.F. Mette, W. Aufsatz, J. van der Winden, M.A. Matzke, and A.J.M. Matzke. Transcriptional silencing and promoter methylation triggered by double-stranded RNA. *The EMBO Journal*, 19(19):5194–5201, October 2000.

[180] Xiaofeng Cao and Steven E. Jacobsen. Role of the Arabidopsis DRM Methyltransferases in De Novo DNA Methylation and Gene Silencing. *Current Biology*, 12(13):1138–1144, July 2002.

[181] Rafal Archacki, Ruslan Yatusevich, Daniel Buszewicz, Katarzyna Krzyczmonik, Jacek Patryn, Roksana Iwanicka-Nowicka, Przemyslaw Biecek, Bartek Wilczynski, Marta Koblowska, Andrzej Jerzmanowski, and Szymon Swiezewski. Arabidopsis SWI/SNF chromatin remodeling complex binds both promoters and terminators to regulate gene expression. *Nucleic Acids Research*, 45(6):3116–3129, 2017.

[182] J. Brzeski, W. Podstolski, K. Olczak, and A. Jerzmanowski. Identification and analysis of the Arabidopsis thaliana BSH gene, a member of the SNF5 gene family. *Nucleic Acids Research*, 27(11):2393–2399, June 1999.

[183] Soon-Ki Han, Yi Sang, Americo Rodrigues, BIOL425 F2010, Miin-Feng Wu, Pedro L. Rodriguez, and Doris Wagner. The SWI2/SNF2 chromatin remodeling ATPase BRAHMA represses abscisic acid responses in the absence of the stress stimulus in Arabidopsis. *The Plant Cell*, 24(12):4892–4906, December 2012.

[184] Chenlong Li, Lianfeng Gu, Lei Gao, Chen Chen, Chuang-Qi Wei, Qi Qiu, Chih-Wei Chien, Suikang Wang, Lihua Jiang, Lian-Feng Ai, Chia-Yang Chen, Song-guang Yang, Vi Nguyen, Yanhua Qi, Michael P. Snyder, Alma L. Burlingame, Susanne E. Kohalmi, Shangzhi Huang, Xiaofeng Cao, Zhi-Yong Wang, Keqiang Wu, Xuemei Chen, and Yuhai Cui. Concerted genomic targeting of H3K27 demethylase REF6 and chromatin-remodeling ATPase BRM in Arabidopsis. *Nature Genetics*, 48(6):687–693, 2016.

[185] Sebastian P. Sacharowski, Dominika M. Gratkowska, Elzbieta A. Sarnowska, Paulina Kondrak, Iga Jancewicz, Aimone Porri, Ernest Bucior, Anna T. Rolicka, Rainer Franzen, Justyna Kowalczyk, Katarzyna Pawlikowska, Bruno Huettel, Stefano Torti, Elmon Schmelzer, George Coupland, Andrzej Jerzmanowski, Csaba Koncz, and Tomasz J. Sarnowski. SWP73 Subunits of Arabidopsis SWI/SNF Chromatin Remodeling Complexes Play Distinct Roles in Leaf and Flower Development. *The Plant Cell*, 27(7):1889–1906, July 2015.

[186] Tomasz J. Sarnowski, Gabino Ríos, Jan Jásik, Szymon Swiezewski, Szymon Kaczanowski, Yong Li, Aleksandra Kwiatkowska, Katarzyna Pawlikowska, Marta Koźbiał, Piotr Koźbiał, Csaba Koncz, and Andrzej Jerzmanowski. SWI3 subunits of putative SWI/SNF chromatin-remodeling complexes play distinct roles during Arabidopsis development. *The Plant Cell*, 17(9):2454–2472, September 2005.

[187] Doris Wagner and Elliot M. Meyerowitz. SPLAYED, a novel SWI/SNF ATPase homolog, controls reproductive development in Arabidopsis. *Current biology: CB*, 12(2):85–94, January 2002.

[188] Ramakrishna K. Chodavarapu, Suhua Feng, Yana V. Bernatavichute, Pao-Yang Chen, Hume Stroud, Yanchun Yu, Jonathan A. Hetzel, Frank Kuo, Jin Kim, Shawn J. Cokus, David Casero, Maria Bernal, Peter Huijser, Amander T. Clark, Ute Krämer, Sabeeha S. Merchant, Xiaoyu Zhang, Steven E. Jacobsen, and Matteo Pellegrini. Relationship between nucleosome positioning and DNA methylation. *Nature*, 466(7304):388–392, July 2010.

[189] Jason T. Huff and Daniel Zilberman. Dnmt1-Independent CG Methylation Contributes to Nucleosome Positioning in Diverse Eukaryotes. *Cell*, 156(6):1286–1297, March 2014. Publisher: Elsevier.

[190] David B Lyons and Daniel Zilberman. DDM1 and Lsh remodelers allow methylation of DNA wrapped in nucleosomes. *eLife*, 6:e30674, November 2017.

[191] Max Felle, Helen Hoffmeister, Julia Rothammer, Andreas Fuchs, Josef H. Exler, and Gernot Längst. Nucleosomes protect DNA from DNA methylation in vivo and in vitro. *Nucleic Acids Research*, 39(16):6956–6969, September 2011. Publisher: Oxford Academic.

[192] Clayton K. Collings, Peter J. Waddell, and John N. Anderson. Effects of DNA methylation on nucleosome stability. *Nucleic Acids Research*, 41(5):2918–2931, March 2013.

[193] S. Vinod Kumar and Philip A. Wigge. H2A.Z-Containing Nucleosomes Mediate the Thermosensory Response in Arabidopsis. *Cell*, 140(1):136–147, January 2010.

[194] Israel Ausin, Maxim V. C. Greenberg, Dhirendra K. Simanshu, Christopher J. Hale, Ajay A. Vashisht, Stacey A. Simon, Tzuu-fen Lee, Suhua Feng, Sophia D. Española, Blake C. Meyers, James A. Wohlschlegel, Dinshaw J. Patel, and Steven E. Jacobsen. 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, May 2012.

[195] Meng Xie, Guodong Ren, Chi Zhang, and Bin Yu. 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*, 72(3):491–500, 2012. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-313X.2012.05092.x.

[196] Tomasz J. Sarnowski, Szymon Swiezewski, Katarzyna Pawlikowska, Szymon Kaczanowski, and Andrzej Jerzmanowski. AtSWI3B, an Arabidopsis homolog of SWI3, a core subunit of yeast Swi/Snf chromatin remodeling complex, interacts with FCA, a regulator of flowering time. *Nucleic Acids Research*, 30(15):3412–3421, August 2002.

[197] Cuijun Zhang, Xuan Du, Kai Tang, Zhenlin Yang, Li Pan, Peipei Zhu, Jinyan Luo, Yuwei Jiang, Hui Zhang, Huafang Wan, Xingang Wang, Fengkai Wu, W. Andy Tao, Xin-Jian He, Heng Zhang, Ray A. Bressan, Jiamu Du, and Jian-Kang Zhu. Arabidopsis AGDP1 links H3K9me2 to DNA methylation in heterochromatin. *Nature Communications*, 9(1):4547, October 2018. Number: 1 Publisher: Nature Publishing Group.

[198] Lisa M. Smith, Olga Pontes, Iain Searle, Nataliya Yelina, Faridoon K. Yousafzai, Alan J. Herr, Craig S. Pikaard, and David C. Baulcombe. 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, May 2007. Publisher: American Society of Plant Biologists Section: Research Article.

[199] Clayton K. Collings and John N. Anderson. Links between DNA methylation and nucleosome occupancy in the human genome. *Epigenetics & Chromatin*, 10(1):18, April 2017.

[200] Sarah K. Bowman, Matthew D. Simon, Aimee M. Deaton, Michael Tolstorukov, Mark L. Borowsky, and Robert E. Kingston. Multiplexed Illumina sequencing libraries from picogram quantities of DNA. *BMC Genomics*, 14(1):466, July 2013.

[201] Kaifu Chen, Yuanxin Xi, Xuewen Pan, Zhaoyu Li, Klaus Kaestner, Jessica Tyler, Sharon Dent, Xiangwei He, and Wei Li. DANPOS: dynamic analysis of nucleosome position and occupancy by sequencing. *Genome Research*, 23(2):341–351, February 2013.

[202] Weizhong Chen, Yi Liu, Shanshan Zhu, Christopher D. Green, Gang Wei, and Jing-Dong Jackie Han. Improved nucleosome-positioning algorithm iNPS for accurate nucleosome positioning from sequencing data. *Nature Communications*, 5:4909, September 2014.

[203] Felix Krueger and Simon R. Andrews. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics (Oxford, England)*, 27(11):1571–1572, June 2011.

[204] Altuna Akalin, Matthias Kormaksson, Sheng Li, Francine E. Garrett-Bakelman, Maria E. Figueroa, Ari Melnick, and Christopher E. Mason. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biology*, 13(10):R87, October 2012.

[205] M Hafiz Rothi, Shriya Sethuraman, Jakub Dolata, Alan P Boyle, and Andrzej T Wierzbicki. DNA methylation directs nucleosome positioning in RNA-mediated transcriptional silencing. preprint, Plant Biology, October 2020.

[206] Yukio Kurihara, Akihiro Matsui, Kousuke Hanada, Makiko Kawashima, Junko Ishida, Taeko Morosawa, Maho Tanaka, Eli Kaminuma, Yoshiki Mochizuki, Akihiro Matsushima, Tetsuro Toyoda, Kazuo Shinozaki, and Motoaki Seki. Genome-wide suppression of aberrant mRNA-like noncoding RNAs by NMD in Arabidopsis. *Proceedings of the National Academy of Sciences*, 106(7):2453–2458, February 2009. Publisher: National Academy of Sciences Section: Biological Sciences.

[207] Marcel E. Dinger, Paulo P. Amaral, Timothy R. Mercer, and John S. Mattick. Pervasive transcription of the eukaryotic genome: functional indices and conceptual implications. *Briefings in Functional Genomics*, 8(6):407–423, November 2009. Publisher: Oxford Academic.

[208] Jürgen Brosius. Waste not, want not – transcript excess in multicellular eukaryotes. *Trends in Genetics*, 21(5):287–288, May 2005.

[209] Assaf Zemach, M. Yvonne Kim, Ping-Hung Hsieh, Devin Coleman-Derr, Leor Eshed-Williams, Ka Thao, Stacey L. Harmer, and Daniel Zilberman. The Arabidopsis nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin. *Cell*, 153(1):193–205, March 2013.